




Proceedings of the International Congress of Mathematicians

Hyderabad 2010



Volume IV
Invited Lectures

Editor
Rajendra Bhatia

 HINDUSTAN
BOOK AGENCY

Proceedings of the

**International Congress
of Mathematicians**

Hyderabad, August 19–27, 2010

This page is intentionally left blank



Proceedings of the

International Congress of Mathematicians

Hyderabad, August 19–27, 2010

Volume IV

Invited Lectures

Editor

Rajendra Bhatia

Co-Editors

Arup Pal

G. Rangarajan

V. Srinivas

M. Vanninathan

Technical Editor

Pablo Gastesi

 HINDUSTAN
BOOK AGENCY

Editor

Rajendra Bhatia, Indian Statistical Institute, Delhi

Co-editors

Arup Pal, Indian Statistical Institute, Delhi

G. Rangarajan, Indian Institute of Science, Bangalore

V. Srinivas, Tata Institute of Fundamental Research, Mumbai

M. Vanninathan, Tata Institute of Fundamental Research, Bangalore

Technical editor

Pablo Gastesi, Tata Institute of Fundamental Research, Mumbai

Published by

Hindustan Book Agency (India)

P 19, Green Park Extension

New Delhi 110 016

India

email: info@hindbook.com

<http://www.hindbook.com>

Copyright © 2010, Authors of individual articles.

No part of the material protected by this copyright notice may be reproduced or utilized in any form or by any means, electronic or mechanical, including photocopying, recording or by any information storage and retrieval system, without written permission from the copyright owner, who has also the sole right to grant licences for translation into other languages and publication thereof.

ISBN 978-81-85931-08-3

Exclusive distribution worldwide except India

World Scientific Publishing Co. Pte. Ltd.

5 Toh Tuck Link, Singapore 596224

Contents

13 Probability and Statistics

Itai Benjamini

Random Planar Metrics 2177

Alexei Borodin

Growth of Random Surfaces 2188

Arup Bose*, Rajat Subhra Hazra, and Koushik Saha

Patterned Random Matrices and Method of Moments 2203

David Brydges* and Gordon Slade

Renormalisation Group Analysis of Weakly Self-avoiding Walk in
Dimensions Four and Higher 2232

Frank den Hollander

A Key Large Deviation Principle for Interacting Stochastic Systems 2258

Steven N. Evans

Time and Chance Happeneth to Them all: Mutation, Selection and
Recombination 2275

Claudia Neuhauser

Coevolution in Spatial Habitats 2297

Jeremy Quastel

Weakly Asymmetric Exclusion and KPZ 2310

Qi-Man Shao

Stein's Method, Self-normalized Limit Theory and Applications 2325

Sara van de Geer

ℓ_1 -regularization in High-dimensional Statistical Models 2351

Aad van der Vaart

Bayesian Regularization 2370

In case of papers with several authors, invited speakers at the Congress are marked with an asterisk.

14 Combinatorics

Louis J. Billera

Flag Enumeration in Polytopes, Eulerian Partially Ordered Sets and
Coxeter Groups 2389

Henry Cohn

Order and Disorder in Energy Minimization 2416

Sergei K. Lando

Hurwitz Numbers: On the Edge Between Combinatorics and
Geometry 2444

Bernard Leclerc

Cluster Algebras and Representation Theory 2471

Brendan D. McKay

Subgraphs of Random Graphs with Specified Degrees 2489

J. Nešetřil* and P. Ossona de Mendez

Sparse Combinatorial Structures: Classification and Applications 2502

Eric M. Rains

Elliptic Analogues of the Macdonald and Koornwinder Polynomials 2530

Oliver Riordan

Percolation on Sequences of Graphs 2555

Benny Sudakov

Recent Developments in Extremal Combinatorics: Ramsey and Turán
Type Problems 2579

15 Mathematical Aspects of Computer Science

Peter Bürgisser

Smoothed Analysis of Condition Numbers 2609

Cynthia Dwork

Privacy Against Many Arbitrary Low-sensitivity Queries 2634

Venkatesan Guruswami

Bridging Shannon and Hamming: List Error-correction with
Optimal Rate 2648

Subhash Khot

Inapproximability of NP-complete Problems, Discrete Fourier
Analysis, and Geometry 2676

Daniel A. Spielman

Algorithms, Graph Theory, and Linear Equations in Laplacian
Matrices 2698

Salil Vadhan

The Unified Theory of Pseudorandomness 2723

16 Numerical Analysis and Scientific Computing**Bernardo Cockburn**

The Hybridizable Discontinuous Galerkin Methods 2749

Peter A. Markowich

Numerical Analysis of Schrödinger Equations in the Highly
Oscillatory Regime 2776

Ricardo H. Nochetto

Why Adaptive Finite Element Methods Outperform Classical Ones 2805

Zuowei Shen

Wavelet Frames and Image Restorations 2834

Mary F. Wheeler*, Mojdeh Delshad, Xianhui Kong, Sunil**Thomas, Tim Wildey and Guangri Xue**

Role of Computational Science in Protecting the Environment:
Geological Storage of CO₂ 2864

Jinchao Xu

Fast Poisson-based Solvers for Linear and Nonlinear PDEs 2886

17 Control Theory and Optimization**Hélène Frankowska**

Optimal Control under State Constraints 2915

Satoru Iwata

Submodular Functions: Optimization and Approximation 2943

Yurii Nesterov

Recent Advances in Structural Optimization 2964

Alexander Shapiro

Computational Complexity of Stochastic Programming: Monte
Carlo Sampling Approach 2979

Robert Weismantel

A Cutting Plane Theory for Mixed Integer Optimization 2996

Xu Zhang

A Unified Controllability/Observability Theory for Some Stochastic
and Deterministic Partial Differential Equations 3008

18 Mathematics in Science and Technology

Ellen Baake

Deterministic and Stochastic Aspects of Single-crossover Recombination	3037
---	------

Freddy Delbaen

BSDE and Risk Measures	3054
------------------------------	------

Kazufumi Ito and Karl Kunisch*

Novel Concepts for Nonsmooth Optimization and their Impact on Science and Technology	3061
---	------

Philip K. Maini*, Robert A. Gatenby and Kieran Smallbone

Modelling Aspects of Tumour Metabolism	3091
--	------

Natasa Djurdjevac, Marco Sarich, and Christof Schütte*

On Markov State Models for Metastable Processes	3105
---	------

Nizar Touzi

Second Order Backward SDEs, Fully Nonlinear PDEs, and Applications in Finance	3132
--	------

Zongben Xu

Data Modeling: Visual Psychology Approach and $L_{1/2}$ Regularization Theory	3151
--	------

Xun Yu Zhou

Mathematicalising Behavioural Finance	3185
---	------

19 Mathematics Education and Popularization of Mathematics

Jill Adler

Professional Knowledge Matters in Mathematics Teaching	3213
--	------

20 History of Mathematics

Tinne Hoff Kjeldsen

History of Convexity and Mathematical Programming: Connections and Relationships in Two Episodes of Research in Pure and Applied Mathematics of the 20th Century	3233
--	------

Norbert Schappacher

Rewriting Points	3258
------------------------	------

Author Index	3293
--------------------	------

Section 13

Probability and Statistics

This page is intentionally left blank

Random Planar Metrics

Itai Benjamini*

Abstract

A discussion regarding aspects of several quite different random planar metrics and related topics is presented.

Mathematics Subject Classification (2010). Primary 05C80; Secondary 82B41.

Keywords. First passage Percolation, Quantum gravity, Hyperbolic geometry.

1. Introduction

In this note we will review some aspects of random planar geometry, starting with random perturbation of the Euclidean metric. In the second section we move on to stationary planar graphs, including unimodular random graphs, distributional local limits and in particular the uniform infinite planar triangulation and its scaling limit. The last section is about a non planar random metric, the critical long range percolation, which arises as a discretization of a Poisson process on the space of lines in the hyperbolic plane. Several open problems are scattered throughout the paper. We only touch a small part of this rather diverse and rich topic.

2. Euclidean Perturbed

One natural way to randomly perturb the Euclidean planar metric is that of first passage percolation (FPP), see [25] for background. That is, consider the square grid lattice, denoted \mathbb{Z}^2 , and to each edge assign an i.i.d. random positive length. There are other ways to randomly perturb the Euclidean metric and many features are not expected to be model dependent. Large balls converge after rescaling to a convex centrally symmetric shape, the boundary fluctuations are conjectured to have a Tracy-Widom distribution. The variance of the distance

*Department of Mathematics, Weizmann Institute, Israel.
E-mail: itai.benjamini@weizmann.ac.il.

from origin to $(n, 0)$ is conjectured to be of order $n^{2/3}$. So far only an upper bound of $\frac{n}{\log n}$ was established, see [7]. It is still not known how stable is the shortest path and its length to random perturbation as considered in noise sensitivity theory, see [8, 21]. Also what are the most efficient algorithms to find the shortest path or to estimate its length? When viewed as a random electrical network it is conjectured that the variance of the resistance from the origin to $(n, 0)$ is uniformly bounded, see [10].

Consider random lengths chosen as follows: 1 with probability $p > 1/2$ and ∞ otherwise. Look at the convex hull of all vertices with distance less than n to the origin (assuming the origin is in the infinite cluster). Simulations suggest that as $p \searrow 1/2$ the limiting shape converges to a Euclidean ball. This is still open but heuristically supported by the conformal invariance of critical Bernoulli percolation.

The structure of geodesic rays and two sided infinite geodesics in first passage percolation is still far from understood. Furstenberg asked in the 80's (following a talk by Kesten) to show that almost surely there are no two sided infinite geodesics for natural FPP's, e.g. exponential length on edges.

Häggström and Pemantle introduced [22] competitions based on FPP, see [18] for a survey. Here is a related problem. Start two independent simple random walks on \mathbb{Z}^2 walking with the same clock, with the one additional condition, the walkers are not allowed to step on vertices already visited by the other walk, and otherwise chose uniformly among allowed vertices. Show that almost surely, one walker will be trapped in a finite domain. Prove that this is not the case in higher dimensions.

3. Unimodular Random Graphs, Uniform Random Triangulations

There is a recent growing interest in graph limits, see e.g. [31] for a diversity of viewpoints. In parallel the theory of random triangulations was developed as a toy model of quantum gravity, initially by physicists. Angel and Schramm [2, 3] constructed the uniform infinite planar triangulation (UIPT), a rooted infinite unimodular random triangulation which is the limit (in the sense of [11]) of finite random triangulations (the uniform measure on all non isomorphic triangulations of the sphere of size n), a model that was studied extensively by many (see e.g. [26]). Exponential of the Gaussian free field (GFF) provides a model of random measure on the plane, see [19].

Therefore in the theory of random uniform planar graphs and triangulations we encounter several view points and many missing links. The general theory of unimodular random graphs [11, 1] is useful in deducing certain properties, giving a notion of “stationary” graph in the spirit of stationary process. This is a measure on graphs rooted at a directed edge which is invariant for rerooting along a random walk path. This rather minimal assumption turned out to be

a surprisingly strong generalization of Cayley graphs, or transitive unimodular graphs. Conformal geometry is useful in the bounded degree set up. Enumeration is useful when no restriction on the degree is given. See the recent work [13] and references there, for the success of enumeration techniques. The key to the success of enumeration is the spatial Markov property. It is of interest to classify which other distributions on rooted infinite triangulations enjoys this property? The links to the Gaussian free field is only a conjecture at the moment, and a method of constructing a conformal invariant random path metric on the real plane from the Gaussian free field is still eluding. There are many open problems in any of the models. Here are a few, for more in particular regarding extensions unimodular planar triangulation's other then the UIPT or the UIPQ see [6]:

1. Angel and Schramm [3] conjectured that the UIPT is a.s. recurrent. At what rate does the resistance grow? Note that the local limit of bounded degree finite planar graphs is recurrent [11]. The degree distribution of UIPT has an exponential tail. It is of interest to understand the structure of large random triangulations conditioned on having degree smaller than some fixed constant. Adapt the enumeration techniques to the bounded degree set up. [6] subdiffusivity of the simple random walk on the UIPQ was established with exponent 1/3 as upper bound on the displacement by time n . Denote the SRW by X_k . What is the true exponent $1/3 \geq \alpha \geq 0$ such that

$$\sup_{0 \leq k \leq n} \text{dist}(o, X_k) \asymp n^\alpha?$$

Does $\alpha = 1/4$?

2. The UIPT is Liouville (no non constant bounded harmonic functions) and any unimodular graph of subexponential asymptotic volume growth, see [6]. Show that if G is planar, Liouville and unimodular then G is recurrent?
3. View a large finite triangulation as an electrical network. Understanding the effective resistance will make it possible to study the Gaussian free field on the triangulation. The Laplacian spectrum and eigenfunctions nodal domain and level sets are of interest, see [20] for background.
4. Show that if G is a distributional limit (in the sense of [11]) of finite planar graphs then the critical probability for percolation on G satisfies $p_c(G, \text{site}) \geq 1/2$ a.s. and no percolation at the critical probability. This last fact should hold for any unimodular planar graph.
5. Consider the $n \times n$ grid equipped with the Gaussian free field with no boundary conditions. The exponential of the field gives a positive "length" to each vertex. We get a random metric on the square grid. Let $\gamma_1(n)$ be the shortest path between the top corners and $\gamma_2(n)$ the shortest path

between the bottom corners. Show there is $c > 0$, so that for any n , $P(\{\gamma_1(n) \cap \gamma_2(n) \neq \emptyset\}) > c$. Identify the scaling limit of $\gamma_1(n)$? Establish and study the scaling limit of these metric spaces. How do geodesics concentrate around a fixed height of the field? What is the dimension of the geodesics? Since scaling limits of geodesics likely have Euclidean dimension strictly bigger than one, it suggests that geodesics wind a every scale and therefore “forget” the starting point. Thus likely the limit is rotationally invariant and maybe close to Schramm’s SLE_κ curve, for what κ ?

6. The following deterministic statement, just proved together with Panagiotis Papazoglou, might help clear up several aspects regarding the geometry of the random triangulation. G planar with polynomial volume growth r^β , then there are arbitrarily large finite domains Ω_n in G with boundary of size at most $|\Omega_n|^{1/\beta}$? Does having such an isoperimetric upperbound implies an upperbound on the exponent α in the displacement of SRW from certain starting points and certain times ($\text{dist}(o, X_n) \asymp n^\alpha$)?
7. Adapt the enumeration techniques to the bounded degree set up. Devise an algorithm to sample uniformly a large finite triangulation, chosen uniformly among planar maps of size n and degree at most d . Try to formulate and/or prove something in higher dimensions, see [5].

The coming three subsections discuss the scaling limit of finite random planar maps and harmonic measure for random walks on random triangulations.

3.1. Scaling limit of Planar maps. A planar map m is a proper embedding of a planar graph into the two dimensional sphere \mathbb{S}_2 seen up to deformations. A *quadrangulation* is a rooted planar map such that all faces have degree 4. For sake of simplicity we will only deal with these maps (see universality results). Let m_n be a uniform variable on the set \mathcal{Q}_n of quadrangulations with n faces and v_n be a vertex picked uniformly in m_n . The radius of (m_n, v_n) is

$$r_n = \max_{v \in \text{Vertices}(m_n)} d_{\text{gr}}(v_n, v).$$

In their pioneer work, Chassaing and Schaeffer [17] showed that the rescaled radii converge in law toward the radius r of the one-dimensional Integrated Super Brownian Excursion (ISE),

$$n^{-1/4} r_n \xrightarrow{(\text{law})} \left(\frac{8}{9}\right)^{1/4} r.$$

The key ingredient is a bijective encoding of rooted quadrangulations by labelled trees due to Cori-Vauquelin and Schaeffer [33]. This was the first proof of the physicist’s conjecture that the distance in a typical map of size n should behave like $n^{1/4}$. Nevertheless this convergence does not allow us to understand the

whole metric structure of a large map. To do this, we should consider a map endowed with its graph distance d_{gr} as a metric space and ask for convergence in the sense of Gromov-Hausdorff metric (see [15]). In other words, if m_n is uniform on \mathcal{Q}_n , we wonder whether the following weak convergence for the Gromov-Hausdorff metric occurs

$$(m_n, n^{-1/4}d_{gr}) \xrightarrow{?} (m_\infty, d_\infty), \tag{3.1}$$

where (m_∞, d_∞) is a random compact metric space. Unfortunately, the convergence (3.1) is still unproved and constitutes the main open problem in this area. Nevertheless, Le Gall has shown in [28] that (3.1) is true along subsequences. Thus we are left with a family of random metric spaces called Brownian maps which are precisely the limiting points of the the sequence $(m_n, n^{-1/4}d_{gr})$ for the weak convergence of probability measures with respect to Gromov-Hausdorff distance. One conjectures that there is no need to take a subsequence, that is all Brownian maps have the same law. Still one can establish properties shared by all Brownian maps e.g.

Theorem 3.1 ([28],[30]). *Let (m_∞, d_∞) be a Brownian map. Then*

- (a) *Almost surely, the Hausdorff dimension of (m_∞, d_∞) is 4.*
- (b) *Almost surely, (m_∞, d_∞) is homeomorphic to \mathbb{S}_2 .*

In a recent work [29], Le Gall completely described the geodesics toward a distinguished point and the description is independent of the Brownian map considered. Here are some extensions and open problems:

1. Although we know that Brownian maps share numerous properties, they do not seem sufficient to identify the law and thus prove (3.1). In a forthcoming paper by Curien, Le Gall and Miermont, they show the convergence (without taking any subsequence) of the so-called ‘‘Cactus’’ associated to m_n .
2. The law of mutual distances between p -points is sufficient to characterize the law of a random metric space. For $p = 2$, the distance in any Brownian map between two random independent points can be expressed in terms of ISE. Recently the physicists Bouttier and Guitter [16] solved the case $p = 3$. Unfortunately their techniques do not seem to extend to four points.

3.2. QG and GFF. Let \mathcal{T}_n be the set of all triangulations of the sphere \mathbb{S}_2 with n faces with no loops or multiple edges. We recall the well known circle packing theorem (see Wikipedia, [23]):

Theorem 3.2. *If T is a finite triangulation without loops or multiple edges then there exists a circle packing $P = (P_c)_{c \in C}$ in the sphere \mathbb{S}_2 such that the contact graph of P is T . This packing is unique up to M\"obius transformations.*

Recall that the group of Möbius transformations $z \mapsto \frac{az+b}{cz+d}$ for $a, b, c, d \in \mathbb{C}$ with $ad - bc \neq 0$ can be identified with $\mathrm{PSL}_2(\mathbb{C})$ and act transitively on triplets (x, y, z) of \mathbb{S}_2 . The circle packing enables us to take a “nice” representation of a triangulation $T \in \mathcal{T}_n$, nevertheless the non-uniqueness is somehow disturbing because to fix a representation we can, for example, fix the images of three vertices of a distinguished face of T . This specification breaks all the symmetry, because sizes of some circles are chosen arbitrarily. Here is how to proceed:

Barycenter of a measure on \mathbb{S}_2 . The action on \mathbb{S}_2 of an element $\gamma \in \mathrm{PSL}_2(\mathbb{C})$ can be continuously extended to $\mathbb{B}_3 := \{(x, y, z) \in \mathbb{R}^3, x^2 + y^2 + z^2 \leq 1\}$: this is the Poincaré-Beardon extension. We will keep the notation γ for transformations $\mathbb{B}_3 \rightarrow \mathbb{B}_3$. The action of $\mathrm{PSL}_2(\mathbb{C})$ on \mathbb{B}_3 is now transitive on points. The group of transformations that leave 0 fixed is precisely the group $\mathrm{SO}_2(\mathbb{R})$ of rotations of \mathbb{R}^3 .

Theorem 3.3 (Douady-Earle). *Let μ be a measure on \mathbb{S}_2 such that $\#\mathrm{supp}(\mu) \geq 2$. Then we can associate to μ a “barycenter” denoted by $\mathrm{Bar}(\mu) \in \mathbb{B}_3$ such that for all $\gamma \in \mathrm{PSL}_2(\mathbb{C})$ we have*

$$\mathrm{Bar}(\gamma^{-1}\mu) = \gamma(\mathrm{Bar}(\mu)).$$

We can now describe the renormalization of a circle packing. If P is a circle packing associated to a triangulation $T \in \mathcal{T}_n$, we can consider the atomic measure μ_P formed by the Dirac’s at centers of the spheres in P

$$\mu_P := \frac{1}{\#P} \sum_{x \text{ centers of } P} \delta_x.$$

By transitivity there exists a conformal map $\gamma \in \mathrm{PSL}_2(\mathbb{C})$ such that $\mathrm{Bar}(\gamma^{-1}\mu_P) = 0$. The renormalized circle packing is by Definition $\gamma(P)$, this circle packing is unique up to rotation of $\mathrm{SO}_2(\mathbb{R})$, we will denote it by \mathbf{P}_T . This constitutes a canonical discrete conformal structure for the triangulation.

Problems. If T_n is a random variable uniform over the set \mathcal{T}_n , then the variable $\mu_{\mathbf{P}_{T_n}}$ is a random probability measure over \mathbb{S}_2 seen up to rotations of $\mathrm{SO}_2(\mathbb{R})$. By standard arguments there exist weak limits μ_∞ of $\mu_{\mathbf{P}_{T_n}}$.

1. (Schramm [Talk about QG]) Determine coarse properties (invariant under $\mathrm{SO}_2(\mathbb{R})$) of μ_∞ , e.g. what is the dimension of the support? Start with showing singularity.
2. Uniqueness (in law) of μ_∞ ? In particular can we describe μ_∞ in terms of GFF? Is it $\exp((8/3)^{1/2}GFF)$, does KPZ hold? see [19].
3. The random measure μ_∞ can come together with d_∞ a random distance on \mathbb{S}_2 (in the spirit of [28]). Can you describe links between μ_∞ and d_∞ ? Does one characterize the other? Is it a path metric space?

3.3. Harmonic measure and recurrence. Our goal in this subsection is to remark that if a graph is recurrent then harmonic measure on boundaries of domains can not be very spread and supported uniformly on (too) large sets. We have in mind random triangulations. We first discuss general graphs.

Let G denote a bounded degree infinite graph. Fix a base vertex v and denote by $B(r)$ the ball of radius r centered at v , by $\partial B(r)$ the boundary of the ball, that is vertices with distance r from v . Denote by μ_r the harmonic measure for simple random walk starting at v on $\partial B(r)$.

Assume simple random walk (SRW) on G is recurrent. Further assume that there are arbitrarily large excursions attaining the maximum distance once, this happens in many natural examples but not always (e.g. consider the graph obtained by starting with a ray and adding to the ray a full n levels binary tree rooted at the vertex on the ray with distance n to the root, for all n). The maximum of SRW excursion on \mathbb{Z} is attained a tight number of times. It is reasonable to believe that if each of the vertices in $\partial B(r)$ admit a neighbor in $\partial B(r + 1)$, then the same conclusion will hold.

Proposition 3.4. *Under the stronger further assumption above, for infinitely many r 's,*

$$\sum_{u \in \partial B(r)} \mu_r(u)^2 > \frac{1}{r \log^2 r}.$$

Note that for the uniform measure, U_r , on $\partial B(r)$, $\sum_{u \in \partial B(r)} U_r(u)^2 = |\partial B(r)|^{-1}$.

Gady Kozma constructed a recurrent bounded degree planar graph (not a triangulation) for which harmonic measure on any minimal cutsets outside $B(r)$ for any r is supported on a set of size at least $r^{4/3}$, or even larger exponents. The example is very “irregular”, it will be useful to come up with a natural general condition that will guarantee a linear support.

Proof. What is the probability SRW will reach maximal distance r once, before returning back to v ? By summing all paths from v to $\partial B(r)$ and back to v visiting $\partial B(r)$ and v once, we get that up to a constant depending on the degree the answer is

$$\sum_{u \in \partial B(r)} \mu_r(u)^2.$$

But by our assumptions then,

$$\sum_r \sum_{u \in \partial B(r)} \mu_r(u)^2 = \infty.$$

Observe that the events “excursion to maximal distance n from the origin” are independent for different n 's. \square

We next consider planar triangulations. Rather than working in the context of abstract graph it is natural to circle pack them and use conformal geometry. Assume G is a bounded degree recurrent infinite planar triangulation. By He and Schramm [23], G admits a circle packing in the whole Euclidean plane. Fix a root for G .

Question: Is it the case that for arbitrarily large radii r , there are domains containing a ball of radius r around the root, so that harmonic measure on the domain boundary is supported on $r^{1+o(1)}$ circles?

By *supported* we mean $1-o(1)$ of the measure is supported on the set. Here is a possible approach: Consider a huge ball in the infinite recurrent triangulation. Circle pack the infinite recurrent triangulation in the whole plane [23]. Look at the Euclidean domain which is the image of this ball. Random walk on the triangulation will be close to SRW on hexagonal packing inside this domain. By the discrete adaptation of Makarov's theorem [27, 32], harmonic measure on the boundary circles will be supported on a linear number of hexagonal circles. How can we see that no more original circles are needed for some of the domains, using recurrence? Note that for hyperbolic triangulations this is not the case.

It might be the case that this is not true for general triangulation but further assuming unimodularity will do the job. In particular is it true for the UIPT?

4. Random Hyperbolic Lines

Following the Euclidean random graph and the conjecturally recurrent UIPT we move on to the hyperbolic plane.

In [9] it was shown that a.s. the components of the complement of a Poisson process on the space of hyperbolic geodesics in the hyperbolic plane are bounded iff the intensity of the process is bigger or equal one, when the hyperbolic plane is scaled to have -1 curvature. This sharp transition and rapid mixing of the geodesic flow suggests that when removing from a compact hyperbolic surface the initial segment of a random geodesic, then the size of the largest component of the complement drops in a sharp transition from order the size of the surface to a logarithmic in the size of the surface. In the coming subsections we will discuss two different directions inspired by this Poisson process of hyperbolic lines.

4.1. Vacant sets. Random geodesics on an hyperbolic surface mix rapidly, this further suggests that the vacant set of non backtracking or even simple walk path on a “well connected” graph will also admit a sharp percolation-like transition. That is, the amount of randomness and independence in processes

such as random walk on uniformly transient graphs are sufficient to create phase transitions usually seen in the context of independent percolation or other spin systems such as the random cluster, or Potts models. In [12] there are initial results towards understanding this phenomena.

Let G_n be a sequence of finite transitive graphs $|G_n| \rightarrow \infty$ which are uniformly transient (that is, when viewing the edges as one Ohm conductors the electric resistance between any pair of vertices in any of the G_n 's is uniformly bounded).

Conjecture 4.1. *Show that the size of the largest vacant component of simple random walk on G_n 's drops from order $|G_n|$ to $o(|G_n|)$ after less than $C|G_n|$ steps, for some $C < \infty$ fixed and in an interval of width $o(|G_n|)$.*

Note that the n^d -Euclidean grid tori satisfies the assumption when $d > 2$.

The following *conjecture* which is still open, is relevant for this problem. The probability to cover a graph by SRW in order size steps is exponentially small. Formally, for any $C < \infty$ there is $c < 1$, so that for any graph G of size n and no double edges, the probability Simple Random Walk covers G in Cn steps is smaller than c^n .

4.2. Long range percolation. Consider this Poisson line process (from [9]) with intensity λ on the upper half plane model for the hyperbolic plane. For each pair $x, y \in \mathbb{Z}$, let there be an edge between x and y (independently for different pairs) iff there is a line in the line process with one endpoint in $[x, x+1]$ and the other in $[y, y+1]$. Then a calculation shows that the probability that there is an edge between x and y is asymptotic to $\lambda/|x-y|^2$ as $|x-y| \rightarrow \infty$. We just recovered the standard *long range percolation* model on \mathbb{Z} with critical exponent 2 (see [4]). The critical case of long range percolation is not well understood. The fact that it is a discretization of the Möbius invariant process hopefully will be useful and already indicates that the process is somewhat natural.

Here is a direct formulation. Start with the one dimensional grid \mathbb{Z} with the nearest neighbor edges, add to it additional edges as follows. Between, i and j add an edge with probability $\beta|i-j|^{-2}$, independently for each pair. The main open problem is how does the distance between 0 and n grow typically in this random graph? The answer is believed to be of the form $\theta(n^{f(\beta)})$, where f is strictly between 0 and 1 and is strictly monotone in β . When -2 is replaced by another exponent the answers are known, see [4, 14].

Acknowledgements. Thanks to Gady Kozma and Scott Sheffield for useful discussions. Thanks to Nicolas Curien for composing the subsection on scaling limits. Thanks to Nicolas Curien and Frederic Paulin for reconstructing Oded's suggestion in the subsection on quantum gravity and the GFF.

References

- [1] D. Aldous and R. Lyons, Processes on unimodular random networks, *Electron. J. Probab.* paper 54, pages 1454-1508 (2007).
- [2] O. Angel, Growth and percolation on the uniform infinite planar triangulation, *Geometric And Functional Analysis* **13** 935–974 (2003).
- [3] O. Angel and O. Schramm, Uniform infinite planar triangulations, *Comm. Math. Phys.* **241** , 191–213 (2003).
- [4] I. Benjamini and N. Berger, The diameter of long-range percolation clusters on finite cycles, *Random Structures Algorithms* **19**, 102–111 (2001).
- [5] I. Benjamini and N. Curien, Local limit of Packable graphs, <http://arxiv.org/abs/0907.2609>
- [6] I. Benjamini and N. Curien, The Simple Random Walk on the Uniform Infinite Planar Quadrangulation is Subdiffusive, (2010).
- [7] I. Benjamini, G. Kalai and O. Schramm, First Passage Percolation Has Sublinear Distance Variance, *Ann. of Probab.* **31** 1970–1978 (2003).
- [8] I. Benjamini, G. Kalai and O. Schramm, Noise sensitivity of Boolean functions and applications to percolation, *Publications Mathématiques de L’IHES* **90** 5–43 (1999).
- [9] I. Benjamini, J. Jonasson, O. Schramm and J. Tykesson, Visibility to infinity in the hyperbolic plane, despite obstacles, *ALEA* **6**, 323–342 (2009).
- [10] I. Benjamini and R. Rossignol, Submean Variance Bound for Effective Resistance of Random Electric Networks, *Comm. Math. Physics* **280**, 445–462 (2008).
- [11] I. Benjamini and O. Schramm, Recurrence of Distributional Limits of Finite Planar Graphs, *Electron. J. Probab.* **6** 13 pp. (2001).
- [12] I. Benjamini and A-S. Sznitman, Giant Component and Vacant Set for Random Walk on a Discrete Torus, *J. Eur. Math. Soc.* **10**, 133–172 (2008).
- [13] O. Bernardi and M. Bousquet-Mélou, Counting colored planar maps: algebraicity results, arXiv:0909.1695
- [14] M. Biskup, On the scaling of the chemical distance in long-range percolation models, *Ann. Probab.* **32** 2938–2977 (2004).
- [15] D. Burago, Y. Burago, and S. Ivanov, *A course in metric geometry*, volume 33 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, RI, (2001).
- [16] J. Bouttier and E. Guitter, The three-point function of planar quadrangulations, *J. Stat. Mech. Theory Exp.*, (7):P07020, 39, (2008).
- [17] P. Chassaing and G. Schaeffer, Random planar lattices and integrated super-Brownian excursion, *Prob. Theor. and Rel. Fields* **128**, 161–212 (2004).
- [18] M. Deijfen, O. Häggström, The pleasures and pains of studying the two-type Richardson model, *Analysis and Stochastics of Growth Processes and interface models*, 39–54 (2008).
- [19] B. Duplantier and S. Sheffield, Liouville Quantum Gravity and KPZ, arXiv:0808.1560

-
- [20] Y. Elon, Eigenvectors of the discrete Laplacian on regular graphs a statistical approach, *J. Phys. A: Math. Theor.* **41** 435203 (17pp)(2008).
- [21] C. Garban, Oded Schramm's contributions to Noise Sensitivity, *Ann. Prob.* (2009) To appear
- [22] O. Häggström and R. Pemantle, First passage percolation and a model for competing spatial growth, *Jour. Appl. Proba.* **35** 683–692 (1998)
- [23] Z-X. He and O. Schramm, Hyperbolic and parabolic packings, *Jour. Discrete and Computational Geometry* **14** 123–149, (1995).
- [24] C. Hoffman, Geodesics in first passage percolation, *Ann. Appl. Prob.* **18**,1944–1969 (2008).
- [25] H. Kesten, Aspects of first passage percolation, *Lecture Notes in Math*, Springer (1986)
- [26] M. A. Krikun. A uniformly distributed infinite planar triangulation and a related branching process, *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov. (POMI)*, 307(Teor. Predst. Din. Sist. Komb. i Algoritm. Metody. 10):141–174, 282–283 (2004).
- [27] G. Lawler, A Discrete Analogue of a Theorem of Makarov, *Combinatorics, Probability and Computing* **2** 181–199, (1993).
- [28] J.F. Le Gall, The topological structure of scaling limits of large planar maps, *Invent. Math.*, 169(3):621–670 (2007).
- [29] J.F. Le Gall, Geodesics in large planar maps and in the brownian map, *preprint available on arxiv*, (2009).
- [30] J.F. Le Gall and F. Paulin, Scaling limits of bipartite planar maps are homeomorphic to the 2-sphere, *Geom. Funct. Anal.*, 18(3):893–918 (2008).
- [31] L. Lovász, Very large graphs, *Current Developments in Mathematics* (2008), to appear
- [32] N. Makarov, On the distortion of boundary sets under conformal mappings, *Proc. London Math. Soc.* **51** 369–384 (1985).
- [33] G. Schaeffer, Conjugaison d'arbres et cartes combinatoires aléatoires, phd thesis. (1998).

Growth of Random Surfaces

Alexei Borodin*

Abstract

We describe a class of exactly solvable random growth models of one and two-dimensional interfaces. The growth is local (distant parts of the interface grow independently), it has a smoothing mechanism (fractal boundaries do not appear), and the speed of growth depends on the local slope of the interface.

The models enjoy a rich algebraic structure that is reflected through closed determinantal formulas for the correlation functions. Large time asymptotic analysis of such formulas reveals asymptotic features of the emerging interface in different scales. Macroscopically, a deterministic limit shape phenomenon can be observed. Fluctuations around the limit shape range from universal laws of Random Matrix Theory to conformally invariant Gaussian processes in the plane. On the microscopic (lattice) scale, certain universal determinantal random point processes arise.

Mathematics Subject Classification (2010). Primary 82C41; Secondary 60B10, 60G55, 60K35.

Keywords. Random growth, determinantal point processes, Gaussian free field

1. Introduction

In recent years there has been a lot of progress in understanding large time fluctuations of driven interacting particle systems on the one-dimensional lattice. Evolution of such systems is commonly interpreted as random growth of a one-dimensional interface, and if one views the time as an extra variable, the evolution produces a random surface. In a different direction, substantial progress has also been achieved in studying the asymptotics of random surfaces arising from dimers on planar bipartite graphs.

Although random surfaces of these two kinds were shown to share certain asymptotic properties, also common to random matrix models, no direct

*Mathematics 253-37, Caltech, Pasadena CA 91125, USA, and Dobrushin Mathematics Laboratory, IITP RAS, Moscow 101447, Russia. E-mail: borodin@caltech.edu.

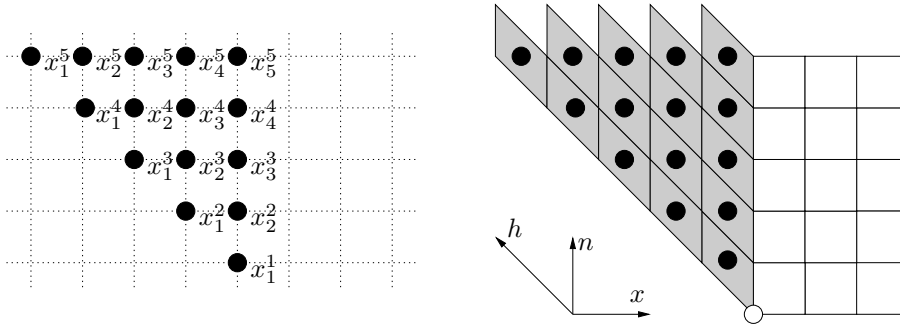


Figure 2.1. The densely packed initial conditions.

connection between them was known. Our original motivation was to find such a connection.

We were able to construct a class of two-dimensional random growth models that in two different projections yield random surfaces of these two kinds (one projection reduces the spatial dimension by one, the second projection is fixing time). It became clear that studying such models of random surface growth should be viewed as a subject on its own, and the goal of this note is to survey its different faces. In Section 2 we define one model of random surface growth and explain what is known about it. This model can be approached from several different directions, and in Section 3 we show how different viewpoints naturally lead to various generalizations of the original model.

2. A Two-dimensional Growth Model

Consider a continuous time Markov chain on the state space of interlacing variables

$$\mathcal{S}^{(n)} = \left\{ \{x_k^m\}_{\substack{k=1, \dots, m \\ m=1, \dots, n}} \subset \mathbb{Z}^{\frac{n(n+1)}{2}} \mid x_{k-1}^m < x_{k-1}^{m-1} \leq x_k^m \right\}, \quad n = 1, 2, \dots \quad (2.1)$$

As initial condition, we consider the fully-packed one, namely at time moment $t = 0$ we have $x_k^m(0) = k - m - 1$ for all k, m , see Figure 2.1.

The particles evolve according to the following dynamics. Each of the particles x_k^m has an independent exponential clock of rate one, and when the x_k^m -clock rings the particle attempts to jump to the right by one. If at that moment $x_k^m = x_k^{m-1} - 1$ then the jump is blocked. If that is not the case, we find the largest $c \geq 1$ such that $x_k^m = x_{k+1}^{m+1} = \dots = x_{k+c-1}^{m+c-1}$, and all c particles in this string jump to the right by one. A Java simulation of this dynamics can be found at <http://www-wt.iam.uni-bonn.de/~ferrari/animations/AnisotropicKPZ.html>. For any $t \geq 0$ denote by $\mathcal{M}^{(n)}(t)$ the resulting measure on $\mathcal{S}^{(n)}$ at time moment t .

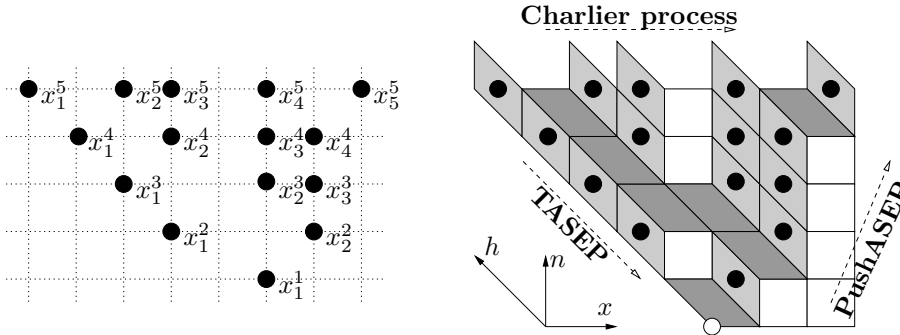


Figure 2.2. From particle configurations (left) to 3d visualization (right).

Informally speaking, the particles with smaller upper indices are heavier than those with larger upper indices, so that the heavier particles block and push the lighter ones in order for the interlacing conditions to be preserved.

Let us illustrate the dynamics using Figure 2.2, which shows a possible configuration of particles obtained from our initial condition. If in this state of the system the x_1^3 -clock rings, then particle x_1^3 does not move, because it is blocked by particle x_1^2 . If it is the x_2^2 -clock that rings, then particle x_2^2 moves to the right by one unit, but to keep the interlacing property satisfied, also particles x_3^3 and x_4^4 move by one unit at the same time.

Observe that $\mathcal{S}^{(n_1)} \subset \mathcal{S}^{(n_2)}$ for $n_1 \leq n_2$, and the definition of the evolution implies that $\mathcal{M}^{(n_1)}(t)$ is a marginal of $\mathcal{M}^{(n_2)}(t)$ for any $t \geq 0$. Thus, we can think of $\mathcal{M}^{(n)}$'s as marginals of the measure $\mathcal{M} = \lim_{\leftarrow} \mathcal{M}^{(n)}$ on $\mathcal{S} = \lim_{\leftarrow} \mathcal{S}^{(n)}$. In other words, $\mathcal{M}(t)$ are measures on the space \mathcal{S} of infinite point configurations $\{x_k^m\}_{k=1, \dots, m, m \geq 1}$.

The Markov chain described above has different interpretations. Also, some projections of the Markov chain to subsets of $\mathcal{S}^{(n)}$ are Markovian.

1. The evolution of x_1^1 is the one-dimensional Poisson process of rate one.
2. The set of left-most particles $\{x_1^m\}_{m \geq 1}$ evolves as a Markov chain on \mathbb{Z} known as the *Totally Asymmetric Simple Exclusion Process* (TASEP), and the initial condition $x_1^m(0) = -m$ is commonly referred to as *step initial condition*. In this case, particle x_1^k jumps to its right with unit rate, provided the arrival site is empty (exclusion constraint).
3. The set of right-most particles $\{x_m^m\}_{m \geq 1}$ also evolves as a Markov chain on \mathbb{Z} that is sometimes called "long range TASEP"; it was also called PushASEP in [6]. It is convenient to view $\{x_m^m + m\}_{m \geq 1}$ as particle locations in \mathbb{Z} . Then, when the x_k^k -clock rings, the particle $x_k^k + k$ jumps to its right and pushes by one unit the (maybe empty) block of particles sitting next to it. If one disregards the particle labeling, one can think

of particles as independently jumping to the next free site on their right with unit rate.

4. For our initial condition, the evolution of each row $\{x_k^m\}_{k=1,\dots,m}$, $m = 1, 2, \dots$, is also a Markov chain. It was called *Charlier process* in [30] because of its relation to the classical orthogonal Charlier polynomials. It can be defined as Doob h -transform for m independent rate one Poisson processes with the harmonic function h equal to the Vandermonde determinant.
5. Infinite point configurations $\{x_k^n\} \in \mathcal{S}$ can be viewed as *Gelfand-Tsetlin schemes*. Then $\mathcal{M}(t)$ is the “Fourier transform” of a suitable irreducible character of the infinite-dimensional unitary group $U(\infty)$, see [16]. Interestingly enough, increasing t corresponds to a *deterministic* flow on the space of irreducible characters of $U(\infty)$.
6. Elements of \mathcal{S} can also be viewed as lozenge tiling of a sector in the plane. To see that one surrounds each particle location by a rhombus of one type and draws edges through locations where there are no particles, see Figure 2.2. Our initial condition corresponds to a perfectly regular tiling, see Figure 2.1.
7. The random tiling defined by $\mathcal{M}(t)$ is the limit of the uniformly distributed lozenge tilings of hexagons with side lengths (a, b, c) , when $a, b, c \rightarrow \infty$ so that $ab/c \rightarrow t$, and we observe the hexagon tiling at finite distances from the corner between sides of lengths a and b .
8. Finally, Figure 2.2 has a clear three-dimensional connotation. Given the random configuration $\{x_k^n(t)\} \in \mathcal{S}$ at time moment t , define the random *height function*

$$\begin{aligned} h : (\mathbb{Z} + \tfrac{1}{2}) \times \mathbb{Z}_{>0} \times \mathbb{R}_{\geq 0} &\rightarrow \mathbb{Z}_{\geq 0}, \\ h(x, n, t) &= \#\{k \in \{1, \dots, n\} \mid x_k^n(t) > x\}. \end{aligned} \tag{2.2}$$

In terms of the tiling on Figure 2.2, the height function is defined at the vertices of rhombi, and it counts the number of particles to the right from a given vertex. (This definition differs by a simple linear function of (x, n) from the standard definition of the height function for lozenge tilings, see e.g. [26, 27].) The initial condition corresponds to starting with perfectly flat facets.

Thus, our Markov chain can be viewed as a random growth model of the surface given by the plot of the height function. In terms of the stepped surface of Figure 2.2, the evolution consists of removing all columns of (x, n, h) -dimensions $(1, *, 1)$ that could be removed, independently with exponential waiting times of rate one. For example, if x_2^2 jumps to its right, then three consecutive cubes (associated to x_2^2, x_3^3, x_4^4) are removed.

Clearly, in this dynamics the directions x and n do not play symmetric roles. Indeed, this model belongs to the $2 + 1$ anisotropic KPZ class of stochastic growth models, see Section 2.4.

2.1. Determinantal formula. The first nontrivial result about the Markov chain $\mathcal{M}(t)$ is the (partial) determinantal structure of the correlation functions.

Theorem 2.1 (Theorem 1.1 of [7]). *For any $N = 1, 2, \dots$, pick N triples*

$$\varkappa_j = (x_j, n_j, t_j) \in \mathbb{Z} \times \mathbb{Z}_{>0} \times \mathbb{R}_{\geq 0}$$

such that $t_1 \leq t_2 \leq \dots \leq t_N$, $n_1 \geq n_2 \geq \dots \geq n_N$. Then

$$\mathbb{P}\{\text{For each } j = 1, \dots, N \text{ there exists a } k_j, \\ 1 \leq k_j \leq n_j \text{ such that } x_{k_j}^{n_j}(t_j) = x_j\} = \det [\mathcal{K}(\varkappa_i, \varkappa_j)]_{i,j=1}^N, \quad (2.3)$$

where

$$\mathcal{K}(x_1, n_1, t_1; x_2, n_2, t_2) = -\frac{1}{2\pi i} \oint_{\Gamma_0} \frac{dw}{w^{x_2-x_1+1}} \frac{e^{(t_1-t_2)/w}}{(1-w)^{n_2-n_1}} \mathbb{1}_{[(n_1, t_1) \prec (n_2, t_2)]} \\ + \frac{1}{(2\pi i)^2} \oint_{\Gamma_0} dw \oint_{\Gamma_1} dz \frac{e^{t_1/w}}{e^{t_2/z}} \frac{(1-w)^{n_1}}{(1-z)^{n_2}} \frac{w^{x_1}}{z^{x_2+1}} \frac{1}{w-z}, \quad (2.4)$$

the contours Γ_0, Γ_1 are simple positively oriented closed paths that include the poles 0 and 1, respectively, and no other poles (hence, they are disjoint).

The determinantal structure makes it possible to study the asymptotics.

2.2. Macroscopic scale, one-point fluctuations, and local structure. In the large time limit, under hydrodynamic scaling our model has a limit shape which we now describe, see Figure 2.3.

Since we consider heights at different times, we cannot use time as a large parameter. Instead, we introduce a large parameter L and consider space and time coordinates that are comparable to L . The limit shape consists of three facets interpolated by a curved piece. To describe it, consider the set

$$\mathcal{D} = \{(\nu, \eta, \tau) \in \mathbb{R}_{>0}^3 \mid (\sqrt{\eta} - \sqrt{\tau})^2 < \nu < (\sqrt{\eta} + \sqrt{\tau})^2\}. \quad (2.5)$$

It is exactly the set of triples $(\nu, \eta, \tau) \in \mathbb{R}_{>0}^3$ for which there exists a non-degenerate triangle with side lengths $(\sqrt{\nu}, \sqrt{\eta}, \sqrt{\tau})$. Denote by $(\pi_\nu, \pi_\eta, \pi_\tau)$ the angles of this triangle that are opposite to the corresponding sides.

The following result concerns the limit shape and the Gaussian fluctuations in the curved region, living on a $\sqrt{\ln L}$ scale.

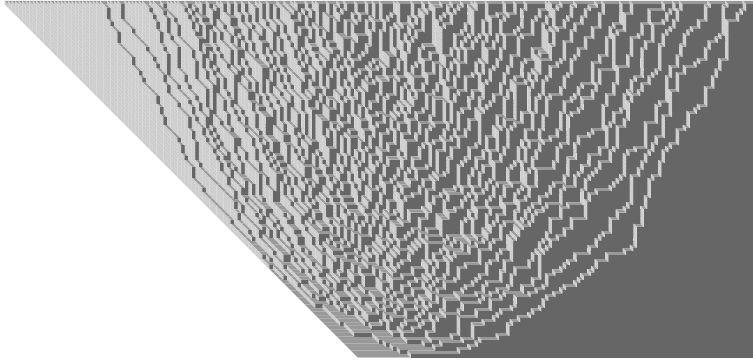


Figure 2.3. A configuration of the model with 100 levels ($m = 1, \dots, 100$) at time $t = 25$, using the same representation as in Figure 2.2.

Theorem 2.2 (Theorem 1.2 of [7]). *For any $(\nu, \eta, \tau) \in \mathcal{D}$ we have the moment convergence of random variables*

$$\lim_{L \rightarrow \infty} \frac{h([\nu - \eta]L + \frac{1}{2}, [\eta]L, \tau L) - \mathbb{E} h([\nu - \eta]L + \frac{1}{2}, [\eta]L, \tau L)}{\sqrt{\ln L / (2\pi^2)}} = \mathcal{N}(0, 1). \tag{2.6}$$

One also has an explicit formula for the limit shape:

$$\begin{aligned} \lim_{L \rightarrow \infty} \frac{\mathbb{E} h([\nu - \eta]L + \frac{1}{2}, [\eta]L, \tau L)}{L} &=: \mathbf{h}(\nu, \eta, \tau) \\ &= \frac{1}{\pi} \left(-\nu\pi_\eta + \eta(\pi - \pi_\nu) + \tau \frac{\sin \pi_\nu \sin \pi_\eta}{\sin \pi_\tau} \right). \end{aligned} \tag{2.7}$$

Theorem 2.2 describes the limit shape \mathbf{h} of our growing surface, and the domain \mathcal{D} describes the points where this limit shape is *curved*. The logarithmic fluctuations are essentially a consequence of the local asymptotic behavior being governed by the discrete sine kernel (this local behavior occurs also in tiling models [21,24,32]). Using the connection with the Charlier ensembles, see above, the formula (2.7) for the limit shape can be read off the formulas of [2].

Using Theorem 2.1 it is not hard to verify that near every point of the limit shape in the curved region, at any fixed time moment the random lozenge tiling approaches the unique translation invariant measure $M_{\pi_\nu, \pi_\eta, \pi_\tau}$ on lozenge tilings of the plane with prescribed slope (see [18,27,29] and references therein for discussions of these measures). The slope is exactly the slope of the tangent plane to the limit shape. This implies in particular, that $(\pi_\nu/\pi, \pi_\eta/\pi, \pi_\tau/\pi)$ are the asymptotic proportions of lozenges of three different types in the neighborhood of the point of the limit shape.

One also computes the growth velocity (see (2.9) for the definition of Ω)

$$\frac{\partial \mathbf{h}}{\partial \tau} = \frac{1}{\pi} \frac{\sin \pi_\nu \sin \pi_\eta}{\sin \pi_\tau} = \frac{\text{Im}(\Omega(\nu, \eta, \tau))}{\pi}. \tag{2.8}$$

Since the right-hand side depends only on the slope of the tangent plane, this suggests that it should be possible to extend the definition of our surface evolution to the random surfaces distributed according to measures $M_{\pi_\nu, \pi_\eta, \pi_\tau}$; these measures have to remain invariant under evolution, and the speed of the height growth should be given by the right-hand side of (2.8).

2.3. Complex structure and multi-point fluctuations. The Gaussian Free Field. To describe the correlations of the fluctuations of our random surface, we first need to introduce a complex structure on the limit shape. Set $\mathbb{H} = \{z \in \mathbb{C} \mid \text{Im}(z) > 0\}$ and define the map $\Omega : \mathcal{D} \rightarrow \mathbb{H}$ by

$$|\Omega(\nu, \eta, \tau)| = \sqrt{\eta/\tau}, \quad |1 - \Omega(\nu, \eta, \tau)| = \sqrt{\nu/\tau}. \tag{2.9}$$

Observe that $\arg \Omega = \pi_\nu$ and $\arg(1 - \Omega) = -\pi_\eta$. The preimage of any $\Omega \in \mathbb{H}$ is a ray in \mathcal{D} that consists of triples (ν, η, τ) with constant ratios $(\nu : \eta : \tau)$. Denote this ray by R_Ω . One sees that R_Ω 's are also the level sets of the slope of the tangent plane to the limit shape. Since $\mathbf{h}(\alpha\nu, \alpha\eta, \alpha\tau) = \alpha\mathbf{h}(\nu, \eta, \tau)$ for any $\alpha > 0$, the height function grows linearly in time along each R_Ω . Note also that the map Ω satisfies

$$(1 - \Omega) \frac{\partial \Omega}{\partial \nu} = \Omega \frac{\partial \Omega}{\partial \eta} = -\frac{\partial \Omega}{\partial \tau}, \tag{2.10}$$

and the first of these relations is the complex Burgers equation, cf. [28].

From Theorem 2.2 one might think that to get non-trivial correlations one needs to consider $(h - \mathbb{E}(h))/\sqrt{\ln L}$. However, this is not true and the division by $\sqrt{\ln L}$ is not needed. To state the precise result, denote by

$$\mathcal{G}(z, w) = -\frac{1}{2\pi} \ln \left| \frac{z - w}{z - \bar{w}} \right| \tag{2.11}$$

the Green function of the Laplace operator on \mathbb{H} with Dirichlet boundary conditions.

Theorem 2.3 (Theorem 1.3 of [7]). *For any $N = 1, 2, \dots$, let $\varkappa_j = (\nu_j, \eta_j, \tau_j) \in \mathcal{D}$ be any distinct N triples such that*

$$\tau_1 \leq \tau_2 \leq \dots \leq \tau_N, \quad \eta_1 \geq \eta_2 \geq \dots \geq \eta_N. \tag{2.12}$$

Denote

$$H_L(\nu, \eta, \tau) := \sqrt{\pi} \left(h([\nu - \eta]L) + \frac{1}{2}, [\eta L], \tau L \right) - \mathbb{E} h([\nu - \eta]L) + \frac{1}{2}, [\eta L], \tau L), \tag{2.13}$$

and $\Omega_j = \Omega(\nu_j, \eta_j, \tau_j)$. Then

$$\lim_{L \rightarrow \infty} \mathbb{E} (H_L(\mathbf{z}_1) \cdots H_L(\mathbf{z}_N)) = \begin{cases} \sum_{\sigma \in \mathcal{F}_N} \prod_{j=1}^{N/2} \mathcal{G}(\Omega_{\sigma(2j-1)}, \Omega_{\sigma(2j)}), & N \text{ is even,} \\ 0, & N \text{ is odd,} \end{cases} \tag{2.14}$$

where the summation is taken over all fixed point free involutions σ on $\{1, \dots, N\}$.

The result of the theorem means that as $L \rightarrow \infty$, $H_L(\Omega^{-1}(z))$ is a Gaussian process with covariance given by \mathcal{G} , i.e., it has correlation of the Gaussian Free Field (GFF) on \mathbb{H} . A few additional estimates allow one to prove that the fluctuations indeed converge to GFF, see Section 5.5 of [7].

Conjecture 2.4. *The statement of Theorem 2.3 holds without the assumption (2.12), provided that Ω -images of all the triples are pairwise distinct.*

Theorem 2.3 and Conjecture 2.4 indicate that the fluctuations of the height function along the rays R_Ω vary slower than in any other space-time direction. This statement can be rephrased more generally: the height function has slower fluctuations along the curves where the slope of the limit shape remains constant. Or even more generally: the height function fluctuates slower along the characteristics of the first order PDE (2.8) that governs the macroscopic growth of the interface. This *slow decorrelation* phenomenon has been established for several (1+1)-dimensional growth models in [20], [19].

2.4. Universality class. In the terminology of physics literature, see e.g. [1], our Markov chain falls into the class of local growth models with relaxation and lateral growth, that are believed to be described, on the fluctuation level, by the Kardar-Parisi-Zhang (KPZ) equation

$$\partial_t h = \Delta h + Q(\partial_x h, \partial_y h) + \text{space-time white noise}, \tag{2.15}$$

where Q is a quadratic form. For our model, one easily computes that the determinant of the Hessian of $\partial_t h$, viewed as a function of the slope, is strictly negative, which means that the form Q in our case has signature $(-1, 1)$. In such a situation the equation (2.15) is called *anisotropic KPZ* or *AKPZ* equation.

Using non-rigorous renormalization group analysis based on one-loop expansion, Wolf [35] predicted that large time fluctuations (the roughness) of the growth models described by AKPZ equation should be similar to those of linear models described by the Edwards-Wilkinson equation (heat equation with random term): $\partial_t h = \Delta h + \text{white noise}$.

The above results can be viewed as the first rigorous analysis of a non-equilibrium growth model in the AKPZ class. Indeed, Wolf’s prediction correctly identifies the logarithmic behavior of height fluctuations. However, it

does not (at least explicitly) predict the appearance of the Gaussian Free Field, and in particular the complete structure (map Ω) of the fluctuations described above.

On the other hand, universality considerations imply that analogs of Theorems 2.2 and 2.3, as well as possibly Conjecture 2.4, should hold in any AKPZ growth model.

3. More General Random Growth Models

It turns out that the growth model from the previous section can be substantially generalized from a variety of viewpoints.

Typically, these more general models would still lead to the partial determinantal structure of the correlation functions, as in Section 2.1 above. However, asymptotic analysis becomes more difficult.

On all three levels — macroscopically (hydrodynamic scale), microscopically (lattice scale), and mesoscopically (fluctuation scale) — new phenomena may appear. Below we describe a few general sources of new models together with some of the new effects.

3.1. More general update rules. Here is a list of “simple” twists that one could impose on the growth model discussed above. Note that in all the situations, the interlacing condition is being preserved by the same “block-push” mechanism as above.

1. Clearly, instead of making particles jump to the right, we could let them jump to the left — there is an immediate symmetry that interchanges the two directions. However, why not let particles jump both to the left and to the right, with independent exponential clocks governing jumps in different directions?
2. One can imagine that different particles might jump with different rates. For example, one could assume that the jump rate of x_k^m is a function of m .
3. Instead of using the continuous time, one could consider discrete time dynamics with each particle attempting a Bernoulli jump, or a geometrically distributed jump, independently of the others.

As was shown in [7], neither of these modifications (nor any temporal sequence of them) destroys the determinantal structure of the correlation functions, and this structure paves the way for the asymptotic analysis. In [5] such analysis led to a discovery of new type of critical behavior that may arise at a tacnode singularity of the frozen boundary.

3.2. Special functions and tiling models. It was mentioned in the previous section that Charlier classical orthogonal polynomials¹ can be used to analyze measures $\mathcal{M}(t)$. The Charlier polynomials lie at the bottom of the so-called Askey scheme — the hierarchy of classical orthogonal polynomials of hypergeometric type. Thus, it is natural to ask if more general hypergeometric orthogonal polynomials correspond to meaningful growth models/interlacing particle systems.

One answer to this question was given in [14, 15]. Let us describe the static (one-time) distributions that arise.

For any integers $a, b, c \geq 1$ consider a hexagon with sides a, b, c, a, b, c drawn on the regular triangular lattice. Denote by $\Omega_{a \times b \times c}$ the set of all tilings of this hexagon by rhombi obtained by gluing two of the neighboring elementary triangles together (such rhombi are called *lozenges*). Lozenge tilings of a hexagon can be identified with 3D Young diagrams (equivalently, boxed plane partitions) or with stepped surfaces.

We consider probability distributions on $\Omega_{a \times b \times c}$ of the following form. Fix one of the three lozenge types, and to any $\mathcal{T} \in \Omega_{a \times b \times c}$ assign the weight equal to the product of weights of lozenges of the chosen type in \mathcal{T} . The weight of one lozenge, in its turn, is equal to $\zeta q^j - 1/(\zeta q^j)$, where ζ and q are (generally speaking, complex) parameters, and j is a linear function on the plane that is constant along the long diagonals of the lozenges of chosen type. There are three different ways to restrict ζ and q to suitable regions of the complex plane to make the resulting measure on $\Omega_{a \times b \times c}$ positive. Degenerations include the uniform measure on $\Omega_{a \times b \times c}$ and the measure with weights q^{Vol} , where Vol is the volume of the corresponding plane partition.

In [14, 15], we constructed discrete time Markov chains, quite similar to the growth model described in the previous section, that map measures on $\Omega_{a \times b \times c}$ from this class to similar ones on $\Omega_{a \times (b \pm 1) \times (c \mp 1)}$. Since $\Omega_{a \times b \times 0}$ is a singleton, this provided, in particular, an exact sampling algorithm for these measures. Its computer implementation can be found at <http://www.math.caltech.edu/papers/Borodin-Gorin-Rains.exe>. A gallery of pictures can be found in Section 9 of [15].

The key property that allowed us to access these measures was that they are closely related to the q -Racah hypergeometric orthogonal polynomials (that lie at the top of the q -Askey scheme). We were able to compute the limit shapes in certain limit regimes and prove the microscopic convergence of the measures to the ergodic translation invariant Gibbs measures on lozenge tilings of the plane (they were already mentioned in Section 2.2).

3.3. Symmetric functions and skew plane partitions. The story about $\Omega_{a \times b \times c}$ suggests that it may be possible to do similar things to lozenge tilings with more complex boundary conditions. This is indeed the

¹The orthogonality set is $\{0, 1, 2, \dots\}$ and the weight function is $w(x) = t^x/x!$

case, and in order to describe a result it is more convenient to use the language of plane partitions.

Fix two natural numbers A and B . For a Young diagram $\pi \subset B^A$, set $\bar{\pi} = B^A/\pi$. A (skew) plane partition Π with support $\bar{\pi}$ is a filling of all boxes of $\bar{\pi}$ by nonnegative integers $\Pi_{i,j}$ (we assume that $\Pi_{i,j}$ is located in the i th row and j th column of B^A) such that $\Pi_{i,j} \geq \Pi_{i,j+1}$ and $\Pi_{i,j} \geq \Pi_{i+1,j}$ for all values of i, j . The volume of the plane partition Π is defined as

$$\text{Vol}(\Pi) = \sum_{i,j} \Pi_{i,j}.$$

In [4] we constructed Markov chains (similar to the one described in Section 2) that incrementally grow the support of skew plane partitions and map measures q^{Vol} to similar ones. This produces, in particular, an exact sampling algorithm for q^{Vol} -distributed skew plane partitions with an arbitrary back wall π .

As was shown in [32, 33], q^{Vol} -distributed skew plane partitions form a special case of a much more general class of probability measures on sequences of partitions called the *Schur processes*. As the name suggests, they are defined in terms of Schur symmetric functions. The construction of [4] applies to the general Schur processes as well and has at least one other application, cf. Section 3.5 below.

3.4. Random growth in 1+1 dimensions. As was mentioned above, the restriction of the 2+1 dimensional growth model of the previous section to the row of left-most particles yields the Totally Asymmetric Simple Exclusion Process (TASEP). This is one of the basic models of the 1-dimensional growth that has been extensively studied.

It is known that on the macroscopic scale, the particle density of the TASEP evolves deterministically according to the nonviscid Burgers equation. Therefore, one natural question is to study the fluctuation properties which show rather interesting features.

For simplicity, let us restrict ourselves to deterministic initial conditions. The densely packed initial condition for the 2+1 dimensional model of the previous section induces the so-called step initial condition for TASEP, when the m th particle starts off $-m$, $m \geq 1$. Historically, this was the first initial condition for which the fluctuations were understood. Johansson [23] showed that asymptotic fluctuations of the position of a given particle are governed by the so-called GUE (Gaussian Unitary Ensemble) Tracy-Widom distribution F_2^2 . This is the asymptotic distribution of the largest eigenvalue of the Gaussian Hermitian random matrices in the limit of large matrix dimension. In a later work [25], he showed the convergence of multi-particle fluctuations to the

²The correct scaling of the fluctuations had been predicted more than 10 years earlier by physicists via the KPZ equation.

so-called Airy_2 process, that also arises from the evolution of GUE random matrices via the Dyson Brownian motion.

While for a few years this result remained very surprising, from the (2+1)-dimensional point of view the appearance of random matrices is very clear: Particles $\{x_k^m\}_{k=1}^m$ with fixed m form a random matrix type object – the m -point orthogonal polynomial ensemble with the Charlier weight. In the large time limit, the Charlier polynomials converge to the Hermite polynomials, and the distribution of $\{x_k^m\}_{k=1}^m$ converges to that for the eigenvalues of the m -point GUE.

In a series of papers [8–13], we have used the (2+1)-dimensional perspective to analyze the fluctuation structure in a variety of situations. In particular,

1. The so-called flat initial condition for TASEP, long range TASEP, and polynuclear growth model (PNG) in (1+1)-dimensions. An example of the flat initial condition for TASEP is when initially every second site (or every third site) is occupied by a particle. The one-particle asymptotic fluctuations are given by GOE Tracy-Widom distribution F_1 , multi-particle fluctuations are given by the Airy_1 process.
2. Half-flat initial condition for TASEP, when every second site on the negative semiaxis is occupied, all other sites are free. A nontrivial transition from Airy_1 to Airy_2 processes occurs.
3. Half-flat initial condition with a slow first particle induces a shock — macroscopic discontinuity of the particle density. The fluctuation picture near the shock was obtained. When the speed of the first particle is equal to speed of the flow in the completely flat case, one suddenly obtains a matrix analog of P. Lévy's theorem on the maximum of the Brownian motion [10].

Despite obvious successes, there is still a lot to be understood. Physical universality arguments predict that Airy_1 and Airy_2 processes should arise for virtually any deterministic initial conditions. At this moment we cannot even show that for a periodic initial condition of the form $\dots \text{xx000xx000xx000}\dots$. A further investigation of the shock phenomenon is under way. What we have considered is the $t^{1/3} - t^{1/2}$ shock — the fluctuations on the left of shock have magnitude $t^{1/3}$, while on the right the fluctuations scale as $t^{1/2}$. It is very interesting to look into $t^{1/3} - t^{1/3}$ shocks as nothing is known about them. In particular, it is unclear if the shocks persist on the mesoscopic level.

In a different direction, even though the (2+1)-dimensional perspective turned out to be very useful for TASEP (Sasamoto [34] has to be credited for the idea), for many TASEP initial conditions (e.g., flat ones) the measure on the corresponding system of interlacing particles is not positive. In particular, the relation of TASEP to GOE (not GUE!) random matrices remains highly mysterious, especially since numerical experiments indicate that Airy_1 process does not arise from the GOE via the Dyson Brownian motion [3].

3.5. Representation Theory. As was mentioned above, the measures $\mathcal{M}(t)$ from the previous section can be viewed as Fourier transforms of certain characters (called *Plancherel* characters) of the infinite-dimensional unitary group $U(\infty)$. It would be natural to try to extend the construction to other families of characters of $U(\infty)$, as well as to inductive limits of other classical Lie groups (or, more generally, Riemannian symmetric spaces of classical type).

In the first direction, in [4] we have employed the general construction for the Schur processes mentioned in Section 3.3 to produce Markov chains on infinite Gelfand-Tsetlin schemes that represent deterministic flows on the space of extreme characters of the infinite-dimensional unitary group.

In the second direction, in [17] we considered the case of the Plancherel characters of the infinite-dimensional orthogonal group $O(\infty)$. Similarly to $U(\infty)$ case, one also obtains Markov dynamics on an interlacing particle system that can be viewed as a stepped surface or a lozenge tiling of a suitable domain. The difference is that this system has a *reflecting wall*; a Java simulation can be found at <http://www.math.caltech.edu/papers/OrthPlanch.html>.

We proved the determinantal structure of the correlation functions, computed the limit shape, and analyzed the local asymptotic behavior near the wall. This led to three different determinantal processes; one arises through antisymmetric Gaussian Hermitian random matrices [22], while two others are new. All three are likely to be universal (i.e. arising from many models with suitable symmetries).

A much more challenging problem is to do similar things for all the Riemannian symmetric spaces of BC-type in one stroke, cf. [31]. This will likely lead to a particle system with a generic reflection/absorption condition at the wall.

In yet another direction, representation theoretic background naturally leads to a very intriguing set of questions. What we have been doing so far lies in the realm of “commutative” probability. In other words, the representations have been restricted to a maximal commutative subalgebra of a suitable group algebra, and then the asymptotics has been studied. It is entirely possible that one can do similar things with a larger (non-commutative) subalgebra and it is very intriguing to see how much deeper we will be able to go.

References

- [1] A.L. Barabási and H.E. Stanley, *Fractal concepts in surface growth*, Cambridge University Press, Cambridge, 1995.
- [2] P. Biane, *Approximate factorization and concentration for characters of symmetric groups*, Int. Math. Res. Not. **4** (2001), 179–192.
- [3] F. Bornemann, P. L. Ferrari, and M. Prähofer, *The Airy₁ process is not the limit of the largest eigenvalue in GOE matrix diffusion*, J. Stat. Phys. **133** (2008), 405–415.
- [4] A. Borodin, *Schur dynamics of the Schur processes*, arXiv:1001.3442.

-
- [5] A. Borodin and M. Duits, *Limits of determinantal processes near a tacnode*, arXiv:0911.1980.
- [6] A. Borodin and P.L. Ferrari, *Large time asymptotics of growth models on space-like paths I: PushASEP*, Electron. J. Probab. **13** (2008), 1380-1418.
- [7] A. Borodin and P. L. Ferrari, *Anisotropic growth of random surfaces in $2 + 1$ dimensions*, arXiv:0804.3035.
- [8] A. Borodin, P.L. Ferrari, and M. Prähofer, *Fluctuations in the discrete TASEP with periodic initial configurations and the Airy_1 process*, Int. Math. Res. Papers **2007** (2007), rpm002.
- [9] A. Borodin, P.L. Ferrari, M. Prähofer, and T. Sasamoto, *Fluctuation properties of the TASEP with periodic initial configuration*, J. Stat. Phys. **129** (2007), 1055–1080.
- [10] A. Borodin, P.L. Ferrari, M. Prähofer, T. Sasamoto, and J. Warren, *Maximum of Dyson Brownian motion and non-colliding systems with a boundary*, to appear in Elect. Comm. Prob., arXiv:0905.3989.
- [11] A. Borodin, P.L. Ferrari, and T. Sasamoto, *Transition between Airy_1 and Airy_2 processes and TASEP fluctuations*, Comm. Pure Appl. Math. **61** (2008), 1603–1629.
- [12] A. Borodin, P.L. Ferrari, and T. Sasamoto, *Large time asymptotics of growth models on space-like paths II: PNG and parallel TASEP*, Comm. Math. Phys. **283** (2008), 417–449.
- [13] A. Borodin, P.L. Ferrari, and T. Sasamoto, *Two-speed TASEP*, Journal of Statistical Physics, **137** (2009), 936–977.
- [14] A. Borodin and V. Gorin, *Shuffling algorithm for boxed plane partitions*. Adv. Math. **220** (2009), 1739–1770.
- [15] A. Borodin, V. Gorin, and E. M. Rains, *q -Distributions on boxed plane partitions*, arXiv:0905.0679.
- [16] A. Borodin and J. Kuan, *Asymptotics of Plancherel measures for the infinite-dimensional unitary group*, Adv. Math. **219** (2008), 894–931.
- [17] A. Borodin and J. Kuan, *Random surface growth with a wall and Plancherel measures for $O(\infty)$* , to appear in Comm. Pure Appl. Math., arXiv:0904.2607.
- [18] A. Borodin and S. Shlosman, *Gibbs ensembles of nonintersecting paths*, Comm. Math. Phys. **293** (2010), 145–170.
- [19] I. Corwin, P. L. Ferrari, and S. Péché, *Universality of slow decorrelation in KPZ growth*, arXiv:1001.5345.
- [20] P.L. Ferrari, *Slow decorrelations in KPZ growth*, J. Stat. Mech. (2008), P07022.
- [21] P.L. Ferrari and H. Spohn, *Step fluctuations for a faceted crystal*, J. Stat. Phys. **113** (2003), 1–46.
- [22] P. J. Forrester and E. Nordenstam, *The Anti-Symmetric GUE Minor Process*, Mosc. Math. J. **9** (2009), 749–774.
- [23] K. Johansson, *Shape fluctuations and random matrices*, Comm. Math. Phys. **209** (2000), 437–476.

-
- [24] K. Johansson, *Non-intersecting paths, random tilings and random matrices*, Probab. Theory Related Fields **123** (2002), 225–280.
- [25] K. Johansson, *Discrete polynuclear growth and determinantal processes*, Comm. Math. Phys. **242** (2003), 277–329.
- [26] R. Kenyon, *Height fluctuations in the honeycomb dimer model*, Comm. Math. Phys. **281** (2008), 675–709.
- [27] R. Kenyon, *Lectures on dimers*, Statistical mechanics, 191–230, IAS/Park City Math. Ser. **16**, Amer. Math. Soc., Providence, RI, 2009.
- [28] R. Kenyon and A. Okounkov, *Limit shapes and the complex Burgers equation*, Acta Math. **199** (2007), 263–302.
- [29] R. Kenyon, A. Okounkov, and S. Sheffield, *Dimers and amoebae*, Ann. of Math. **163**, 1019–1056.
- [30] W. König, N. O’Connell, and S. Roch, *Non-colliding Random Walks, Tandem Queues and Discrete Orthogonal Polynomial Ensembles*, Electron. J. Probab. **7** (2002), 1–24.
- [31] A. Okounkov and G. Olshanski, *Limits of BC-type orthogonal polynomials as the number of variables goes to infinity*, Jack, Hall-Littlewood and Macdonald Polynomials (E. B. Kuznetsov and S. Sahi, eds). Amer. Math. Soc., Contemporary Math. vol. 417, 2006.
- [32] A. Okounkov and N. Reshetikhin, *Correlation function of Schur process with application to local geometry of a random 3-dimensional Young diagram*, J. Amer. Math. Soc. **16** (2003), 581–603.
- [33] A. Okounkov and N. Reshetikhin *Random skew plane partitions and the Pearcey process*, Comm. Math. Phys. **269** (2007). 571–609.
- [34] T. Sasamoto, *Spatial correlations of the 1D KPZ surface on a flat substrate*, J. Phys. A **38** (2005), L549–L556.
- [35] D.E. Wolf, *Kinetic roughening of vicinal surfaces*, Phys. Rev. Lett. **67** (1991), 1783–1786.

Patterned Random Matrices and Method of Moments

Arup Bose*, Rajat Subhra Hazra, and Koushik Saha

In memory of Ashok Prasad Maitra

Abstract

We present a unified approach to limiting spectral distribution (LSD) of patterned matrices via the moment method. We demonstrate relatively short proofs for the LSD of common matrices and provide insight into the nature of different LSD and their interrelations. The method is flexible enough to be applicable to matrices with appropriate dependent entries, banded matrices, and matrices of the form $A_p = \frac{1}{n}XX'$ where X is a $p \times n$ matrix with real entries and $p \rightarrow \infty$ with $n = n(p) \rightarrow \infty$ and $p/n \rightarrow y$ with $0 \leq y < \infty$.

This approach raises interesting questions about the class of patterns for which LSD exists and the nature of the possible limits. In many cases the LSD are not known in any explicit forms and so deriving probabilistic properties of the limit are also interesting issues.

Mathematics Subject Classification (2010). Primary 60B20; Secondary 60F05, 62E20, 60G57, 60B10.

Keywords. Moment method, large dimensional random matrix, eigenvalues, empirical and limiting spectral distributions, Wigner, Toeplitz, Hankel, circulant, reverse circulant, symmetric circulant, sample covariance and XX' matrices, band matrix, balanced matrix, linear dependence.

1. Introduction

Consider a sequence of patterned matrices with random entries. Examples include Wigner, sample variance covariance, Toeplitz and Hankel matrices. Finding the asymptotic properties of the spectrum as the dimension increases has been a major focus of research. We concentrate on such real symmetric matrices

*Research supported by J.C. Bose National Fellowship, DST, Govt. of India.
Stat.-Math. Unit, Indian Statistical Institute, 203 B.T Road, Kolkata 700108.
E-mails: bosearu@gmail.com, rajat_r@isical.ac.in, koushik_r@isical.ac.in.

and provide an overview of a unified moment approach in deriving their limiting spectral distribution (LSD). After developing a unified framework, we present selective sketches of proofs for a few of these matrices. We also discuss extensions to situations where the entries come from a dependent sequence or the matrix is of the form XX' , thus generalizing the sample variance covariance matrix. Finally we discuss in brief a few other matrices as well as methods for deriving the LSD.

2. Moment Method

Suppose $\{Y_n\}$ is a sequence of real valued random variables. Suppose that there exists some (nonrandom) sequence $\{\beta_h\}$ such that $E[Y_n^h] \rightarrow \beta_h$ for every positive integer h where $\{\beta_h\}$ satisfies *Carleman's condition*:

$$\sum_{h=1}^{\infty} \beta_{2h}^{-1/2h} = \infty.$$

Then there exists a distribution function F such that for all h

$$\beta_h = \int x^h dF(x) \text{ and } Y_n \text{ converges to } F \text{ in distribution.}$$

As an illustration suppose $\{x_i\}$ are i.i.d. random variables with mean zero and variance one and all moments finite. Let $Y_n = n^{-1/2}(x_1 + x_2 + \dots + x_n)$. By using binomial expansion and taking term by term expectation and then using elementary order calculations, $E[Y_n^{2k+1}] \rightarrow 0$ and $E[Y_n^{2k}] \rightarrow \frac{2k!}{2^k k!}$. Using Stirling's approximation it can be easily checked that $\{\beta_{2k} = \frac{2k!}{2^k k!}\}$ satisfies Carleman's condition. Since β_{2k} is the $2k$ -th moment of the standard Normal distribution, $Y_n \xrightarrow{D} N(0, 1)$.

This idea has traditionally been used for establishing the LSD for example of the Wigner and the sample variance covariance matrices. There the trace formula replaces the binomial expansion. However, the calculation/estimation of the leading term and the bounding of the lower order terms lead to combinatorial issues which usually have been addressed on a case by case basis.

3. Limiting Spectral Distribution and Moments

For any random $n \times n$ matrix A_n , if $\lambda_1, \lambda_2, \dots, \lambda_n$ are all its eigenvalues, then its *empirical spectral distribution function (ESD)* is given by

$$F^{A_n}(x, y) = n^{-1} \sum_{i=1}^n I\{\operatorname{Re}\lambda_i \leq x, \operatorname{Im}\lambda_i \leq y\}.$$

The *expected spectral distribution function* of A_n is defined as $E[F^{A_n}(\cdot)]$. The limiting spectral distribution (LSD) of a sequence $\{A_n\}$, as $n \rightarrow \infty$, is the

weak limit of the sequence $\{F^{A_n}\}$ if it exists, either almost surely (a.s.) or in probability. We shall deal with only real symmetric matrices and hence all eigenvalues are real. The h -th moment of the ESD of A_n has the following nice form:

$$h\text{-th moment of the ESD of } A_n = \frac{1}{n} \sum_{i=1}^n \lambda_i^h = \frac{1}{n} \text{tr}(A_n^h) = \beta_h(A_n) \text{ (say).}$$

The following easy Lemma links convergence of moments and LSD. Consider the following conditions:

(M1) For every $h \geq 1$, $E[\beta_h(A_n)] \rightarrow \beta_h$ and $\{\beta_h\}$ satisfies Carleman’s condition.

(M2) $\text{Var}[\beta_h(A_n)] \rightarrow 0$ for every $h \geq 1$.

(M4) $\sum_{n=1}^\infty E[\beta_h(A_n) - E(\beta_h(A_n))]^4 < \infty$ for every $h \geq 1$.

Lemma 1. *If (M1) and (M2) hold then $\{F^{A_n}\}$ converges in probability to F determined by $\{\beta_h\}$. If further (M4) holds then this convergence is a.s.*

4. A Unified Approach

The sequence of variables which is used to construct the matrix will be called the **input sequence**. It shall be of the form $\{x_i; i \geq 0\}$ or $\{x_{ij}; i, j \geq 1\}$.

4.1. Link function. Let \mathbb{Z} be the set of all integers and let \mathbb{Z}_+ denote the set of all nonnegative integers. Let $L_n : \{1, 2, \dots, n\}^2 \rightarrow \mathbb{Z}^d, n \geq 1$ be a sequence of functions such that $L_{n+1}(i, j) = L_n(i, j)$ whenever $1 \leq i, j \leq n$. We shall write $L_n = L$ and call it the **link** function and by abuse of notation we write \mathbb{Z}_+^2 as the common domain of $\{L_n\}$. The matrices we consider will be of the form $((x_{L(i,j)}))$. Here are some well known matrices and their link functions:

(i) Wigner matrix $W_n^{(s)}$. $L : \mathbb{Z}_+^2 \rightarrow \mathbb{Z}^2$ where $L(i, j) = (\min(i, j), \max(i, j))$.

$$W_n^{(s)} = \begin{bmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1(n-1)} & x_{1n} \\ x_{12} & x_{22} & x_{23} & \dots & x_{2(n-1)} & x_{2n} \\ & & & \vdots & & \\ x_{1n} & x_{2n} & x_{3n} & \dots & x_{(n-1)n} & x_{nn} \end{bmatrix}.$$

(ii) Symmetric Toeplitz matrix $T_n^{(s)}$. $L : \mathbb{Z}_+^2 \rightarrow \mathbb{Z}$ where $L(i, j) = |i - j|$.

$$T_n^{(s)} = \begin{bmatrix} x_0 & x_1 & x_2 & \dots & x_{n-2} & x_{n-1} \\ x_1 & x_0 & x_1 & \dots & x_{n-3} & x_{n-2} \\ x_2 & x_1 & x_0 & \dots & x_{n-4} & x_{n-3} \\ & & & \vdots & & \\ x_{n-1} & x_{n-2} & x_{n-3} & \dots & x_1 & x_0 \end{bmatrix}.$$

(iii) Symmetric Hankel matrix $H_n^{(s)}$. $L : \mathbb{Z}_+^2 \rightarrow \mathbb{Z}$ where $L(i, j) = i + j$.

$$H_n^{(s)} = \begin{bmatrix} x_2 & x_3 & x_4 & \dots & x_n & x_{n+1} \\ x_3 & x_4 & x_5 & \dots & x_{n+1} & x_{n+2} \\ x_4 & x_5 & x_6 & \dots & x_{n+2} & x_{n+3} \\ & & & \vdots & & \\ x_{n+1} & x_{n+2} & x_{n+3} & \dots & x_{2n-1} & x_{2n} \end{bmatrix}.$$

(iv) Reverse Circulant $R_n^{(s)}$. $L : \mathbb{Z}_+^2 \rightarrow \mathbb{Z}$ where $L(i, j) = (i + j) \bmod n$.

$$R_n^{(s)} = \begin{bmatrix} x_2 & x_3 & x_4 & \dots & x_0 & x_1 \\ x_3 & x_4 & x_5 & \dots & x_1 & x_2 \\ x_4 & x_5 & x_6 & \dots & x_2 & x_3 \\ & & & \vdots & & \\ x_1 & x_2 & x_3 & \dots & x_{n-1} & x_0 \end{bmatrix}.$$

(v) Symmetric circulant $C_n^{(s)}$. $L : \mathbb{Z}_+^2 \rightarrow \mathbb{Z}$ where $L(i, j) = n/2 - |n/2 - |i - j||$.

$$C_n^{(s)} = \begin{bmatrix} x_0 & x_1 & x_2 & \dots & x_2 & x_1 \\ x_1 & x_0 & x_1 & \dots & x_3 & x_2 \\ x_2 & x_1 & x_0 & \dots & x_2 & x_3 \\ & & & \vdots & & \\ x_1 & x_2 & x_3 & \dots & x_1 & x_0 \end{bmatrix}.$$

(vi) Doubly symmetric Hankel matrix DH_n . The symmetric circulant is also a “doubly symmetric” Toeplitz matrix. The doubly symmetric Hankel matrix DH_n with link function $L(i, j) = n/2 - |n/2 - (i + j) \bmod n|$, $0 \leq i, j \leq n$ is

$$DH_n = \begin{bmatrix} x_0 & x_1 & x_2 & \dots & x_3 & x_2 & x_1 \\ x_1 & x_2 & x_3 & \dots & x_2 & x_1 & x_0 \\ x_2 & x_3 & x_4 & \dots & x_1 & x_0 & x_1 \\ & & & \vdots & & & \\ x_2 & x_1 & x_0 & \dots & x_5 & x_4 & x_3 \\ x_1 & x_0 & x_1 & \dots & x_4 & x_3 & x_2 \end{bmatrix}.$$

(vii) Palindromic matrices PT_n and PH_n . For these symmetric matrices the first row is a palindrome. PT_n is given below and PH_n is defined similarly.

$$PT_n = \begin{bmatrix} x_0 & x_1 & x_2 & \dots & x_2 & x_1 & x_0 \\ x_1 & x_0 & x_1 & \dots & x_3 & x_2 & x_1 \\ x_2 & x_1 & x_0 & \dots & x_4 & x_3 & x_2 \\ & & & \vdots & & & \\ x_1 & x_2 & x_3 & \dots & x_1 & x_0 & x_1 \\ x_0 & x_1 & x_2 & \dots & x_2 & x_1 & x_0 \end{bmatrix}.$$

(viii) Sample variance covariance matrix: often called the S matrix, is defined as

$$A_p(W) = n^{-1}W_pW'_p \text{ where } W_p = ((x_{ij}))_{1 \leq i \leq p, 1 \leq j \leq n}. \tag{1}$$

It is convenient in this case to think of the link function as a *pair*, given by:

$$L1, L2 : \mathbb{Z}_+^2 \times \mathbb{Z}_+^2 \rightarrow \mathbb{Z}^2, L1(i, j) = (i, j), L2(i, j) = (j, i).$$

(ix) Taking a cue from (viii) one may consider XX' where X is a suitable nonsymmetric matrix.

All the link functions above possess Property B given below with $f(x) = x$. Unless otherwise specified we shall assume that $f(x) = x$. The general form for f is needed to deal with matrices with dependent entries.

Property B: Let $f : \mathbb{Z}^d \rightarrow \mathbb{Z}$. Then (L, f) is said to satisfy Property B if

$$\Delta(L, f) = \sup_n \sup_{t \in \mathbb{Z}_+^d} \sup_{1 \leq k \leq n} \#\{l : 1 \leq l \leq n, f(|L(k, l) - t|) = 0\} < \infty. \tag{2}$$

4.2. Scaling. Assume that $\{x_i\}$ have mean zero and variance 1. Let F_n denote the ESD of $T_n^{(s)}$ and let X_n be the corresponding random variable. Then

$$E_{F_n}(X_n) = \frac{1}{n} \sum_{i=1}^n \lambda_{i,n} = \frac{1}{n} \text{Tr}(T_n^{(s)}) = x_0 \text{ and } E[E_{F_n}(X_n)] = 0,$$

$$\begin{aligned} E_{F_n}(X_n^2) &= \frac{1}{n} \sum_{i=1}^n \lambda_{i,n}^2 = \frac{1}{n} \text{Tr}(T_n^{(s)2}) \\ &= \frac{1}{n} [nx_0^2 + 2(n-1)x_1^2 + \dots + 2x_{n-1}^2] \text{ and } E[E_{F_n}(X_n^2)] = n. \end{aligned}$$

Hence it is appropriate to consider $n^{-1/2}T_n^{(s)}$. The same holds for all matrices except XX' , for which the issue is more complicated.

4.3. Trace formula and circuits. Let $A_n = ((a_{L(i,j)}))$. Then the h -th moment of $F^{n^{-1/2}A_n}$ is given by

$$\frac{1}{n} \text{Tr} \left(\frac{A_n}{\sqrt{n}} \right)^h = \frac{1}{n^{1+h/2}} \sum_{1 \leq i_1, i_2, \dots, i_h \leq n} a_{L(i_1, i_2)} a_{L(i_2, i_3)} \cdots a_{L(i_{h-1}, i_h)} a_{L(i_h, i_1)}. \tag{3}$$

Circuit: $\pi : \{0, 1, 2, \dots, h\} \rightarrow \{1, 2, \dots, n\}$ with $\pi(0) = \pi(h)$ is called a **circuit** of **length** $l(\pi) := h$. The dependence of a circuit on h and n will be suppressed. Clearly (M1), (M2) and (M4) may be written in terms of circuits.

For example,

$$(M1) \quad E[\beta_h(n^{-1/2}A_n)] = E\left[\frac{1}{n} \text{Tr} \left(\frac{A_n}{\sqrt{n}} \right)^h\right] = \frac{1}{n^{1+h/2}} \sum_{\pi: \pi \text{ circuit}} E X_\pi \rightarrow \beta_h$$

where

$$X_\pi = x_{L(\pi(0),\pi(1))} x_{L(\pi(1),\pi(2))} \cdots x_{L(\pi(h-2),\pi(h-1))} x_{L(\pi(h-1),\pi(h))}.$$

Matched Circuits: For any π , any $L(\pi(i-1), \pi(i))$ is an *L-value*. If an *L-value* is repeated exactly e times we say that it has an **edge of order** e ($1 \leq e \leq h$). If π has all $e \geq 2$ then it is called **L-matched** (in short *matched*). For any nonmatched π , $E[X_\pi] = 0$ and hence only matched π are relevant. If π has only order 2 edges then it is called **pair matched**.

To verify (M2) we need multiple circuits: k circuits $\pi_1, \pi_2, \dots, \pi_k$ are **jointly matched** if each *L-value* occurs at least twice across all circuits. They are **cross matched** if each circuit has at least one *L-value* which occurs in at least one of the other circuits.

To deal with dependent inputs we need the following: a π is **(L, f)-matched** if for each i , there is at least one $j \neq i$ such that $f(|L(\pi(i-1), \pi(i)) - L(\pi(j-1), \pi(j))|) = 0$. The earlier *L* matching is a special case with $f(x) = x$. The concepts of jointly matching and cross matching can be similarly extended.

Equivalence of circuits: The following defines an equivalence relation between the circuits: π_1 and π_2 are **equivalent** if and only if

$$\{L(\pi_1(i-1), \pi_1(i)) = L(\pi_1(j-1), \pi_1(j)) \Leftrightarrow L(\pi_2(i-1), \pi_2(i)) = L(\pi_2(j-1), \pi_2(j))\}.$$

4.4. Words. Any equivalence class induces a partition of $\{1, 2, \dots, h\}$. To any partition we associate a **word** w of length $l(w) = h$ of letters where the first occurrence of each letter is in alphabetical order. For example if $h = 5$ then the partition $\{\{1, 3, 5\}, \{2, 4\}\}$ is represented by the word *ababa*.

The class $\Pi(w)$: let $w[i]$ denote the i -th entry of w . The equivalence class corresponding to w will be denoted by

$$\Pi(w) = \{\pi : w[i] = w[j] \Leftrightarrow L(\pi(i-1), \pi(i)) = L(\pi(j-1), \pi(j))\}.$$

The number of partition blocks corresponding to w will be denoted by $|w|$. If $\pi \in \Pi(w)$, then clearly, $\#\{L(\pi(i-1), \pi(i)) : 1 \leq i \leq h\} = |w|$.

The above notions carry over to words. For instance *ababa* is matched. The word *abcadbaa* is nonmatched with edges of order 1, 2 and 4 and the corresponding partition is $\{\{1, 4, 7, 8\}, \{2, 6\}, \{3\}, \{5\}\}$.

For technical reasons it becomes easier to deal with a class bigger than Π . Let

$$\Pi^*(w) = \{\pi : w[i] = w[j] \Rightarrow L(\pi(i-1), \pi(i)) = L(\pi(j-1), \pi(j))\}.$$

4.5. Reduction to bounded case. We first show that in general, it is enough to work with input sequences which are uniformly bounded. The proof of the following lemma is available in Bose and Sen (2008)[21].

Assumption I $\{x_i, x_{ij}\}$ are independent and uniformly bounded with mean zero and variance 1.

Assumption II $\{x_i, x_{ij}\}$ are i.i.d. with mean zero and variance 1.

Let $\{A_n\}$ be a sequence of $n \times n$ random matrices with link function L_n . Let

$$k_n = \#\{L_n(i, j) : 1 \leq i, j \leq n\}, \quad \alpha_n = \max_k \#\{(i, j) : L_n(i, j) = k, 1 \leq i, j \leq n\}.$$

Lemma 2. *Suppose $k_n \rightarrow \infty$ and $k_n \alpha_n = O(n^2)$. If $\{F^{n^{-1/2}A_n}\}$ converges to a nonrandom F a.s. when the input sequence satisfies Assumption I. Then the same limit holds if it satisfies Assumption II.*

4.6. Only pair matched words contribute. From the discussion in Section 4.3 it is enough to consider matched circuits. The next lemma shows that we can further restrict attention to pair matched words. Its proof is easy and is available in Bose and Sen (2008)[21].

Let $N_{h,3+}$ be the number of (L, f) matched circuits on $\{1, 2, \dots, n\}$ of length h with at least one edge of order ≥ 3 .

Lemma 3. *(a) If (L, f) satisfies Property B then there is a constant C such that*

$$N_{h,3+} \leq Cn^{\lfloor (h+1)/2 \rfloor} \quad \text{and as } n \rightarrow \infty, \quad n^{-(1+h/2)}N_{h,3+} \rightarrow 0. \tag{4}$$

(b) Suppose $\{A_n\}$ is a sequence of $n \times n$ random matrices with input sequence $\{x_i\}$ satisfying Assumption I and (L, f) with $f(x) = x$ satisfying Property B. Then

$$\text{if } h \text{ is odd, } \lim_n \mathbb{E}[\beta_h(n^{-1/2}A_n)] = \lim_n \mathbb{E} \left[\frac{1}{n} \text{Tr} \left(\frac{A_n}{\sqrt{n}} \right)^h \right] = 0 \tag{5}$$

$$\text{and if } h = 2k \text{ then } \sum_{\substack{w \text{ has only} \\ \text{order 2 edges}}} \lim_n \frac{1}{n^{1+k}} |\Pi^*(w) - \Pi(w)| = 0 \tag{6}$$

and provided the limit in the right side below exists,

$$\lim_n \mathbb{E}[\beta_{2k}(n^{-1/2} A_n)] = \sum_{\substack{w \text{ has only} \\ \text{order } 2 \text{ edges}}} \lim_n \frac{1}{n^{1+k}} |\Pi(w)|. \tag{7}$$

Define, for each fixed matched word w of length $2k$ with $|w| = k$,

$$p(w) = \lim_n \frac{1}{n^{1+k}} |\Pi(w)| = \lim_n \frac{1}{n^{1+k}} |\Pi^*(w)| \tag{8}$$

whenever the limit exists. This limit will be positive and finite only if the number of elements in the set is of exact order n^{k+1} . In that case Lemma 3 implies that the limiting $(2k)$ -th moment is

$$\beta_{2k} = \sum_{w: |w|=k, l(w)=2k} p(w).$$

The next Lemma verifies (M4). Its proof is easy and is available in Bose and Sen (2008)[21]. Let $Q_{h,4}$ be the number of quadruples of circuits $(\pi_1, \pi_2, \pi_3, \pi_4)$ of length h which are jointly matched and cross matched with respect to (L, f) .

Lemma 4. (a) If (L, f) obeys Property B, $Q_{h,4} \leq Kn^{2h+2}$ for some constant K .

(b) If $\{A_n\}$ is a sequence of $n \times n$ random matrices with the input sequence $\{x_i\}$ satisfying Assumption I and (L, f) with $f(x) = x$ satisfying Property B then the following holds. As a consequence (M4) holds.

$$\mathbb{E} \left[\frac{1}{n} \text{Tr} \left(\frac{A_n}{\sqrt{n}} \right)^h - \mathbb{E} \frac{1}{n} \text{Tr} \left(\frac{A_n}{\sqrt{n}} \right)^h \right]^4 = O(n^{-2}). \tag{9}$$

4.7. Vertex, generating vertex and Carleman’s condition.

Any $\pi(i)$ is a **vertex**. It is **generating** if either $i = 0$ or $w[i]$ is the *first* occurrence of a letter. For example if $w = abbcab$ then $\pi(0), \pi(1), \pi(2), \pi(4)$ are generating. By Property B a circuit is completely determined, *up to finitely many choices*, by its generating vertices. The number of generating vertices is $|w| + 1$ and hence $|\Pi^*(w)| = O(n^{|w|+1})$. The following result is due to Bose and Sen (2008)[21].

Theorem 4.1. Let $\{A_n = ((x_{L(i,j)}))_{i,j=1}^n\}$ with the input sequence satisfying Assumption I and (L, f) satisfying Property B with $f(x) = x$. Then $\{F^{n^{-1/2} A_n}\}$ is tight a.s. Any subsequential limit G satisfies, for all nonnegative integers k , (i) $\beta_{2k+1}(G) = 0$ and (ii) $\beta_{2k}(G) \leq \frac{(2k)! \Delta(L, f)^k}{k! 2^k}$. Hence G is sub Gaussian. The (nonrandom) LSD exists for $\{n^{-1/2} A_n\}$ iff for every h , a.s.,

$$\lim \beta_h(n^{-1/2} A_n) = \beta_h \text{ (say)}. \tag{10}$$

In particular $\{\beta_h\}$ automatically satisfies Carleman’s condition.

5. The LSD for Some Specific Matrices

To derive any LSD it is enough to obtain (10) or (8). It turns out that for $C_n^{(s)}$, PT_n , PH_n and DH_n , $p(w) = 1$ for all w . For other matrices only certain words contribute in the limit. Properties of $p(w)$ for different matrices is given in Tables 1 and 2. Two special type of words which arise are the following:

Symmetric and Catalan words: A pair matched word is *symmetric* if each letter occurs once each in an odd and an even position. A pair matched word is *Catalan* if sequentially deleting all double letters leads to the empty word. For example *abccbdda* is a Catalan word whereas *abab* is not. The following result gives the count of these words. The proof of the first part of (a) is also available in Chapter 2 of Anderson, Guionnet and Zeitouni (2009)[3] and Bose and Sen (2008)[21].

Lemma 5. (a) *The number of Catalan words and symmetric words of length $2k$ are respectively, $\frac{(2k)!}{(k+1)!k!}$ and $k!$.*

(b) *Let $M_{t,k} = \#\{ \text{Catalan words of length } 2k \text{ with } (t + 1) \text{ even generating vertices and } (k - t) \text{ odd generating vertices} \}$. Then*

$$M_{t,k} = \binom{k-1}{t}^2 - \binom{k-1}{t+1} \binom{k-1}{t-1} = \frac{1}{t+1} \binom{k}{t} \binom{k-1}{t}.$$

Proof. (a) For any Catalan word mark the first and second occurrences of a letter by +1 and -1 respectively. For example *abba* and *abccbdda* are represented respectively by (1, 1, -1, -1) and (1, 1, 1, -1, -1, 1, -1, -1). This provides a bijection between the Catalan words of length $2k$ and sequences $\{u_l\}_{1 \leq l \leq 2k}$ satisfying: each $u_l = \pm 1$, $S_l = \sum_{j=1}^l u_j \geq 0 \forall l \geq 1$ and $S_{2k} = 0$. By reflection principle, the total number of such paths is easily seen to be $\frac{(2k)!}{(k+1)!k!}$. We omit the details. The proof of the second part is trivial.

(b) We know from part (a) that,

$$\#\{\text{Catalan words of length } 2k\} = \#\{\{u_l\}_{1 \leq l \leq 2k} : u_l = \pm 1, S_l \geq 0, S_{2k} = 0\}.$$

Note that the conditions $\{S_l \geq 0 \text{ and } S_{2k} = 0\}$ imply $u_1 = 1$ and $u_{2k} = -1$. Define

$$N_{e,1} := \#\{l : u_l = 1, l \text{ even}\} \text{ and } N_{o,-1} := \#\{l : u_l = -1, l \text{ odd}\}.$$

Clearly, $N_{e,1} \leq k - 1$ and $N_{o,-1} \leq k - 1$. Define

$$\begin{aligned} C_0 &= \{\{u_l\} : S_{2k-1} = 1, N_{e,1} = t, N_{o,-1} = t\}, \\ C_1 &= \{\{u_l\} : S_l < 0 \text{ for some } l, S_{2k-1} = 1, N_{e,1} = t, N_{o,-1} = t\}, \\ C_2 &= \{\{u_l\} : S_{2k-1} = -3, N_{e,1} = t - 1, N_{o,-1} = t + 1\}. \end{aligned}$$

Then it is easy to see that

$$\begin{aligned} \#C_0 &= \binom{k-1}{t}^2 \quad \text{and} \quad \#C_2 = \binom{k-1}{t-1} \binom{k-1}{t+1}, \\ M_{t,k} &= \#\{\{u_l\} : S_l \geq 0 \forall l \text{ and } S_{2k-1} = 1, N_{\epsilon,1} = t, N_{0,-1} = t\} \\ &= \#C_0 - \#C_1. \end{aligned}$$

Now we will show $\#C_1 = \#C_2$. Note that for $\{u_l\} \in C_1$, there exist l such that $S_{l-1} = -1$. Similarly for $\{u_l\} \in C_2$, there exist l such that $S_{l-1} = -1$. Let

$$\begin{aligned} l_1 &= \sup\{l : S_{l-1} = -1, \{u_l\} \in C_1\}, \\ l_2 &= \sup\{l : S_{l-1} = -1, \{u_l\} \in C_2\}. \end{aligned}$$

Then

$$u_{l_1} = u_{l_1+1} = 1 \quad \text{and} \quad u_{l_2} = u_{l_2+1} = -1.$$

Now define a map $f : C_1 \rightarrow C_2$ as follows: $f(\{u_l\}) = \{u'_l\}$ where

$$u'_l = u_l \forall l \neq l_1, l_1 + 1 \text{ and } u'_{l_1} = -u_{l_1}, u'_{l_1+1} = -u_{l_1+1}.$$

Similarly define $g : C_2 \rightarrow C_1$ as $g(\{u_l\}) = \{u'_l\}$ where

$$u'_l = u_l \forall l \neq l_2, l_2 + 1 \text{ and } u'_{l_2} = -u_{l_2}, u'_{l_2+1} = -u_{l_2+1}.$$

It is easy to see that f and g are injective. Hence $\#C_1 = \#C_2$. Therefore

$$M_{t,k} = \#C_0 - \#C_1 = \#C_0 - \#C_2 = \binom{k-1}{t}^2 - \binom{k-1}{t+1} \binom{k-1}{t-1}.$$

□

We now provide brief sketches of the steps verifying the existence of the limit (8) for different matrices.

5.1. Wigner matrix: the semicircular law. The semi-circular law \mathcal{L}_W arises as the LSD of $n^{-1/2}W_n^{(s)}$. It has the density function

$$p_W(s) = \begin{cases} \frac{1}{2\pi} \sqrt{4 - s^2} & \text{if } |s| \leq 2, \\ 0 & \text{otherwise.} \end{cases} \tag{11}$$

All its odd moments are zero. The even moments are given by

$$\beta_{2k}(\mathcal{L}_W) = \int_{-2}^2 s^{2k} p_W(s) ds = \frac{(2k)!}{k!(k+1)!}. \tag{12}$$

Wigner (1958)[47] assumed the entries $\{x_i\}$ to be i.i.d. real Gaussian and established the convergence of $E[F^{n^{-1/2}W_n^{(s)}}(\cdot)]$ to the semi-circular law (11). Subsequent improvements and extensions can be found in Grenander (1963, pages 179 and 209)[26], Arnold (1967)[2] and Bai (1999)[6].

Theorem 5.1. *Let $W_n^{(s)}$ be the $n \times n$ Wigner matrix with the entries $\{x_{ij} : 1 \leq i \leq j, j \geq 1\}$ satisfying Assumption I or II. Then with probability one, $\{F^{n^{-1/2}}W_n\}$ converges weakly to the semicircular law \mathcal{L}_W given in (11).*

Proof. By Lemma 3, Lemma 5 and Theorem 4.1, it is enough to show that for pair matched word w ,

$$\frac{1}{n^{1+k}}|\Pi^*(w)| \rightarrow 1 \text{ or } 0 \text{ according as } w \text{ is or is not a Catalan word.} \tag{13}$$

Note that if $\pi \in \Pi^*(w), w[i] = w[j] \Rightarrow L(\pi(i - 1), \pi(i)) = L(\pi(j - 1), \pi(j))$. Then

$$(\pi(i - 1), \pi(i)) = \begin{cases} (\pi(j - 1), \pi(j)) & \text{(constraint } C1) \text{ or} \\ (\pi(j), \pi(j - 1)) & \text{(constraint } C2). \end{cases}$$

For any matched word w , there are k such constraints. Since each constraint is either $C1$ or $C2$, there are at most 2^k choices in all. Let λ be a typical choice of k constraints and $\Pi_\lambda^*(w)$ be the subset of $\Pi^*(w)$ corresponding to λ and so,

$$\Pi^*(w) = \bigcup_{\lambda} \Pi_\lambda^*(w), \text{ a disjoint union.} \tag{14}$$

Fix w and λ . For $\pi \in \Pi_\lambda^*(w)$, consider the graph with vertices $\pi(0), \pi(1), \dots, \pi(2k)$. The edge set is defined as follows:

- (i) if $w[i] = w[j]$ yields constraint $C1$, connect $(\pi(i - 1), \pi(j - 1))$ and $(\pi(i), \pi(j))$.
- (ii) if $w[i] = w[j]$ yields constraint $C2$, connect $(\pi(i - 1), \pi(j))$ and $(\pi(i), \pi(j - 1))$.
- (iii) connect $(\pi(0), \pi(2k))$, ensuring that π is indeed a circuit.

So, the graph has a total of $(2k + 1)$ edges. These may include loops and double edges. By abuse of notation, $\pi(i)$ thus denotes both, a vertex and its numerical value. One shows by an easy argument that the graph has $(k + 1)$ connected components if and only if w is Catalan and all constraints are $C2$. See Bose and Sen (2008)[21] for details. Denote by λ_0 the case when all constraints are $C2$. Note that

$$|\Pi_{\lambda_0}^*(w)| = n^{k+1}. \tag{15}$$

If w is Catalan and not all constraints are $C2$, or, w is not Catalan and λ is arbitrary, then the corresponding graph has at most k connected components and hence $|\bigcup_{\lambda \neq \lambda_0} \Pi_\lambda^*(w)| \leq 2^k n^k$ implying

$$\frac{1}{n^{k+1}}|\bigcup_{\lambda \neq \lambda_0} \Pi_\lambda^*(w)| \rightarrow 0. \tag{16}$$

Now (13) follows by combining (14), (15) and (16), and the proof is complete. □

Remark 1. *Robustness of the semicircle law:*

(i) For the Wigner matrix, $\Delta(L, f) = 1$ with $f(x) = x$ and $\alpha_n = 2$. The following can be proved using the approach described here: If A_n is symmetric where L satisfies $\Delta(L, f) = 1$ with $f(x) = x$ and $\alpha_n = O(1)$ and the input sequence satisfies Assumption I or II, then $F^{n^{-1/2}A_n}$ converges a.s. to the semicircle law. This and other related results on the Wigner matrix may be found in Bannerjee (2010) [4].

(ii) Consider Wigner matrices with the input random variables having possibly different variances which repeat periodically modulo some integer m_n . Then the LSD turns out to be a scaled semicircular distribution. The details are available in Sen (2010) [43].

(iii) Anderson and Zeitouni (2006)[1] considers an $n \times n$ symmetric random matrix with on-or-above-diagonal terms of the form $\frac{1}{\sqrt{n}}f(\frac{i}{n}, \frac{j}{n})\xi_{ij}$ where ξ_{ij} are zero mean unit variance i.i.d. random variables with all moments bounded and f is a continuous function on $[0, 1]^2$ such that $\int_0^1 f^2(x, y)dy = 1$. They show that the empirical distribution of eigenvalues converges weakly to the semi-circular law.

5.2. Toeplitz and Hankel matrices.

5.2.1. Standard Toeplitz and Hankel. Their LSD were established by Bryc, Dembo and Jiang (2006)[22] and Hammond and Miller (2005)[28].

Theorem 5.2. *If $\{x_i\}$ satisfies Assumption I or II then a.s., $\{F^{n^{-1/2}T_n^{(s)}}\}$ and $\{F^{n^{-1/2}H_n^{(s)}}\}$ converge to symmetric distributions, \mathcal{L}_T and \mathcal{L}_H respectively.*

$\mathcal{L}_T, \mathcal{L}_H$ have unbounded support. Their moments may be expressed as volumes of certain subsets of hypercubes. Obtaining further properties of the LSD is an open problem.

Proof of Theorem 5.2. We sketch the main steps in the proof for the Toeplitz matrix. Since the L -function satisfies Property B with $f(x) = x$, it is enough to obtain $\lim_{n \rightarrow \infty} \frac{1}{n^{1+k}} |\Pi^*(w)|$. From Bryc, Dembo and Jiang (2006)[22], this limit is equal to $\lim_{n \rightarrow \infty} \frac{1}{n^{1+k}} |\Pi^{**}(w)|$, where

$$\Pi^{**}(w) = \{ \pi : w[i] = w[j] \Rightarrow \pi(i-1) - \pi(i) + \pi(j-1) - \pi(j) = 0 \}.$$

Let $v_i = \pi(i)/n$ and $U_n = \{0, 1/n, 2/n, \dots, (n-1)/n\}$. The number of elements in $\Pi^{**}(w)$ then equals

$$\# \left\{ (v_0, v_1, \dots, v_{2k}) : v_0 = v_{2k}, v_i \in U_n \text{ and } v_{i-1} - v_i + v_{j-1} - v_j = 0 \text{ if } w[i] = w[j] \right\}.$$

Let $S = \{0\} \cup \{\min(i, j) : w[i] = w[j], i \neq j\}$ be the set of all indices corresponding to the generating vertices of word w and clearly, $|S| = k + 1$. If $\{v_i\}$ satisfy k equations then each v_i is a unique linear combination of $\{v_j\}$ where $j \in S$ and $j \leq i$. Denoting $\{v_i : i \in S\}$ by v_S , we write $v_i = L_i^T(v_S) \forall i = 0, 1, \dots, 2k$. Note that these linear functions $\{L_i^T\}$ also depend on the word w . Clearly, $L_i^T(v_S) = v_i$ if $i \in S$ and also summing the k equations would imply $L_{2k}^T(v_S) = v_0$. So

$$|\Pi^{**}(w)| = \#\{v_S : L_i^T(v_S) \in U_n \text{ for all } i = 0, 1, \dots, n\}. \tag{17}$$

Since $\frac{1}{n^{1+k}}|\Pi^{**}(w)|$ is nothing but the $(k + 1)$ dimensional Riemann sum for the function $I(0 \leq L_i^T(v_S) \leq 1, \forall i \notin S \cup \{2k\})$ over $[0, 1]^{k+1}$,

$$\lim_{n \rightarrow \infty} \frac{1}{n^{1+k}}|\Pi^{**}(w)| = \underbrace{\int_0^1 \dots \int_0^1}_{k+1} I(0 \leq L_i^T(v_S) \leq 1, \forall i \notin S \cup \{2k\}) dv_S := p_T(w) \tag{18}$$

and $\beta_{2k}(\mathcal{L}_T) = \sum_{l(w)=2k, |w|=k} w \text{ matched, } p_T(w)$.

Similarly $\beta_{2k}(\mathcal{L}_H) = \sum_{l(w)=2k, |w|=k} w \text{ matched, } p_H(w)$, where $p_H(w)$ is given by

$$\underbrace{\int_0^1 \dots \int_0^1}_{k+1} I(0 \leq L_i^H(v_S) \leq 1, \forall i \notin S \cup \{2k\}) I(v_0 = L_{2k}^H(v_S)) dv_S. \tag{19}$$

□

5.2.2. Balanced Toeplitz and Hankel matrices. Excluding the diagonal, each variable in the Wigner matrix appears equal number of times (twice). The Toeplitz and Hankel matrices do not have this property and it seems natural to consider the following balanced versions, first considered by Sen (2006)[42]. For proof of the next theorem see Basak and Bose (2009)[12]. Let

$$BH_n = \begin{bmatrix} \frac{x_1}{\sqrt{1}} & \frac{x_2}{\sqrt{2}} & \frac{x_3}{\sqrt{3}} & \dots & \frac{x_{n-1}}{\sqrt{n-1}} & \frac{x_n}{\sqrt{n}} \\ \frac{x_2}{\sqrt{2}} & \frac{x_3}{\sqrt{3}} & \frac{x_4}{\sqrt{4}} & \dots & \frac{x_n}{\sqrt{n}} & \frac{x_{n+1}}{\sqrt{n+1}} \\ \frac{x_3}{\sqrt{3}} & \frac{x_4}{\sqrt{4}} & \frac{x_5}{\sqrt{5}} & \dots & \frac{x_{n+1}}{\sqrt{n+1}} & \frac{x_{n+2}}{\sqrt{n+2}} \\ & & & \vdots & & \\ \frac{x_n}{\sqrt{n}} & \frac{x_{n+1}}{\sqrt{n+1}} & \frac{x_{n+2}}{\sqrt{n+2}} & \dots & \frac{x_{2n-2}}{\sqrt{2}} & \frac{x_{2n-1}}{\sqrt{1}} \end{bmatrix}.$$

$$BT_n = \begin{bmatrix} \frac{x_0}{\sqrt{n}} & \frac{x_1}{\sqrt{n-1}} & \frac{x_2}{\sqrt{n-2}} & \dots & \frac{x_{n-2}}{\sqrt{2}} & \frac{x_{n-1}}{\sqrt{1}} \\ \frac{x_1}{\sqrt{n-1}} & \frac{x_0}{\sqrt{n}} & \frac{x_1}{\sqrt{n-1}} & \dots & \frac{x_{n-3}}{\sqrt{3}} & \frac{x_{n-2}}{\sqrt{2}} \\ \frac{x_2}{\sqrt{n-2}} & \frac{x_1}{\sqrt{n-1}} & \frac{x_0}{\sqrt{n}} & \dots & \frac{x_{n-4}}{\sqrt{4}} & \frac{x_{n-3}}{\sqrt{3}} \\ & & & \vdots & & \\ \frac{x_{n-1}}{\sqrt{1}} & \frac{x_{n-2}}{\sqrt{2}} & \frac{x_{n-3}}{\sqrt{3}} & \dots & \frac{x_1}{\sqrt{n-1}} & \frac{x_0}{\sqrt{n}} \end{bmatrix}.$$

Theorem 5.3. *If $\{x_i\}$ satisfies Assumption I or II then $\{F^{BT_n}\}, \{F^{BH_n}\}$ converge a.s. to symmetric distributions having unbounded support and finite moments.*

5.3. The reverse circulant and the palindromic matrices.

Bose and Mitra (2002)[18] studied the LSD of $n^{-1/2}R_n^{(s)}$ under finiteness of the third moment. Massey, Miller and Sinsheimer (2007)[34] established the Gaussian limit for $F^{n^{-1/2}PT_n}$ and $F^{n^{-1/2}PH_n}$. The following result may be proved using arguments similar but simpler than those given earlier for the Wigner and Toeplitz matrices. See Bose and Sen (2008)[21] for details. Let \mathcal{L}_R be the distribution with density and moments

$$f_R(x) = |x| \exp(-x^2), \quad -\infty < x < \infty, \quad \beta_{2k+1}(\mathcal{L}_R) = 0 \text{ and } \beta_{2k}(\mathcal{L}_R) = k! \quad k \geq 0.$$

Theorem 5.4. *If $\{x_i\}$ satisfies Assumption I or II, then a.s., $\{F^{n^{-1/2}R_n^{(s)}}\}$ converges to \mathcal{L}_R and $\{F^{n^{-1/2}A_n}\}$ for $A_n = PT_n, C_n^{(s)}, PH_n$ and DH_n , converge to the standard Gaussian distribution.*

Proof. First consider $R_n^{(s)}$. It is enough to show that

(i) If w is pair matched and not symmetric then $\lim_{n \rightarrow \infty} \frac{1}{n^{k+1}} |\Pi^*(w)| = 0$.

(ii) If w is symmetric then for every choice of the generating vertices there is exactly one choice for the nongenerating vertices.

Proof (i) Since w is pair matched let $\{(i_s, j_s), 1 \leq s \leq k\}$ be such that $w[i_l] = w[j_l]$, $j_s, 1 \leq s \leq k$ is in ascending order and $j_k = 2k$. We use the notation from the proof of Theorem 5.2. So, $|\Pi^*(w)| = \sum_{\nu=(\nu_1, \nu_2, \dots, \nu_k) \in \{-1, 0, 1\}^k}$

$$\#\left\{ (v_0, v_1, \dots, v_{2k}) : v_0 = v_{2k}, v_i \in U_n, \text{ and } v_{i_s-1} + v_{i_s} - v_{j_s-1} - v_{j_s} = \nu_s \right\}.$$

Observe that $v_i = L_i^H(v_S) + a_i^{(\nu)}$, $i \notin S$ for some integer $a_i^{(\nu)}$. As in the Hankel case, we easily reach the following equality (compare with (19)):

$$\lim_{n \rightarrow \infty} \frac{1}{n^{k+1}} |\Pi^*(w)| = \sum_{\nu} \underbrace{\int_0^1 \dots \int_0^1}_{k+1} \mathbb{I}(0 \leq L_i^H(v_S) + a_i^{(\nu)} \leq 1, \forall i \notin S \cup \{2k\}) \mathbb{I}(v_0 = L_{2k}^H(v_S) + a_{2k}^{(\nu)}) dv_S.$$

For the integral to be non zero, we must have $v_0 = L_{2k}^H(v_S) + a_{2k}^{(\nu)}$.

Let $t_i = v_{i-1} + v_i$. From the definition of $\Pi^*(w)$, $v_{2k} = v_{2k} + \sum_{s=1}^k \alpha_s (t_{i_s} - t_{j_s} - \nu_s)$ for some $\{\alpha_i\}$. We choose them as follows: Let $\alpha_k = 1$. Having fixed $\alpha_k, \alpha_{k-1}, \dots, \alpha_{s+1}$, we choose α_s as follows: (a) if $j_s + 1 \in \{i_m, j_m\}$ for some $m > s$, then set $\alpha_s = \pm \alpha_m$ according as $j_s + 1$ equals i_m or j_m , (b) if there is no such m , choose α_s arbitrarily. By this choice of $\{\alpha_s\}$, $v_{2k} = v_{2k} +$

$\sum_{s=1}^k \alpha_s(t_{i_s} - t_{j_s} - \nu_s) = L_{2k}^H(v_S) + a_{2k}^{(\nu)}$. Hence $v_{2k} + \sum_{s=1}^k \alpha_s(t_{i_s} - t_{j_s}) + a_{2k}^{(\nu)} - v_0 = 0$ and thus coefficient of each v_i in the left side has to be zero including the constant term. This implies that w is symmetric, proving (i).

(ii) First fix the generating vertices. Then we determine the nongenerating vertices from left to right. Consider $L(\pi(i - 1), \pi(i)) = L(\pi(j - 1), \pi(j))$ where $i < j$ and $\pi(i - 1), \pi(i)$ and $\pi(j - 1)$ have been determined. We rewrite it as

$$\pi(j) \bmod n = Z \text{ where } Z = (L(\pi(i - 1), \pi(i)) - \pi(j - 1)) \bmod n \in \{0, 1, \dots, n\}.$$

This determines $\pi(j)$ uniquely, since $1 \leq \pi(j) \leq n$. Continuing, we obtain the whole circuit uniquely and the result is proved for $\{n^{-1/2}R_n^{(s)}\}$.

For other matrices, similar arguments show that (ii) holds for all pair-matched words. We omit the details. This completes the proof. \square

Remark 2. In a recent paper, Jackson, Miller and Pham (2010) [29] studied the situation when there is more than one palindrome in the first row of a symmetric Toeplitz matrix and used method of moments to show that under certain moment assumptions, the limiting spectral distribution exists and has an unbounded support.

Table 1. Words and moments for symmetric X .

MATRIX	w Cat.	w sym. not Cat.	Other w	β_{2k} or LSD
$C_n^{(s)}$	1	1	1	$\frac{(2k)!}{2^k k!}, N(0, 1)$
PT_n	1	1	1	ditto
PH_n	1	1	1	ditto
DH_n	1	1	1	ditto
$R_n^{(s)}$	1	1	0	$k!, \mathcal{L}_R$
$T_n^{(s)}$	1	$0 < p_T(w) < 1$	$0 < p_T(w) < 1$	$\frac{(2k)!}{k!(k+1)!} \leq \beta_{2k} \leq \frac{(2k)!}{k!2^k}$
$H_n^{(s)}$	1	$0 < p_H(w) = p_T(w) < 1$	0	$\frac{(2k)!}{k!(k+1)!} \leq \beta_{2k} \leq k!$
$W_n^{(s)}$	1	0	0	$\frac{(2k)!}{k!(k+1)!}, \mathcal{L}_W$

5.4. XX' matrices.

5.4.1. Sample covariance matrix. For historical information on the LSD of $S = A_p(W)$, see Bai and Yin (1988)[9], Marčenko, and Pastur (1967)[33], Grenander and Silverstein (1977)[27], Wachter (1978)[46], Jonsson (1982)[30], Yin and Krishnaiah (1985)[49], Yin (1986)[48] and Bai and Zhou (2008)[7].

We first describe the Marčenko-Pastur law denoted by \mathcal{L}_{MPy} : it has a positive mass $1 - \frac{1}{y}$ at the origin if $y > 1$. Elsewhere it has a density:

$$p_{MPy}(x) = \begin{cases} \frac{1}{2\pi xy} \sqrt{(b-x)(x-a)} & \text{if } a \leq x \leq b, \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

where $a = a(y) = (1 - \sqrt{y})^2$ and $b = b(y) = (1 + \sqrt{y})^2$.

Moments of \mathcal{L}_{MPy} are (see Bai (1999)[6] or Bai and Silverstein (2006)[8]):

$$\beta_k(\mathcal{L}_{MPy}) = \sum_{t=0}^{k-1} \frac{1}{t+1} \binom{k}{t} \binom{k-1}{t} y^t, \quad k \geq 1.$$

Theorem 5.5. (a) Suppose $\{x_{ij}\}$ satisfy Assumption I or II and $p \rightarrow \infty$. If $p/n \rightarrow y \in (0, \infty)$, then $\{F^{A_p(W)}\}$ converges to \mathcal{L}_{MPy} a.s..

(b) Suppose $\{x_{ij}\}$ satisfy Assumption I or they are i.i.d. with mean 0, variance 1 and bounded fourth moment. If $p \rightarrow \infty$ and $p/n \rightarrow 0$ then $\{F^{\sqrt{\frac{p}{p}}(A_p(W)-I_p)}\}$ converges to \mathcal{L}_W a.s. where I_p is the identity matrix of order p .

Proof. (a) We apply mutas mutandis the proof given for the Wigner matrix.

$$\beta_k(S) = p^{-1}n^{-k} \sum_{\pi} x_{L1(\pi(0),\pi(1))}x_{L2(\pi(1),\pi(2))}x_{L1(\pi(2),\pi(3))} \cdots x_{L2(\pi(2k-1),\pi(2k))}.$$

A circuit π now has the *non uniform* range $1 \leq \pi(2m) \leq p, 1 \leq \pi(2m+1) \leq n$. It is said to be matched if it is matched within the same $Li, i = 1, 2$ or across. For any w , let $\tilde{\Pi}(w)$ be the possibly larger class of circuits with the range $1 \leq \pi(i) \leq \max(p, n), 1 \leq i \leq 2k$. Likewise define $\tilde{\Pi}^*(w)$.

Lemma 3 remains valid in the present case. See Bose and Sen (2008)[21]. Hence only the pair matched circuits potentially contribute and we need to calculate

$$\lim_n \sum_w \sum_{\pi \in \Pi(w)} \frac{1}{n^k p} \mathbb{E}[x_{L1(\pi(0),\pi(1))} \cdots x_{L2(\pi(2k-1),\pi(2k))}] = \lim_n \sum_{\substack{w: \text{matched,} \\ |w|=k}} \frac{|\tilde{\Pi}^*(w)|}{n^k p}.$$

We need *exactly* $(k+1)$ generating vertices (hence k nongenerating vertices) for a contribution. There is an obvious similarity between the matching restrictions here and the ones we encountered for the Wigner link function. Note that $L1(\pi(i-1), \pi(i)) = L2(\pi(j-1), \pi(j))$ for $i \neq j$ implies a $C2$ constraint. On the other hand, $Lt(\pi(i-1), \pi(i)) = Lt(\pi(j-1), \pi(j)), t = 1$ or 2 , yields a $C1$ constraint. However, unlike the Wigner matrix, $w[i] = w[j]$ implies exactly one of the constraints is satisfied ($C1$ if i and j have same parity and $C2$ otherwise). Hence there is a *unique* $\bar{\lambda}$ (depending on w) such that $\Pi^*(w) = \Pi_{\bar{\lambda}}^*(w)$.

As before, let λ_0 be the index when all constraints in $\Pi_{\lambda_0}^*(w)$ are $C2$. Let $\tilde{\Pi}_{\lambda}^*(w)$ denote the class $\Pi_{\lambda}^*(w)$ but where $1 \leq \pi(i) \leq \max(p, n)$, $i = 0, 1, \dots, 2k$.

If w is not Catalan then it is easy to see that $\bar{\lambda} \neq \lambda_0$. Hence it follows that

$$n^{-k}p^{-1}|\Pi_{\bar{\lambda}}^*(w)| \leq C[\max(p, n)]^{-(k+1)} \left| \bigcup_{\lambda \neq \lambda_0} \tilde{\Pi}_{\lambda}^*(w) \right| \rightarrow 0.$$

For any $0 \leq t \leq k - 1$, if w is Catalan with $(t + 1)$ generating even vertices (with range p) and $(k - t)$ generating odd vertices (with range n) then

$$\lim_{n \rightarrow \infty} n^{-k}p^{-1}|\Pi_{\lambda_0}^*(w)| = \lim_{n \rightarrow \infty} n^{-k}p^{-1}(p^{t+1}n^{k-t}) = y^t.$$

Hence $\lim E[\beta_k(S)] = \sum_{t=0}^{k-1} M_{t,k}y^t$ and the result follows from Lemma 5 (b).

(b) For ease of presentation, we assume that the input sequence satisfies Assumption I. The proof when it is i.i.d. with finite fourth moment requires an appropriate truncation lemma and is available in Bai (1999)[6]. We sketch briefly how the Wigner link function and hence the semicircle law appears. The details can be found in Bose and Sen (2008)[21] Theorem 5. Note that $E[\beta_k(\sqrt{\frac{n}{p}}(A_p(W) - I_p))] = \frac{1}{n^{k/2}p^{1+k/2}} \sum_{\pi} E[X_{\pi}]$, where X_{π} is equal to

$$(x_{\pi(0),\pi(1)}x_{\pi(2),\pi(1)} - \delta_{\pi(0),\pi(2)}) \cdots (x_{\pi(2k-2),\pi(2k-1)}x_{\pi(2k),\pi(2k-1)} - \delta_{\pi(2k-2),\pi(2k)}),$$

with $\delta_{ij} = I\{i = j\}$. Now $E[X_{\pi}] = 0$ if $(\pi(2i), \pi(2i - 1))$ or $(\pi(2i), \pi(2i + 1))$ occurs only once in the product and if for some j , the value of $\pi(2j + 1)$ does not occur at least twice among $\pi(1), \pi(3), \dots, \pi(2k - 1)$. Define a graph $G = (V, E)$ with $V = V_1 \cup V_2$ and $V_1 = \{\pi(2j), 0 \leq j \leq k\}$ and $V_2 = \{\pi(2j - 1), 1 \leq j \leq k\}$. Let $E = \{(\pi(2l), \pi(2l + 1)), (\pi(2l + 2), \pi(2l + 1)) : 0 \leq l \leq k - 1\}$ (multiple edges count as one). Fix a matching (out of finitely many choices) among the even vertices and one among the odd vertices, such that $E[X_{\pi}] \neq 0$. There are at most $p^{|V_1|}n^{|V_2|}$ corresponding circuits. So the contribution of that term is

$$O\left(\frac{p^{|V_1|}n^{|V_2|}}{p^{k/2+1}n^{k/2}}\right) = O\left(\left(\frac{p}{n}\right)^{k/2-|V_2|}\right). \tag{21}$$

If k is odd then $|V_2| < k/2$ and since $p/n \rightarrow 0$, (21) immediately implies that $E[\beta_k(A_p)] \rightarrow 0$.

If $k = 2m$, we look for π which produce nontrivial contribution. From (21), we must have $|V| = 2m + 1, |E| = 2m, |V_2| = m$ and $|V_1| = m + 1$. Observe that:

- (i) $|V_2| = m$ implies a pair partition of odd vertices. Denote it by a word w of length k . So, $\pi(2i - 1) = \pi(2j - 1)$ iff $w[i] = w[j]$.
- (ii) Each pair in E must occur exactly twice.

(iii) If $(\pi(2l), \pi(2l+1)) = (\pi(2l+2), \pi(2l+1))$ or equivalently $\pi(2l) = \pi(2l+2)$, then $E[X_\pi] = 0$. So, consecutive even vertices cannot be equal.

(iv) Note that (ii) and (iii) together imply that $E[X_\pi] = 1$. Suppose

$$w[i] = w[j] \text{ i.e. } \pi(2i - 1) = \pi(2j - 1) \tag{22}$$

and they are different from the rest of the odd vertices. If we fixed w , then independent of w , there are exactly $N_1(n) = n(n - 1) \dots (n - m + 1)$ choices of odd vertices satisfying the pairing imposed by w .

Consider the four pairs of vertices from E , $(\pi(2i - 2), \pi(2i - 1))$, $(\pi(2i), \pi(2i - 1))$, $(\pi(2j - 2), \pi(2j - 1))$ and $(\pi(2j), \pi(2j - 1))$.

By (22) and (ii), they have to be matched in pairs among themselves. Also, (iii) rules out the possibility that the first pair is matched with the second and the third is matched with the fourth. So the other two combinations are the only possibilities. It is easy to verify that this is the same as saying that

$$L(\pi(2i - 2), \pi(2i)) = L(\pi(2j - 2), \pi(2j)) \tag{23}$$

where L is the Wigner link function. Let $\pi^*(i) = \pi(2i)$. Equation (23) implies that π^* is a matched circuit of length k . Let $\Pi^*(w) =$ all circuits π^* satisfying Wigner link function. Then $\lim_{p \rightarrow \infty} \frac{1}{p^{m+1}} |\Pi^*(w)| = 1$ or 0 according as w is or is not Catalan. Hence, the following equalities hold and (M1) is established.

$$\begin{aligned} \lim_{n,p} E[\beta_k(A_p)] &= \lim_{n,p} \frac{1}{p^{m+1} n^m} \sum_{w:\text{matched}, |w|=m} N_1(n) |\Pi^*(w)| \\ &= \lim_p \frac{1}{p^{m+1}} \sum_{w:\text{matched}, |w|=m} |\Pi^*(w)| = \frac{(2m)!}{(m + 1)!m!}. \end{aligned}$$

□

Remark 3. *Simulated eigenvalue distribution of the sample autocovariance matrix and a close cousin of it were given in Sen (2006)[42]. The former is defined as $\Gamma_n = n^{-1}((\sum_{t=1}^{n-|i-j|} x_t x_{t+|i-j|}))_{i,j=1,\dots,n}$. This is also a Toeplitz matrix but with a dependent input sequence. Assuming that $\{x_t\}$ satisfies Assumption II, Basak (2009)[10] showed that the LSD exists, and Basak, Bose and Sen (2010)[14] showed that the LSD exists when $x_t = \sum_{j=0}^d a_j \epsilon_{t-j}$ with $\{\epsilon_t\}$ satisfying Assumption II. They also showed that the modified matrix $\bar{\Gamma}_n = n^{-1}((\sum_{t=1}^n x_t x_{t+|i-j|}))_{i,j=1,\dots,n}$ which is not nonnegative definite also has an LSD.*

5.4.2. XX' matrices with Toeplitz, Hankel and reverse circulant structures. Let $L_p : \{1, 2, \dots, p\} \times \{1, 2, \dots, n = n(p)\} \rightarrow \mathbb{Z}$ be a sequence of link functions. Define the following generalization of the S matrix:

$$A_p(X) = (1/n)XX', \text{ where } X = ((x_{L_p(i,j)}))_{1 \leq i \leq p, 1 \leq j \leq n}.$$

In particular, consider the following choices for X :

(Asymmetric) Toeplitz $T = ((x_{i-j}))_{p \times n}$.

(Asymmetric) Hankel H with (i, j) th entry x_{i+j} if $i > j$ and $x_{-(i+j)}$ if $i \leq j$.

(Asymmetric) reverse circulant R with $L(i, j) = (i + j) \bmod n$ for $i \leq j$ and $-[(i + j) \bmod n]$ for $i > j$.

(Asymmetric) circulant C where $L(i, j) = (n - i + j) \bmod n$.

Also let $H_p^{(s)}$ and $R_p^{(s)}$ be the $p \times n$ rectangular versions of $H_n^{(s)}$ and $R_n^{(s)}$.

Assumption III. $\{x_i\}$ are independent with mean zero and variance 1. Further, $\lambda \geq 1$ is such that $p = O(n^{1/\lambda})$ and $\sup_i E(|x_i|^{4(1+1/\lambda)+\delta}) < \infty$ for some $\delta > 0$.

For a proof of the following theorem see Bose, Gangopadhyay and Sen (2009)[20].

Theorem 5.6. (a) If Assumption I or II holds and $\frac{p}{n} \rightarrow y \in (0, \infty)$, then $\{F^{A_p(X)}\}$, where X is T, H, R or C , converge in distribution a.s. to nonrandom distributions which do not depend on the distribution of $\{x_i\}$.

(b) If Assumption III holds, $p \rightarrow \infty$ and $p/n \rightarrow 0$, then $F^{\sqrt{\frac{p}{n}}(A_p(X) - I_p)} \rightarrow \mathcal{L}_T$ a.s. for X equal to $T, H, R, C, H_p^{(s)}$ or $R_p^{(s)}$.

5.5. Band matrices. If the top right corner and the bottom left corner elements of a matrix are zeroes, we call it a band matrix. The amount of banding may change with the dimension of the matrix and may alter the LSD drastically. See for example Popescu (2009)[41]. In this section we discuss the Toeplitz, Hankel and circulant band matrices. Similar band matrices have been considered by Kargin (2009)[31] and Liu and Wang (2009)[32]. Proofs of the next two theorems are available in Basak and Bose (2009)[11]. Let $\{m_n\}$ be a sequence of integers. For simplicity we write m for m_n . Consider the following assumptions.

Assumption I* $\{x_i\}$ are independent with mean 0 and variance 1 and satisfy

- (i) $\sup E[|x_i|^{2+\delta}] < \infty$ for some $\delta > 0$,
- (ii) For all large t , $\lim n^{-2} \sum_{i=0}^n E[|x_i|^4 I(|x_i| > t)] = 0$.

Assumption IV $\{m_n\} \rightarrow \infty$ and $\lim_{n \rightarrow \infty} m_n/n = \alpha < \infty$.

Assumption V $\sum_{n=1}^{\infty} m_n^{-2} < \infty$. (Holds trivially when $\alpha \neq 0$).

Table 2. Words and moments for XX' matrices.

MATRIX	w Cat.	Other w	β_k and LSD
$p/n \rightarrow 0$ $\sqrt{\frac{n}{p}}(S - I_p)$ $(S = n^{-1}W_pW'_p)$ $\sqrt{\frac{n}{p}}(A_p(X) - I_p)$ $(X = T, H, R, C)$	1 (Cat. in p)	0	$\frac{(2k)!}{k!(k+1)!}, \mathcal{L}_W$ \mathcal{L}_T
$p/n \rightarrow y \neq 0, \infty$ $S = n^{-1}W_pW'_p$ $A_p(X)$ $(X = T, H, R_p, C_p)$	1	0	$\sum_{t=0}^{k-1} \frac{1}{t+1} \binom{k}{t} \binom{k-1}{t} y^t, \mathcal{L}_{MPy}$ different, but universal

(i) **Type I banding.** For any A_n , the Type I band matrix A_n^b is the matrix A_n with input $\{x_i I(i \leq m) + 0I(i > m)\}$.

Let $N(0, 2)$ denote a normal random variable with mean zero and variance 2. Let \Rightarrow denote weak convergence of probability measures.

Theorem 5.7. *Suppose Assumption IV holds and one of the following holds: (A) Assumption I, (B) Assumption II or (C) Assumption I*(i), (ii). Then in probability,*

(a) *If $m_n \leq n/2$ then $F^{m_n^{-1/2}A} \Rightarrow N(0, 2)$ for $A = C_n^{(s)b}, DH_n^b, PT_n^b$ and PH_n^b .*

(b) *If $m_n \leq n$ then $F^{m_n^{-1/2}R_n^{(s)b}} \Rightarrow \mathcal{L}_R$.*

(c) *If $m_n \leq 2n$ then $F^{m_n^{-1/2}H_n^{(s)b}} \Rightarrow H_\alpha^b$ which is symmetric and H_0^b is the degenerate distribution at zero.*

(d) *If $m_n \leq n$ then $F^{m_n^{-1/2}T_n^{(s)b}} \Rightarrow T_\alpha^b$ which is symmetric and $T_0^b = N(0, 2)$.*

If Assumption V holds, the convergence are a.s. in cases (A) and (B).

(ii) **Type II banding.** The Type II band version H_n^B of $H_n^{(s)}$ is defined with the input sequence $\{x_i I(|i - n| \leq m) + 0I(|i - n| > m)\}$. The Type II band versions R_n^B of $R_n^{(s)}$ and T_n^B of $T_n^{(s)}$ are defined with the input sequence $\{x_i I(\{i \leq$

$m\} \cup \{i \geq n - m\}) + 0I(m < i < n - m)\}$. Type II banding does not yield any nontrivial situations for symmetric, doubly symmetric and palindromic matrices.

Theorem 5.8. *Suppose Assumption IV holds and any one of the following holds: (A) Assumption I, (B) Assumption II or (C) Assumption I*(i), (ii). Then in probability,*

- (a) *If $m_n \leq n/2$ then $F^{(2m_n)^{-1/2}R_n^B} \Rightarrow \mathcal{L}_R$.*
- (b) *If $m_n \leq n$ then $F^{(2m_n)^{-1/2}H_n^B} \Rightarrow H_\alpha^B$ which is symmetric and $H_0^B = \mathcal{L}_R$.*
- (c) *If $m_n \leq n/2$ then $F^{m_n^{-1/2}T_n^B} \Rightarrow T_\alpha^B$ which is symmetric and $T_0^B = N(0, 2)$.*

If Assumption V holds, the convergence are a.s. in cases (A) and (B).

5.6. Matrices with dependent entries. Let $x_t = \epsilon_t \epsilon_{t+1} \cdots \epsilon_{t+d-1}$ where $\{\epsilon_i\}$ are i.i.d. To deal with this kind of dependence between $\{x_i\}$, we extend the concept of matching.

Matched circuits: Let L be a link function. Let $M^\pi = ((m_{i,j}))$ be the $d \times h$ matrix where $m_{i,j} = L(\pi(j - 1), \pi(j)) + i - 1$. We say that π is d -**matched** if every element of M^π appears at least twice. This notion is extended to d -joint matching and d -cross matching in the obvious way. Note the following facts:

1. No two entries belonging to the same column of M^π can be equal.
2. If some entry in the j_1 -th column of M^π is equal to some entry in its j_2 -th column then $|L(\pi(j_1 - 1), \pi(j_1)) - L(\pi(j_2 - 1), \pi(j_2))| \leq d - 1$.

Let $N_{h,3+}^M =$ Number of d -matched circuits of length h with at least one entry of M^π repeated at least thrice, and let $Q_{h,4}^M =$ Number of circuits $(\pi_1, \pi_2, \pi_3, \pi_4)$ of length h which are jointly d -matched and jointly d -cross matched. The following lemma was proved in Bose and Sen (2008)[21].

Lemma 5.9. *Suppose (L, f) with $f(x) = x$ satisfies Property B.*

(a) *There are constants $C_{h,d}$ and $K_{h,d}$ such that*

$$N_{h,3+}^M \leq C_{h,d} n^{\lfloor (h+1)/2 \rfloor} \text{ and } Q_{h,4}^M \leq K_{h,d} n^{2h+2}. \tag{24}$$

(b) *Suppose $x_t = \epsilon_t \epsilon_{t+1} \cdots \epsilon_{t+d-1}$ where $\{\epsilon_i\}$ satisfies Assumption I. Let $A_{n,d} = ((x_{L(i,j)}))_{n \times n}$ where (L, f) satisfies Property B with $f(x) = x$. Then for every h*

$$\mathbb{E} \left[\frac{1}{n} \text{Tr} \left(\frac{A_{n,d}}{\sqrt{n}} \right)^h - \mathbb{E} \frac{1}{n} \text{Tr} \left(\frac{A_{n,d}}{\sqrt{n}} \right)^h \right]^4 = O(n^{-2}). \tag{25}$$

As a consequence (M4) holds too.

Lemma 5.10. *Each d -matched circuit π with only pair matchings is also pair-matched w.r.t. L and vice versa. Hence if $l(\pi) = h$ is odd then no d -matched circuit π can be pair-matched.*

Detailed proof of the following theorem is given in Bose and Sen (2008)[21].

Theorem 5.11. *Let $x_t = \epsilon_t \epsilon_{t+1} \cdots \epsilon_{t+d-1}$ where $\{\epsilon_i\}$ satisfies Assumption I. Let $A_{n,d} = ((x_{L(i,j)}))_{n \times n}$ where (L, f) satisfies Property B with $f(x) = x$. If LSD of $\{F^{n^{-1/2}} A_{n,d}\}$ exists a.s. for $d = 1$, then the same LSD holds a.s. for $d \geq 2$.*

Sketch of proof of Theorem 5.11. Let $F_{n,d}$ denote the ESD of $n^{-1/2} A_{n,d}$. Lemma 5.9 and 5.10 imply that for every h, d ,

$$\beta_h(F_{n,d}) - E[\beta_h(F_{n,d})] \rightarrow 0 \text{ almost surely.}$$

On the other hand

$$\begin{aligned} E[\beta_h(F_{n,d})] &= \frac{1}{n} E[\text{Tr}(n^{-1/2} A_{n,d})^h] = \frac{1}{n^{1+h/2}} \sum_{\pi} E[X_{\pi}] \\ &= \frac{1}{n^{1+h/2}} \sum_{\pi \text{ } d\text{-matched}} E[X_{\pi}] \end{aligned}$$

where $X_{\pi} = \prod_{i=1}^h \epsilon_{L(\pi(i-1), \pi(i))} \epsilon_{L(\pi(i-1), \pi(i))+1} \cdots \epsilon_{L(\pi(i-1), \pi(i))+d-1}$.

Lemma 5.9(a) and Lemma 5.10 imply that if h is odd then $\lim E[\beta_h(F_{n,d})] = 0$, and hence for every d , $\lim \beta_h(F_{n,d}) = 0$ a.s..

Now suppose $h = 2k$ is even. Let $\Pi(w)$ be as defined in Section 4 for ordinary L -matching. From Theorem 4.1, a.s.,

$$\lim n^{-(k+1)} \sum_{w:|w|=k} \Pi(w) = \lim \beta_{2k}(n^{-1/2} A_{n,1}) = \lim E[\beta_{2k}(n^{-1/2} A_{n,1})].$$

On the other hand, Lemma 5.9 and Lemma 5.10 imply that for all d

$$\lim n^{-(k+1)} \sum_{w:|w|=k} \Pi(w) = \lim E[\beta_{2k}(F_{n,d})] = \lim \beta_{2k}(F_{n,d}), \text{ almost surely.}$$

□

Here is another result in a dependent situation. For proof see Bose and Sen (2008)[21].

Theorem 5.12. *Let $x_t = \sum_{j=0}^{\infty} a_j \epsilon_{t-j}$ with $\{a_j\}$ satisfying $\sum_j |a_j| < \infty$ and $\{\epsilon_i\}$ satisfying Assumption I and $\sum_j j a_j^2 < \infty$. Then with the input sequence $\{x_t\}$, $\{F^{n^{-1/2}} T_n^{(s)}\}$ and $\{F^{n^{-1/2}} H_n^{(s)}\}$ converge weakly to nonrandom symmetric*

probability measures T_a and H_a respectively. These LSD do not depend on the distribution of ϵ_1 . The $(2k)$ -th moment of T_a and H_a are given by

$$\beta_{2k}(T_a) = A_{2k}\beta_{2k}(\mathcal{L}_T) \text{ and } \beta_{2k}(H_a) = A_{2k}\beta_{2k}(\mathcal{L}_H)$$

where \mathcal{L}_T and \mathcal{L}_H are as in Theorem 5.2 and $A_{2k} = \sum_{d=0}^{\infty} \left(\sum_{\substack{m_1, \dots, m_k \geq 0: \\ \sum_1^k m_i = d}} \prod_{j=1}^k a_{m_j} \right)^2$.

6. Moment Method Applied to Other Matrices

6.1. Mod $[np]$ link functions. Recall that the Hankel and reverse circulant link functions are respectively, $L(i, j) = i + j$ and $L(i, j) = i + j \pmod n$. Define a class of link functions $\{L_p : p \in (0, 2]\}$, where $L_p(i, j) = i + j \pmod [np]$. Then the previous two link functions are special cases. Some results on the LSD with i.i.d. inputs and link function L_p have been established by Basak and Bose (2010)[13]. In particular, when $p = \frac{1}{m}$ for some integer m , the LSD is $(1-p)\delta_0 + p\sqrt{m}R$, where δ_0 is the degenerate distribution at 0 and R has distribution \mathcal{L}_R . Similar extensions were also obtained for mod $[np]$ versions of Toeplitz and symmetric circulant link functions.

6.2. Tridiagonal matrices. Let A_n be the tridiagonal random matrix

$$A_n = \begin{bmatrix} d_n & b_{n-1} & 0 & 0 & \dots & 0 & 0 \\ b_{n-1} & d_{n-1} & b_{n-2} & 0 & \dots & 0 & 0 \\ 0 & b_{n-2} & d_{n-2} & b_{n-3} & \dots & 0 & 0 \\ & & & \vdots & & & \\ 0 & 0 & 0 & \dots & b_2 & d_2 & b_1 \\ 0 & 0 & 0 & \dots & 0 & b_1 & d_1 \end{bmatrix}.$$

Popescu (2009)[41] used the trace formula and the moment method to obtain many interesting limit distributions. He assumed that the sequence $\{d_j, b_j\}$ is independent, moments of d_n are bounded uniformly in n and, $E[(b_n/n^\alpha)^k] \rightarrow m_k$ for every k , as $n \rightarrow \infty$. Let $X_n = A_n/n^\alpha$ and tr denotes the normalized trace i.e., $tr(I) = 1$. He showed that $E[tr(X_n^k)] \rightarrow L_k$ for some $\{L_k\}$ which can be expressed in terms of $\{m_i\}$. Moreover, if $X_{i,n}$ are several independent matrices then the joint moments $E[tr(X_{i_1,n}^{k_1} \dots X_{i_s,n}^{k_s})]$ converge and the limits can be expressed in terms of the corresponding $\{m_r\}$. For example, if $\alpha = 1/2$ and $m_k = 1$ for all k , then the $L_k, k \geq 1$ are moments of the semicircle law. For other values of α , $\{L_k\}$ determines probability distributions whose densities are available in explicit forms.

6.3. Block Matrices. Oraby (2007)[37] discussed the a.s. limiting spectral distributions of some block random matrices. Under the strong assumption that the ESD of the blocks themselves converge a.s. to some limiting spectral distribution, an easy consequence from the theory of polynomials is the a.s. limiting behavior of the spectrum of the block matrix. The proof of the main theorem involves the method of moments.

Let B_k be a block matrix with Hermitian structure of order k (fixed) with blocks formed by independent Wigner matrices of size n . Oraby (2007)[38] showed that its LSD exists and depends only on the structure of the block matrix. When the block structure is circulant, the LSD is a weighted sum of two semicircle laws. In particular, the LSD of a Wigner matrix with k -weakly dependent entries need not be the semicircle law. Bannerjee (2010)[4] considered the case where B_k is symmetric and derived an explicit formula for the moments in terms of the link function L of B_k . In particular, only Catalan words contribute and the support of the LSD lies within $[-2\sqrt{\Delta(L, f)}, 2\sqrt{\Delta(L, f)}]$ with $f(x) = x$.

7. Some Other Methods and Matrices

7.1. Normal approximation and the k circulant matrix. For the circulant matrix, apart from conjugacy, the eigenvalues are asymptotically normal and asymptotically independent. LSD proofs can be developed by appropriate usage of normal approximation methods. See for example, Bose and Mitra (2002)[18] (reverse circulant and symmetric circulant) and Meckes (2009)[35]. Recently Bose, Mitra and Sen (2008)[19] and Bose, Hazra and Saha (2009)[17] used normal approximation to establish LSD for some specific type of k -circulant matrices with independent and dependent inputs respectively.

7.2. Stieltjes transform and the Wigner and sample covariance matrices. Stieltjes transform plays an important role in the study of spectral distribution. For any probability distribution G on the real line, its Stieltjes transform s_G is defined on $\{z : u + iv, v \neq 0\}$ as

$$s_G(z) = \int_{-\infty}^{\infty} \frac{1}{x - z} G(dx).$$

If A has real eigenvalues λ_i , $1 \leq i \leq n$, then the Stieltjes transform of the ESD of A is

$$s_A(z) = \frac{1}{n} \sum_{i=1}^n \frac{1}{\lambda_i - z} = \frac{1}{n} \text{Tr}[(A - zI)^{-1}].$$

Let $\{A_n\}$ be a sequence of random matrices with real eigenvalues and let the corresponding sequence of Stieltjes transform be $\{m_n\}$. If $m_n \rightarrow m$ in some suitable manner, where m is a Stieltjes transform, then the LSD of the sequence

$\{A_n\}$ is the unique probability on the real line whose Stieltjes transform is the function m . It is shown that $\{m_n\}$ satisfies some (approximate) recursion equation. Solving the limiting form of this equation identifies the Stieltjes transform of the LSD. See Bai (1999)[6] for this approach in deriving the LSD for the Wigner and the S matrices and in studying the rate of convergence. Incidentally, no Stieltjes transform based proof is available for the Toeplitz and the Hankel matrices.

7.2.1. Wigner matrix with heavy tailed input. Consider the Wigner matrix $W_n^{(s)}$ with i.i.d. entries belonging to domain of attraction of an α -stable law with $\alpha \in (0, 2)$. Ben Arous and Guionnet (2006)[16] prove that with an appropriate slowly varying function $l(\cdot)$, $\{E[F^{l(n)n^{1/\alpha}-1}W_n^{(s)}]\}$ converges to some law μ_α which depends only on α . This law is symmetric, heavy-tailed, and is absolutely continuous with respect to the Lebesgue measure, except possibly on a compact set of capacity zero. Some similar results for the S matrix and band matrices can be found in Belinschi, Dembo and Guionnet (2009)[15].

7.2.2. I.I.D. Matrix and the Circular Law. Let A_n be the $n \times n$ random matrix with mean 0 and variance 1 i.i.d. complex entries. Then $\{F^{n^{-1/2}A_n}\}$ converges a.s., to the uniform distribution on the unit disk (called the circular law). This was first established for Gaussian entries by Mehta (1967)[36]. Girko (1984)[23] suggested a method of proof for the general case. Bai (1997)[5] considered smooth densities and bounded sixth moment of the entries and showed the result to be true. Gotze and Tikhomirov (2007)[24] showed the result for subgaussian entries and the moment conditions were further relaxed by Pan and Zhou (2010)[39], Gotze and Tikhomirov (2007)[25] and Tao and Vu (2008)[44]. The result in its final form was derived by Tao, Vu and Krishnapur (2009)[45]. The moment method fails for this matrix as all the moments of the circular distribution are zero and they do not determine the distribution uniquely. The Stieltjes transform method was used to show that the ESD converges. The laws of the singular value distribution of $n^{-1/2}A_n - zI$ for complex z also played a crucial role in determining the convergence of the ESD.

8. Discussion

- (i) We have seen that under the boundedness property of the link function, convergence of the moments is a necessary and sufficient condition for the LSD to exist. Moreover, subsequential limits exist. It is not known if suitable restrictions on the link function guarantees the existence of limits of moments.
- (ii) Similarly, given a specific subclass of words, can an appropriate (bounded) link function be devised for which the LSD contribution comes only from these words?

(iii) Under what conditions on the link does the LSD have bounded or unbounded support? Bannerjee (2010)[4] has shown that if Property B is satisfied with $f(x) = x$ and $\alpha_n = o(n)$, then only the Catalan words contribute to the moments and the support of the LSD is a subset of $[-2\sqrt{\Delta(L, f)}, 2\sqrt{\Delta(L, f)}]$.

(iv) We have used the moment method only for real symmetric matrices. Using the moment method for nonsymmetric matrices or for matrices with complex entries does not appear to be convenient. However, more thought on this is needed.

(v) The d -matching helped us to address linear dependence. One can also think of extending the results to input sequences which admit other types of dependence, for example for martingale differences.

(vi) Recall that for the S matrix, there is a positive mass equal to $1 - y^{-1}$ when $p/n \rightarrow y > 1$. It is evident from simulations that a similar phenomenon exists for general XX' matrices. See Bose, Sen and Gangopadhyay (2009)[20]. However, detailed information on the quantum of mass at zero and the gap between 0 and the next point in the support of the LSD is not known.

Acknowledgement

We are grateful to Sayan Bannerjee, Anirban Basak and Sanchayan Sen for helpful discussions.

References

- [1] Anderson, Greg W. and Zeitouni, Ofer (2006). A CLT for a band matrix model. *Probability Theory and Related Fields*, 134, no. 2, 283–338.
- [2] Arnold, L. (1967). On the asymptotic distribution of the eigenvalues of random matrices. *J. Math. Anal. Appl.*, 20, 262–268.
- [3] Anderson, G. W.; Guionnet, A. and Zeitouni, O. (2009). *An Introduction to Random Matrices*. Cambridge University Press.
- [4] Bannerjee, Sayan (2010). Large dimensional random matrices. *M. Stat. Mid Year Project Report*, February 2010. Indian Statistical Institute, Kolkata.
- [5] Bai, Z. D. (1997). Circular law. *Ann. Probab.*, 25, 494–529.
- [6] Bai, Z. D. (1999). Methodologies in spectral analysis of large dimensional random matrices, a review. *Statistica Sinica*, 9, 611–677 (with discussions).
- [7] Bai, Z. D. and Zhou, Wang (2008). Large sample covariance matrices without independent structure in columns. *Statist. Sinica*, 18, 425–442.
- [8] Bai, Z. D. and Silverstein, J. (2006). *Spectral Analysis of Large Dimensional Random Matrices*. Science Press, Beijing.

-
- [9] Bai, Z. D. and Yin, Y. Q. (1988). Convergence to the semicircle law. *Ann. Probab.*, 16, no. 2, 863–875.
- [10] Basak, Anirban (2009). Large dimensional random matrices. *M. Stat. Project Report*, June 2009. Indian Statistical Institute, Kolkata.
- [11] Basak, Anirban and Bose, Arup (2009). Limiting spectral distribution of some band matrices. *Technical Report R16/2009, Stat-Math Unit, Indian Statistical Institute, Kolkata*. To appear in *Periodica Hungarica*.
- [12] Basak, Anirban and Bose, Arup (2010). Balanced random Toeplitz and Hankel Matrices. *Technical Report R01/2010, Stat-Math Unit, Indian Statistical Institute, Kolkata*. To appear in *Elec. Comm. Probab.*
- [13] Basak, Anirban and Bose, Arup (2010). Limiting spectral distribution of a class of patterned matrices. *In preparation*.
- [14] Basak, Anirban; Bose, Arup and Sen, Sanchayan (2010). Limiting spectral distribution of sample autocovariance matrices. *In preparation*.
- [15] Belinschi, Serban; Dembo, Amir and Guionnet, Alice (2009). Spectral measure of heavy tailed band and covariance random matrices. *Comm. Math. Phys.*, 289, no. 3, 1023–1055.
- [16] Ben Arous, Gérard and Guionnet, Alice (2008). The spectrum of heavy tailed random matrices. *Comm. Math. Phys.*, 278, no. 3, 715–751.
- [17] Bose, Arup; Hazra, Rajat Subhra and Saha, Koushik (2009). Limiting spectral distribution of circulant type matrices with dependent inputs. *Electron. J. Probab.*, 14, no. 86, 2463–2491.
- [18] Bose, Arup and Mitra, Joydip (2002). Limiting spectral distribution of a special circulant. *Stat. Probab. Letters*, 60, 1, 111–120.
- [19] Bose, Arup; Mitra, Joydip and Sen, Arnab (2008). Large dimensional random k -circulants. *Technical Report No.R10/2008, Stat-Math Unit, Indian Statistical Institute, Kolkata*.
- [20] Bose, Arup; Gangopadhyay, Sreela and Sen, Arnab (2009). Limiting spectral distribution of XX' matrices. To appear in *Ann. Inst. Henri Poincare Probab. Stat.*
- [21] Bose, Arup and Sen, Arnab (2008). Another look at the moment method for large dimensional random matrices. *Elec. J. Probab.*, 13, 588–628.
- [22] Bryc, W.; Dembo, A. and Jiang, T. (2006). Spectral measure of large random Hankel, Markov and Toeplitz matrices. *Ann. Probab.*, 34, no. 1, 1–38.
- [23] Girko, V. L. (1984). Circular law. *Theory Probab. Appl.*, 4, 694–706.
- [24] Gotze, F. and Tikhomirov, A.N. (2007). On the circular law. *arXiv:math/0702386v1 [math.PR]*.
- [25] Gotze, F. and Tikhomirov, A.N. (2007). The circular law for random matrices. *arXiv:0709.3995v3 [math.PR]*.
- [26] Grenander, U. (1963). *Probabilities on Algebraic Structures*. John Wiley & Sons, Inc., New York-London; Almqvist & Wiksell, Stockholm-Göteborg-Uppsala.
- [27] Grenander, U. and Silverstein, J. W. (1977). Spectral analysis of networks with random topologies. *SIAM J. Appl. Math.*, 32, 499–519.

- [28] Hammond, C. and Miller, S. J. (2005). Distribution of eigenvalues for the ensemble of real symmetric Toeplitz matrices. *J. Theoret. Probab.* 18, no. 3, 537–566.
- [29] Jackson, S., Miller, S. J. and Pham, T. (2010) Distribution of eigenvalues of highly palindromic Toeplitz matrices. *arXiv:1003.2010[math.PR]*
- [30] Jonsson, D. (1982). Some limit theorems for the eigenvalues of a sample covariance matrix. *J. Multivariate Anal.* 12, no. 1, 1–38.
- [31] Kargin, Vladislav (2009). Spectrum of random Toeplitz matrices with band structures. *Elect. Comm. in Probab.* 14 (2009), 412–421.
- [32] Liu, Dang-Zheng and Wang, Zheng-Dong (2009). Limit Distributions for Random Hankel, Toeplitz Matrices and Independent Products. *arXiv:0904.2958v2 [math.PR]*.
- [33] Marčenko, V. A. and Pastur, L. A. (1967). Distribution of eigenvalues in certain sets of random matrices, (Russian) *Mat. Sb. (N.S.)* 72 (114), 507–536.
- [34] Massey, A.; Miller, S. J. and Sinsheimer, J. (2007). Distribution of eigenvalues of real symmetric palindromic Toeplitz matrices and circulant matrices. *J. Theoret. Probab.* 20, 3, 637–662.
- [35] Meckes, Mark W. (2009). Some results on random circulant matrices. *High Dimensional Probability V: The Luminy Volume*, 213–223, IMS Collections 5, Institute of Mathematical Statistics, Beachwood, OH, 2009.
- [36] Mehta, M. L. (1967). *Random matrices and the statistical theory of energy levels*. Academic Press.
- [37] Oraby, Tamer (2007). The limiting spectra of Girko’s block-matrix *J. Theoret. Probab.* 4, 959–970.
- [38] Oraby, Tamer (2007). The spectral laws of Hermitian block-matrices with large random blocks. *Electron. Comm. Probab.* 12, 465–476.
- [39] Pan, G. and Zhou, W. (2010). Circular law, extreme singular values and potential theory. *J. Multivariate Anal.* 101, no. 3, 645–656.
- [40] Pastur, L. (1972). The spectrum of random matrices. (Russian) *Teoret. Mat. Fiz.* 10 no. 1, 102–112.
- [41] Popescu, Ionel (2009). General tridiagonal random matrix models, limiting distributions and fluctuations. *Probab. Theory Relat. Fields*, 144, 179–220.
- [42] Sen, Arnab (2006). Large dimensional random matrices. *M. Stat. Project Report*, May 2006. Indian Statistical Institute, Kolkata.
- [43] Sen, Sanchayan (2010). Large dimensional random matrices. *M. Stat. Mid Year Project Report*, February 2010. Indian Statistical Institute, Kolkata.
- [44] Tao, T. and Vu, V. (2008). Random Matrices: The circular Law, *Communications in Contemporary Mathematics*, 10, 261–307.
- [45] Tao, T.; Vu, V. and Krishnapur M. (Appendix) (2009). Random matrices: universality of the ESDs and the circular law. To appear in *Ann. Probab.*.
- [46] Wachter, K.W. (1978). The strong limits of random matrix spectra for sample matrices of independent elements. *Ann. Probab.* 6, 1–18.
- [47] Wigner, E. P. (1958). On the distribution of the roots of certain symmetric matrices. *Ann. of Math.*, (2), 67, 325–327.

-
- [48] Yin, Y. Q. (1986). Limiting spectral distribution for a class of random matrices. *J. Multivariate Anal.*, 20, no. 1, 50–68.
- [49] Yin, Y. Q. and Krishnaiah, P. R. (1985). Limit theorem for the eigenvalues of the sample covariance matrix when the underlying distribution is isotropic. *Theory Probab. Appl.*, 30, no. 4, 810–816.

Renormalisation Group Analysis of Weakly Self-avoiding Walk in Dimensions Four and Higher

David Brydges* and Gordon Slade†

Abstract

We outline a proof, by a rigorous renormalisation group method, that the critical two-point function for continuous-time weakly self-avoiding walk on \mathbb{Z}^d decays as $|x|^{-(d-2)}$ in the critical dimension $d = 4$, and also for all $d > 4$.

Mathematics Subject Classification (2010). Primary 82B41; Secondary 60K35, 82B28.

Keywords. Self-avoiding walk, Edwards model, renormalization group, supersymmetry, quantum field theory

1. Introduction

We prove $|x|^{-(d-2)}$ decay for the critical two-point function of the continuous-time weakly self-avoiding walk in dimensions $d \geq 4$. This is a summary of the ideas and the steps in the proof. The details are provided in [12]. The proof is based on a rigorous renormalisation group argument. For the case $d > 4$, this provides an approach completely different from the lace expansion methods of [18, 19]. But our main contribution is that our method applies also in the case of the *critical* dimension $d = 4$, where lace expansion methods do not apply.

Renormalisation group methods have been applied previously to study weakly self-avoiding walk on a 4-dimensional *hierarchical* lattice. The continuous-time version of the model has been studied in the series of papers

*Department of Mathematics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2. E-mail: db5d@math.ubc.ca.

†Department of Mathematics, University of British Columbia, Vancouver, BC, Canada V6T 1Z2. E-mail: slade@math.ubc.ca.

[4, 16, 8, 9]; see [5] for a review. More recently, a completely different renormalisation group approach to the discrete-time weakly self-avoiding walk on a 4-dimensional hierarchical lattice has been developed in [20].

The $|x|^{-(d-2)}$ decay for the two-point function for a continuum 4-dimensional Edwards model, with a smoothed delta function, has been proved in [24]; unlike our model, this is not a model of walks taking nearest neighbour steps in the lattice, but it is expected to be in the same universality class as our model. The relation between our model and the Edwards model is discussed in [26]. A big step towards an understanding of the behaviour in dimension $d = 4 - \epsilon$ is taken in [27] (their model is formulated on a lattice in dimension 3 but it mimics the behaviour of the nearest-neighbour model in dimension $4 - \epsilon$).

Our renormalisation group method is a greatly extended and generalised form of work in [4, 8, 9] for the hierarchical lattice and [13, 14, 3, 11] for continuum quantum field theory. Details will appear in [12]. Our method is based on an exact functional integral representation of the two-point function of the continuous-time self-avoiding walk as the two-point function of a quantum field theory containing both bosonic and fermionic fields. Such representations have been recently summarised in [10].

1.1. Background. A self-avoiding walk on the simple cubic lattice \mathbb{Z}^d is an *injective* map

$$\omega : \{0, 1, \dots, n\} \rightarrow \mathbb{Z}^d \tag{1}$$

such that for all i , $\omega(i)$ and $\omega(i + 1)$ are nearest neighbours in \mathbb{Z}^d . We call n the number of steps. The main result of this article will actually be a statement about about random maps $X : [0, T] \rightarrow \mathbb{Z}^d$, but to explain the background we start with self-avoiding walk.

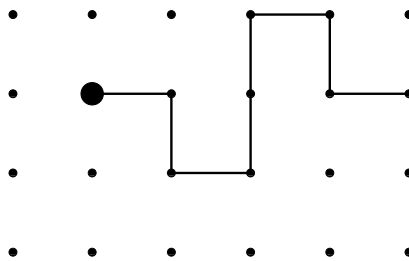


Figure 1. An 8 step self-avoiding walk on \mathbb{Z}^d , $d = 2$.

Let \mathcal{S}_n be the set of all self-avoiding walks with n steps and with $\omega(0) = 0$. Let c_n be the number of elements in \mathcal{S}_n . By declaring that all ω in \mathcal{S}_n have equal probability $1/c_n$ we make \mathcal{S}_n into a probability space with expectation \mathbb{E}_n . The subscript n reminds us that the probability space depends on n . In the sequel “model” means a choice of probability space and law.

This model arose in statistical mechanics. It is, for example, a natural model when one is interested in the conformation of linear polymer molecules. There is another natural model called the *true* or *myopic* self-avoiding walk. Unlike our model, true self-avoiding walk is a stochastic process which at each step looks at its neighbours and chooses uniformly from those visited least often in the past. Recent progress on this model is reported in [23].

The key problem is to determine the growth in n of the mean-square displacement,

$$\mathbb{E}_n |\omega(n)|^2 = c_n^{-1} \sum_{\omega \in \mathcal{S}_n} |\omega(n)|^2, \tag{2}$$

where $|\omega(n)|$ is the Euclidean norm of $\omega(n)$ as an element of \mathbb{Z}^d . More precisely, we want to prove the existence of ν such that

$$\lim_{n \rightarrow \infty} n^{-2\nu} \mathbb{E}_n |\omega(n)|^2 \in (0, \infty), \tag{3}$$

and we want to calculate ν . We will call this the ν problem.

As explained in [26, page 16], there is an easier version of this problem that we will call the *Abelian* ν problem, because proving the existence of ν after solving the Abelian problem is a Tauberian problem. Let $\mathcal{S} = \bigcup_n \mathcal{S}_n$ and let $n(\omega) = n$ for $\omega \in \mathcal{S}_n$. For $z > 0$ we define the *two-point function*

$$G_z(x) = \sum_{\omega \in \mathcal{S}} z^{n(\omega)} \mathbb{1}_{\omega(n(\omega))=x}. \tag{4}$$

Let

$$\chi^{(p)} = \sum_{\omega \in \mathcal{S}} z^{n(\omega)} |\omega(n(\omega))|^p = \sum_{x \in \mathbb{Z}^d} G_z(x) |x|^p. \tag{5}$$

The Abelian version of the ν problem is to determine the growth of $\sqrt{\chi^{(2)}}/\chi^{(0)}$ as $z \uparrow z_c$, where z_c is the common radius of convergence of the power series in this ratio. If ν exists then it equals the Abelian ν . In dimensions $d \geq 5$, according to the following theorem, $\nu = 1/2$.

Theorem 1.1. [21, 22] *For $d \geq 5$, there are positive constants A, D, c, ϵ such that*

$$c_n = A\mu^n [1 + O(n^{-\epsilon})], \tag{6}$$

$$\mathbb{E}_n |\omega(n)|^2 = Dn [1 + O(n^{-\epsilon})], \tag{7}$$

and the rescaled self-avoiding walk converges weakly to Brownian motion:

$$\frac{\omega(\lfloor nt \rfloor)}{\sqrt{Dn}} \Rightarrow B_t. \tag{8}$$

Also [18], as $|x| \rightarrow \infty$,

$$G_{z_c}(x) = c|x|^{-(d-2)} [1 + O(|x|^{-\epsilon})]. \tag{9}$$

The limit in (8) is called a scaling limit. The identification of scaling limits for dimensions $d = 2, 3, 4$ is the grand goal, but the ν problem is a key intermediate objective because $n^{-\nu}\omega(\lfloor nt \rfloor)$ is the candidate sequence for the scaling limit.

If we set up the probability space without imposing the injective condition in the definition of ω , then the mean-square displacement is exactly n , because then the law for ω is that of simple random walk. According to Donsker's Theorem, the scaling limit of simple random walk, with $D = 1$, is also Brownian motion. Thus, in dimensions $d \geq 5$ self-avoiding walk and simple random walk have the same scaling limit. When different models have the same scaling limit, we say the models are in the same *universality class*. One of the goals of mathematical statistical mechanics is to classify universality classes.

Theorem 1.1 will not hold with $\nu = 1/2$ for dimensions four and less. There is a conjecture going back to [2] that, for $d = 4$,

$$c_n \sim A\mu^n(\log n)^{1/4}, \quad \mathbb{E}_n|\omega(n)|^2 \sim Dn(\log n)^{1/4}. \quad (10)$$

This and the next paragraph motivates our interest in four dimensions.

In dimension $d = 3$, nothing is known rigorously about the ν problem. The existence of ν is not proved. It is not known that self-avoiding walk moves away from the origin faster than simple random walk, $\mathbb{E}_n|\omega(n)|^2 \geq n$, nor is it known that self-avoiding walk is slower than ballistic, $\mathbb{E}_n|\omega(n)/n|^2 \rightarrow 0$. In dimension $d = 2$, there is the same basic lack of control as in $d = 3$, but the good news is that there is a candidate for the scaling limit, which tells us that if ν exists it should be equal to $3/4$. In [25], the process known as $\text{SLE}_{8/3}$ is identified as the scaling limit of self-avoiding walk subject to the unproven hypothesis that the scaling limit exists and is conformally invariant.

SLE is a breakthrough discovery because it provides a comprehensive list of possible scaling limits in $d = 2$. It has separated off the issues of existence of limits and universality and made it possible to study candidate limits without first having proved they are limits. On the other hand, theoretical physicists have a profound calculus called the *Renormalisation Group* (RG) that naturally explains when different models are in the same universality class and that can also prove the existence of limits. We will follow this path. RG, in the form that we will develop, was largely invented by Ken Wilson [28, 30, 29]. RG as a rigorous tool originated with [1, 15]. Later developments are reviewed in [6]. The hierarchical lattices mentioned earlier have special properties that greatly simplify RG. The $n(\log n)^{1/4}$ growth of (10) has been shown to hold for continuous-time weakly self-avoiding walk on a four dimensional hierarchical lattice in [4, 8, 9]. Very recently, the corresponding Abelian ν problem has been solved in [20] for a *discrete-time* model on the hierarchical lattice.

1.2. Continuous-time weakly self-avoiding walk and the main result. We now describe a probability law on a space of maps $X : [0, T] \rightarrow \mathbb{Z}^d$. We use the word “time” for the parameter $t \in [0, T]$, but

as for the discrete-time case there is a different space and law for each T . It is not a stochastic process which reveals more about itself as “time” advances, so it is better to think of the interval $[0, T]$ as a continuous parametrisation of a path in \mathbb{Z}^d .

Fix a dimension $d \geq 4$. Let X be the continuous-time simple random walk on \mathbb{Z}^d with $\text{Exp}(1)$ holding times and right-continuous sample paths. In other words, the walk takes its nearest neighbour steps at the events of a rate-1 Poisson process. Let P_a and E_a be the probability law and the expectation for this process started in $X(0) = a$. The local time at x up to time T is given by

$$L_{x,T} = \int_0^T \mathbb{1}_{X(s)=x} ds, \tag{11}$$

and we can measure the amount of self-intersection experienced by X up to time T by

$$\begin{aligned} I(0, T) &= \int_0^T ds_1 \int_0^T ds_2 \mathbb{1}_{X(s_1)=X(s_2)} \\ &= \int_0^T ds_1 \int_0^T ds_2 \sum_{x \in \mathbb{Z}^d} \mathbb{1}_{X(s_1)=x} \mathbb{1}_{X(s_2)=x} = \sum_{x \in \mathbb{Z}^d} L_{x,T}^2. \end{aligned} \tag{12}$$

Then, for $g > 0$, $e^{-gI(0,T)}$ is our substitute for the indicator function supported on self-avoiding X . For $g > 0$, we define a new probability law

$$P_{g,a}(A) = E_a(e^{-gI(0,T)} \mathbb{1}_A) / E_a(e^{-gI(0,T)}) \tag{13}$$

on measurable subsets A of the set of all maps $X : [0, T] \rightarrow \mathbb{Z}^d$ with $X(0) = a$. For this model there is a ν problem¹, but only the Abelian ν problem for \mathbb{Z}^d is currently within the reach of the methods of this paper.

The continuous-time weakly self-avoiding walk two-point function is defined by

$$G_{g,\nu}(a, b) = \int_0^\infty E_a(e^{-gI(0,T)} \mathbb{1}_{X(T)=b}) e^{-\nu T} dT, \tag{14}$$

where ν is a parameter (possibly negative) which is chosen in such a way that the integral converges. For $p \geq 0$ define

$$\chi_g^{(p)}(\nu) = \sum_{b \in \mathbb{Z}^d} G_{g,\nu}(a, b) |b - a|^p. \tag{15}$$

By subadditivity, cf. [26], there exists $\nu_c = \nu_c(g)$ such that $\chi_g^{(0)}(\nu) < \infty$ if and only if $\nu > \nu_c$. We call this ν_c the *critical value* of ν . Our main result is the following theorem.

¹solved on the hierarchical lattice for g small in [4, 8, 9]

Theorem 1.2. *Let $d \geq 4$. There exists $g_{\max} > 0$ such that for each $g \in [0, g_{\max}]$ there exists $c_g > 0$ such that as $|a - b| \rightarrow \infty$,*

$$G_{g, \nu_c(g)}(a, b) = \frac{c_g}{|a - b|^{d-2}} (1 + o(1)). \tag{16}$$

This is the analogue of (9) in Theorem 1.1, but now including dimension $d = 4$. There are no log corrections. Log corrections are only expected in the singular behaviour of $\chi_g^{(p)}(\nu)$ as $\nu \downarrow \nu_c$ for $p \geq 0$. The case $g = 0$ is a standard fact about simple random walk; our proof is given for case $g > 0$.

2. Finite Volume Approximation

In this section we describe the first step in our proof, which is to approximate the *infinite volume* \mathbb{Z}^d by *finite volume*, namely a discrete torus.

We do not make explicit the dependence on g , which is fixed and positive. Let $R \geq 3$ be an integer, and let $\Lambda = \mathbb{Z}^d / R\mathbb{Z}^d$ denote the discrete torus of side R . For $a, b \in \Lambda$, let

$$G_{\Lambda, \nu}(a, b) = \int_0^\infty E_{a, \Lambda} \left(e^{-gI(0, T)} \mathbb{1}_{X(T)=b} \right) e^{-\nu T} dT, \tag{17}$$

where $E_{a, \Lambda}$ denotes the continuous-time simple random walk on Λ , started from a . The following theorem shows that it is possible to study the critical two-point function in the double limit in which first $\Lambda \uparrow \mathbb{Z}^d$ and then $\nu \downarrow \nu_c$. We will follow this route, focusing our analysis on the subcritical finite volume model with sufficient uniformity to take the limits.

Theorem 2.1. *Let $d \geq 1$ and $\nu \geq \nu_c$. Then*

$$G_\nu(a, b) = \lim_{\nu' \downarrow \nu} \lim_{\Lambda \uparrow \mathbb{Z}^d} G_{\Lambda, \nu'}(a, b). \tag{18}$$

3. Integral Representation

The next step in the proof is to represent the two-point function in finite volume by an integral that we will approximate by a Gaussian integral.

Recall that Λ denotes a discrete torus in \mathbb{Z}^d . Given $\varphi \in \mathbb{C}^\Lambda$ and writing $\varphi = (\varphi_x)$, $x \in \Lambda$, we write $d\varphi_x$ and $d\bar{\varphi}_x$ for the differentials, we fix a choice of the square root $\sqrt{2\pi i}$, and we set

$$\psi_x = \frac{1}{\sqrt{2\pi i}} d\varphi_x, \quad \bar{\psi}_x = \frac{1}{\sqrt{2\pi i}} d\bar{\varphi}_x. \tag{19}$$

Define the differential forms

$$\tau_x = \varphi_x \bar{\varphi}_x + \psi_x \wedge \bar{\psi}_x \quad (x \in \Lambda), \tag{20}$$

and

$$\tau_{\Delta,x} = \frac{1}{2} \left(\varphi_x (-\Delta \bar{\varphi})_x + (-\Delta \varphi)_x \bar{\varphi}_x + \psi_x \wedge (-\Delta \bar{\psi})_x + (-\Delta \psi)_x \wedge \bar{\psi}_x \right), \quad (21)$$

where Δ is the lattice Laplacian on Λ defined by $\Delta \varphi_x = \sum_{y:|y-x|=1} (\varphi_y - \varphi_x)$, and \wedge is the standard wedge product. From now on, for differential forms u, v , we will abbreviate by writing $uv = u \wedge v$. In particular $\psi_x \psi_y = -\psi_y \psi_x$ and likewise $\bar{\psi}_x$ anticommutes with $\bar{\psi}_y$ and with ψ_y . The proof of the following proposition is given in [4, 9]; see also [10] for a self-contained proof.

Proposition 3.1. *Given $g > 0$, let ν be such that $G_{\Lambda,\nu}(a, b)$ is finite. Then*

$$G_{\Lambda,\nu}(a, b) = \int_{\mathbb{C}^\Lambda} e^{-\sum_{x \in \Lambda} (\tau_{\Delta,x} + g\tau_x^2 + \nu\tau_x)} \bar{\varphi}_a \varphi_b. \quad (22)$$

The definition of an integral such as the right-hand side of (22) is as follows:

1. Expand the entire integrand in a power series about its degree-zero part (this is a *finite* sum due to the anti-commutativity of the wedge product, and the order of factors in the resulting products is immaterial due to the even degree), e.g.,

$$e^{-\nu\tau_x} = e^{-\nu\varphi_x \bar{\varphi}_x - \frac{\nu}{2\pi i} d\varphi_x d\bar{\varphi}_x} = e^{-\nu\varphi_x \bar{\varphi}_x} \left(1 - \frac{\nu}{2\pi i} d\varphi_x d\bar{\varphi}_x \right). \quad (23)$$

In general, any function of the differentials is defined by its formal power series about its degree-zero part.

2. Keep only terms with one factor $d\varphi_x$ and one $d\bar{\varphi}_x$ for each $x \in \Lambda$, write $\varphi_x = u_x + iv_x$, $\bar{\varphi}_x = u_x - iv_x$ and similarly for the differentials.
3. Rearrange the differentials to $\prod_{x \in \Lambda} du_x dv_x$, using the anti-commutativity of the wedge product.
4. Finally, perform the Lebesgue integral over $\mathbb{R}^{2|\Lambda|}$.

This is explained in more detail in [10]. These integrals have the remarkable self-normalisation property that

$$\int e^{-\sum_{x \in \Lambda} (a_x \tau_{\Delta,x} + b_x \tau_x^2 + c_x \tau_x)} = 1, \quad a_x \geq 0, b_x > 0, c_x \in \mathbb{R}, x \in \Lambda. \quad (24)$$

Self-contained proofs of this, and of generalisations, can be found in [10]. The variables φ_x and the forms ψ_x are called *fields*.

4. Quadratic or Gaussian Approximation

The integral representation of Proposition 3.1 opens a natural route for approximation by non-interacting walk with different parameters. To do this we split the exponent $\tau_{\Delta,x} + g\tau_x^2 + \nu\tau_x$ in (22) into a part which is quadratic in the variables φ and a remainder. When the remainder is ignored the rest of the integral becomes Gaussian and the Gaussian integral represents a non-interacting walk. It is important not to assume that the best approximation is the quadratic terms $\tau_{\Delta,x} + \nu\tau_x$. We even want to allow τ_{Δ} to be divided up. To see what a different coefficient in front of τ_{Δ} means we make the change of variable $\varphi_x \mapsto \sqrt{1+z_0}\varphi_x$, with $z_0 > -1$. This gives

$$G_{\Lambda,\nu}(a, b) = (1+z_0) \int_{\mathbb{C}^{\Lambda}} e^{-\sum_{x \in \Lambda} ((1+z_0)\tau_{\Delta,x} + g(1+z_0)^2\tau_x^2 + \nu(1+z_0)\tau_x)} \bar{\varphi}_a \varphi_b, \tag{25}$$

where the Jacobian is contained in the transformation of $\psi, \bar{\psi}$. Then, for any $m^2 \geq 0$, simple algebra allows us to rewrite this as

$$G_{\Lambda,\nu}(a, b) = (1+z_0) \int e^{-S(\Lambda) - \tilde{V}_0(\Lambda)} \bar{\varphi}_a \varphi_b, \tag{26}$$

where

$$S(\Lambda) = \sum_{x \in \Lambda} (\tau_{\Delta,x} + m^2\tau_x), \tag{27}$$

$$\tilde{V}_0(\Lambda) = \sum_{x \in \Lambda} (g_0\tau_x^2 + \nu_0\tau_x + z_0\tau_{\Delta,x}), \tag{28}$$

$$g_0 = (1+z_0)^2g, \quad \nu_0 = (1+z_0)\nu_c, \quad m^2 = (1+z_0)(\nu - \nu_c), \tag{29}$$

and ν_c was defined below (15). The two-point function $G_{\Lambda,\nu}(a, b)$ in (26) does not depend on (z_0, m^2) so, in the next theorem, these are free parameters that do not get fixed until Section 12. In view of Theorem 2.1 and Proposition 3.1, to prove Theorem 1.2 it suffices to prove the following theorem.

Theorem 4.1. *Let $d \geq 4$. There exists $g_{\max} > 0$ such that for each $g \in [0, g_{\max}]$ there exist $c(g) > 0$ such that as $|a - b| \rightarrow \infty$,*

$$\lim_{\nu \downarrow \nu_c} \lim_{\Lambda \uparrow \mathbb{Z}^d} (1+z_0) \int_{\mathbb{C}^{\Lambda}} e^{-S(\Lambda) - \tilde{V}_0(\Lambda)} \bar{\varphi}_a \varphi_b = \frac{c(g)}{|a - b|^{d-2}} (1 + o(1)). \tag{30}$$

To prove Theorem 4.1, we study the integral on the left-hand side via a renormalisation group analysis, without making further direct reference to its connection with self-avoiding walks. In order to calculate this integral we define, for $\sigma \in \mathbb{C}$,

$$V_0(\Lambda) = \tilde{V}_0(\Lambda) + \sigma \bar{\varphi}_a + \bar{\sigma} \varphi_b \tag{31}$$

and use

$$\int_{\mathbb{C}^{\Lambda}} e^{-S(\Lambda) - \tilde{V}_0(\Lambda)} \bar{\varphi}_a \varphi_b = - \frac{\partial}{\partial \sigma} \frac{\partial}{\partial \bar{\sigma}} \Big|_0 \int_{\mathbb{C}^{\Lambda}} e^{-S(\Lambda) - V_0(\Lambda)}. \tag{32}$$

We will call σ an *external field*.

5. Forms and Test Functions

In this section we introduce notation for handling the differential forms that appear in Theorem 4.1. We will write *form* in place of “differential forms” from now on. We focus on dimension $d = 4$, but leave d in various formulas since 4 can also appear for other reasons.

5.1. The space \mathcal{N} . A form is a polynomial in $\psi, \bar{\psi}$ with coefficients that are functions of $(\varphi, \sigma) \in \mathbb{C}^\Lambda \times \mathbb{C}$.

Given $\sigma \in \mathbb{C}$ we define $\sigma_1 = \sigma$ and $\sigma_2 = \bar{\sigma}$ so that σ can be identified with a function $\sigma : \{1, 2\} \rightarrow \mathbb{C}$. Similarly, let $\Lambda_2 = \Lambda \times \{1, 2\}$ so that given $\varphi \in \mathbb{C}^\Lambda$ we have the function on $x = (s, i) \in \Lambda_2$ defined by

$$\phi_x = \begin{cases} \varphi_s & i = 1, \\ \bar{\varphi}_s & i = 2. \end{cases} \tag{33}$$

Since ϕ and φ are in one to one correspondence and since we are only interested in functions on Λ_2 that arise from some φ we write $\phi \in \mathbb{C}^\Lambda$.

Forms are elements of the algebra \mathcal{N} whose generators are the degree one forms $(\psi_x, \bar{\psi}_x, x \in \Lambda)$ subject to the relations that all generators mutually anticommute. For $x = (s, i) \in \Lambda_2$, we write

$$\psi_x = \begin{cases} \psi_s & i = 1, \\ \bar{\psi}_s & i = 2. \end{cases} \tag{34}$$

Then we introduce the space $\Lambda^* = \cup_{q=0}^\infty \Lambda_2^q$ of all sequences in Λ_2 with finitely many terms so that every monomial in ψ can be written in the form, for some $y \in \Lambda^*$,

$$\psi^y = \begin{cases} 1 & \text{if } q = 0 \\ \psi_{y_1} \cdots \psi_{y_q} & \text{if } q \geq 1. \end{cases} \tag{35}$$

The $q = 0$ term in Λ^* is a set consisting of a single element called the “empty sequence”, which by definition has length zero. Given a sequence $y \in \Lambda^*$, $q = q(y)$ is the length of the sequence and $y! = q(y)!$. Every element of \mathcal{N} has the form

$$F = F(\phi, \sigma) = \sum_{y \in \Lambda^*} \frac{1}{y!} F_y(\phi, \sigma) \psi^y. \tag{36}$$

Given $x = (x_1, \dots, x_p) \in \Lambda_2^p$ and $z = (z_1, \dots, z_r) \in \{1, 2\}^r$, we write

$$F_{x,y,z}(\phi, \sigma) = \frac{\partial^p}{\partial \phi_{x_p} \cdots \partial \phi_{x_1}} \frac{\partial^r}{\partial \sigma_{z_r} \cdots \partial \sigma_{z_1}} F_y(\phi, \sigma). \tag{37}$$

For $X \subset \Lambda$, we define $\mathcal{N}(X)$, which is a subspace of \mathcal{N} , by

$$\mathcal{N}(X) = \{F \in \mathcal{N} : F_{x,y} = 0 \text{ if any component of } x, y \text{ is not in } X\}. \tag{38}$$

For example $\tau_x \in \mathcal{N}(\{x\})$ and $\tau_{\Delta,x} \in \mathcal{N}(X)$ where $X = \{y : |y - x| \leq 1\}$.

By introducing

$$\phi^y = \begin{cases} 1 & \text{if } q = 0 \\ \phi_{y_1} \cdots \phi_{y_q} & \text{if } q \geq 1, \end{cases} \tag{39}$$

we write the formal Taylor expansion of $F(\phi + \xi)$ in powers of ξ and σ as

$$\sum_{x,y \in \Lambda^*, z \in \{1,2\}^*} \frac{1}{x!y!z!} F_{x,y,z}(\phi, 0) \xi^x \psi^y \sigma^z. \tag{40}$$

Functions $f : \Lambda^* \times \Lambda^* \times \{1,2\}^* \rightarrow \mathbb{C}$ are called *test functions*. We define a pairing between elements of \mathcal{N} and the set of test functions as follows: for a test function f , for $\phi \in \mathbb{C}^\Lambda$, let

$$\langle F, f \rangle_\phi = \sum_{x,y \in \Lambda^*, z \in \{1,2\}^*} \frac{1}{x!y!z!} F_{x,y,z}(\phi, 0) f_{x,y,z}. \tag{41}$$

For example, let $F = \varphi_k$ and $F' = \varphi_0 + (k \cdot \nabla \varphi)_0$. Then

$$\langle F, f \rangle_0 = f_k, \quad \langle F', f \rangle_0 = f_0 + (k \cdot \nabla f)_0, \tag{42}$$

and more generally when $\phi = 0$ the effect of the pairing is to replace fields by the test function.

5.2. Local polynomials and localisation. For a function $f : \Lambda \rightarrow \mathbb{C}$ and e a unit vector in \mathbb{Z}^d we define the *finite difference derivative* $(\nabla_e f)_x = f(x + e) - f(x)$. Repeated differences such as $(\nabla_e \nabla_{e'} f)_x$ are called *derivatives*.

A *local monomial* is a product of finitely many fields and derivatives of fields such as $M = \psi \bar{\psi} \nabla_e \bar{\varphi}$. Using this example to introduce a general notation, given $x \in \Lambda$ let $M_x = \psi_x \bar{\psi}_x (\nabla_e \bar{\varphi})_x$, and given $X \subset \Lambda$ let $M(X) = \sum_{x \in X} M_x$. *Local polynomials* are finite sums of local monomials with constant coefficients.

An important example of a local polynomial is

$$V = g\tau^2 + \nu\tau + z\tau_{\Delta,x} + \lambda\mathbb{1}_a \bar{\sigma}\varphi + \lambda\mathbb{1}_b \sigma \bar{\varphi} + (q/2)(\mathbb{1}_a + \mathbb{1}_b) \bar{\sigma}\sigma, \tag{43}$$

which extends the local polynomial of (31) by the addition of the $\bar{\sigma}\sigma$ term. The indicator function $\mathbb{1}_a : \Lambda \rightarrow \{0,1\}$ equals 1 when evaluated on a and is zero otherwise. The parameters (g, ν, z, λ, q) are called *coupling constants*.

Euclidean symmetry: The lattice \mathbb{Z}^d has automorphisms $E : \mathbb{Z}^d \rightarrow \mathbb{Z}^d$. An example for $d = 1$ is $E x = 1 - x$. By letting an automorphism E act on the spatial labels on fields, $\varphi_x \mapsto \varphi_{E x}$, E induces an action, $E : \mathcal{N} \rightarrow \mathcal{N}$. A local polynomial P is *Euclidean invariant* if automorphisms of \mathbb{Z}^d that fix x also fix P_x . For example, $\psi \bar{\psi} \nabla_e \bar{\varphi}$ is not Euclidean invariant because there is a reflection that changes φ_{x+e} into φ_{x-e} so that $(\nabla_e \bar{\varphi})_x \mapsto (\nabla_{-e} \bar{\varphi})_x$. On the other hand, the term τ_Δ in (43) is a Euclidean invariant local monomial.

Gauge invariance: A local polynomial is gauge invariant if it is invariant under the *gauge flow*: $(\sigma, \varphi) \rightarrow (e^{i\theta}\sigma, e^{i\theta}\varphi)$. Thus V of (43) is gauge invariant.

Supersymmetry: There is an antiderivation $\hat{Q} : \mathcal{N} \rightarrow \mathcal{N}$ characterised by

$$\hat{Q}\varphi_x = \psi_x, \quad \hat{Q}\psi_x = -\varphi_x, \quad \hat{Q}\bar{\varphi}_x = \bar{\psi}_x, \quad \hat{Q}\bar{\psi}_x = \bar{\varphi}_x. \tag{44}$$

An element of $F \in \mathcal{N}$ is said to be *supersymmetric* if $\hat{Q}F = 0$. The terms $\tau, \tau_\Delta, \tau^2$ in V are supersymmetric local monomials. The forms $\bar{\sigma}\varphi, \sigma\bar{\varphi}, \bar{\sigma}\sigma$ are gauge invariant, but not supersymmetric. It is straightforward to check that \hat{Q}^2 generates the gauge flow. Therefore supersymmetry implies gauge invariance. Further details can be found in [10].

The pairing (41) defines $F \in \mathcal{N}$ as a linear function, $f \mapsto \langle F, f \rangle_0$, on test functions. The subscript means that we set $\phi = 0$. Let Π be a set of test functions. Two elements F_1 and F_2 of \mathcal{N} are equivalent when they define the same linear function on Π . We say they are *separated* if they are not equivalent.

Example 1. Let Π be the set of test functions that are linear in their Λ arguments. Fix a point $k \in \mathbb{Z}^d$. Let $F = \varphi_k$, and let $F' = \varphi_0 + (k \cdot \nabla\varphi)_0$. Then F and F' are equivalent because a linear test function $f(x) = a + b \cdot x$ cannot separate them, since by (42),

$$\langle F, f \rangle = a + b \cdot k = \langle F', f \rangle. \tag{45}$$

To avoid confusion let us emphasise that two different contexts for “polynomial” are in use: a test functions can be a polynomial in $x \in \Lambda$, while local polynomials are polynomial in fields.

The choice for Π in this example is not the one we want. The details in the definition given below are less important than the objective of the definition, which is that Π should be a minimal space of test functions that separates the terms in (43).

We define Π to be the set of test functions $f(x, y, z)$ that are polynomial in the Λ arguments of $(x, y) \in \Lambda^* \times \Lambda^*$ with restrictions on degree listed below. For $f \in \Pi$, as a polynomial in the x, y components in Λ :

1. The restriction of f to (x, y, z) with $r(z) = 0$ has total degree at most $d - p(x)[\phi] - q(y)[\phi]$; $f(x, y, z) = 0$ when $d - p(x)[\phi] - q(y)[\phi] < 0$. Here

$$[\phi] = (d - 2)/2. \tag{46}$$

For dimension $d = 4$, $[\phi] = 1$.

2. The restriction of f to (x, y, z) with $r(z) = r \in \{1, 2\}$ has total degree at most $r - p(x) - q(y)$; $f(x, y, z) = 0$ if $r - p(x) - q(y) < 0$ or $r > 2$.

Let \mathcal{V} be the vector space of gauge invariant local polynomials that are separated by Π and, for $X \subset \Lambda$, let $\mathcal{V}(X) = \{P(X) : P \in \mathcal{V}\}$. The following proposition associates to any form $F \in \mathcal{N}$ an equivalent local polynomial in $\mathcal{V}(X)$ [12].

Proposition 5.1. *For nonempty $X \subset \mathbb{Z}^d$ there exists a linear map $\overline{\text{Loc}}_X : \mathcal{N} \rightarrow \mathcal{V}(X)$ such that*

$$(a) \quad \langle \overline{\text{Loc}}_X F, f \rangle_0 = \langle F, f \rangle_0 \quad \text{for } f \in \Pi, F \in \mathcal{N}, \tag{47}$$

$$(b) \quad E(\overline{\text{Loc}}_X F) = \overline{\text{Loc}}_{EX}(EF) \quad \text{for automorphisms } E : \mathbb{Z}^d \rightarrow \mathbb{Z}^d, F \in \mathcal{N}, \tag{48}$$

$$(c) \quad \overline{\text{Loc}}_{X'} \circ \overline{\text{Loc}}_X = \overline{\text{Loc}}_{X'} \quad \text{for } X, X' \subset \Lambda. \tag{49}$$

Let $\mathcal{V}_H \subset \mathcal{V}$ be the subspace generated by monomials that are not divisible by σ or $\bar{\sigma}$, and let $\mathcal{V}_O \subset \mathcal{V}$ be the subspace generated by monomials that are divisible by σ or $\bar{\sigma}$. Then $\mathcal{V} = \mathcal{V}_H \oplus \mathcal{V}_O$, and on this direct sum we define

$$\text{Loc}_X = \overline{\text{Loc}}_X \oplus \overline{\text{Loc}}_{X \cap \{a,b\}}, \tag{50}$$

where $\overline{\text{Loc}}_\emptyset$ is interpreted as zero. Symmetry considerations for the integral representation (22) restrict the domain of $\overline{\text{Loc}}$ in our applications so that its range reduces to polynomials of the form V as in (43).

6. Gaussian Integration

6.1. The super-expectation. For a $\Lambda \times \Lambda$ matrix A , we define

$$S_A(\Lambda) = \sum_{x,y \in \Lambda} \left(\varphi_x A_{xy} \bar{\varphi}_x + \psi_x A_{xy} \bar{\psi}_y \right). \tag{51}$$

When $A = m^2 - \Delta$ this is the same as $S(\Lambda)$ which was defined in (27). Let C be a positive-definite $\Lambda \times \Lambda$ matrix. Then $A = C^{-1}$ exists. We introduce the notation

$$\mathbb{E}_C F = \int_{\mathbb{C}^\Lambda} e^{-S_A(\Lambda)} F, \tag{52}$$

for F a form in \mathcal{N} . The integral is defined as described under Proposition 3.1. We call C the covariance because $\mathbb{E}_C \bar{\phi}_a \phi_b = C_{ab}$. More generally, if F is a form of degree zero, i.e., a function of ϕ , then $\mathbb{E}_C F$ is a standard Gaussian expectation for a complex valued random variable ϕ with covariance C [10].

We define a space \mathcal{N}^\times in the same way as \mathcal{N} is defined, but with ϕ doubled to (ϕ, ξ) so that (ϕ, ψ) doubles to the pair $(\phi, \psi), (\xi, \eta)$ with $\eta = (2\pi i)^{-1/2} d\xi$. The external field σ is not doubled. We define $\theta : \mathcal{N} \rightarrow \mathcal{N}^\times$ by

$$(\theta F)(\phi, \xi) = \sum_{y \in \Lambda^*} \frac{1}{y!} F_y(\phi + \xi)(\psi + \eta)^y. \tag{53}$$

We write $\mathbb{E}_C \theta F$ for the element of \mathcal{N} obtained when the integral over \mathbb{C}^Λ in \mathbb{E}_C applies *only* to (ξ, η) . In the general case where F is a form this is not standard probability theory, because $\mathbb{E}_C \theta F$ takes values in \mathcal{N} . To keep this in mind we call this a *super-expectation*. The variables and forms (ξ, η) that are integrated out are called *fluctuation fields*.

6.2. Finite-range decomposition of covariance. Suppose C and $C_j, j = 1, \dots, N$, are positive-definite $\Lambda \times \Lambda$ matrices such that

$$C = \sum_{j=1}^N C_j. \tag{54}$$

Let $C' = \sum_{k=2}^N C_k$. Then, as in the standard theory of Gaussian random variables, the \mathbb{E}_C expectation can be performed progressively:

$$\mathbb{E}_C F = \mathbb{E}_{C'+C_1} F = \mathbb{E}_{C'} (\mathbb{E}_{C_1} \theta F). \tag{55}$$

For further details, see [12].

From now on we work with $C = (m^2 - \Delta)^{-1}$, where Δ is the finite difference Laplacian on the periodic lattice Λ . Given any sufficiently large dyadic integer L , there exists a decomposition $C = \sum_{j=1}^N C_j$ such that C_j is positive-definite and

$$C_j(x, y) = 0 \quad \text{if} \quad |x - y| \geq L^j/2. \tag{56}$$

This is called the *finite range* property. The existence of such a decomposition is established in [7] for the case where Λ is replaced by \mathbb{Z}^d . In [6, Lecture 2] it is briefly explained how the decomposition for the periodic Λ case is obtained from the \mathbb{Z}^d case, for Λ a torus of side L^N . To accommodate this restriction on the side of Λ the infinite volume limit in Theorem 4.1 is taken with a sequence of tori with sides $L^N, N \in \mathbb{N}$.

We conclude this section with an informal discussion of scaling estimates that guide the proof. Equation (55) says that F , which depends on a field with covariance C , can be replaced by $\mathbb{E}_{C_1} \theta F$, which depends on a field characterised by the covariance C' . Repeating this operation j times will replace F by a new F that depends on a *field at scale j* characterised by the covariance $\sum_{k=j+1}^N C_k$. According to estimates in [7], this sum is dominated by the first term which satisfies

$$|\nabla_x^\alpha \nabla_y^\beta C_{j+1}(x, y)| \leq \text{const } L^{-2j[\phi] - |\alpha|_1 j - |\beta|_1 j}, \tag{57}$$

where the symbol $[\phi]$, which is called the *dimension* of the field, was defined in (46). The typical field at scale j behaves like “half a covariance,” and in particular the standard deviation of φ_x is $\approx L^{-j[\phi]}$. Furthermore, the estimate on derivatives in (57) says that typical fields at scale j are roughly constant over distances of order L^j .

We can now explain why the terms in V as defined by (43) play a pre-eminent role. For a cube B of side L^j , which contains L^{dj} points,

$$\sum_{x \in B} \varphi_{j,x}^p \approx L^{(d-p[\phi])j}. \tag{58}$$

In the case of $d = 4$, for which $[\phi] = 1$, this scales down when $p > 4$ and φ^p is said to be *irrelevant*. The power $p = 4$ neither decays nor grows, and is called

marginal. Powers $p < 4$ grow with the scale, and are called *relevant*. Since the derivatives in (57) provide powers of L , the monomial $\varphi(-\Delta)\bar{\varphi}$ is marginal. Thus $\tau, \tau_\Delta, \tau^2$ are the supersymmetric marginal and relevant monomials.

6.3. Progressive integration. To prove Theorem 4.1 using (32) we have to calculate

$$\int_{\mathbb{C}^\Lambda} e^{-S(\Lambda)-V_0(\Lambda)} = \mathbb{E}_C e^{-V_0(\Lambda)}, \tag{59}$$

where V_0 is given by (31). This V_0 equals V as defined in (43), with (g, ν, z, λ, q) replaced by $(g_0, \nu_0, z_0, \lambda_0, q_0)$ with

$$q_0 = 0, \quad \lambda_0 = 1. \tag{60}$$

Sections 6.1 and 6.2 have taught us that we can evaluate $\mathbb{E}_C e^{-V_0(\Lambda)}$ by the following iteration: let

$$Z_0 = e^{-V_0(\Lambda)}. \tag{61}$$

Inductively define $Z_j, j = 0, \dots, N$, by

$$Z_{j+1} = \mathbb{E}_{C_{j+1}} \theta Z_j. \tag{62}$$

Then

$$\mathbb{E}_C e^{-V_0(\Lambda)} = Z_N. \tag{63}$$

Therefore the proof of Theorem 4.1 now depends on the analysis of the sequence Z_j . Our proof will depend on showing that the Z_j simplify as j increases. In fact, in the next section we will see that they become more Gaussian, in the sense that the $g\tau^2$ term becomes smaller. The index j will be called a *scale*.

7. Perturbation Theory and Flow Equations

In this section we start to prove that Z_j becomes more Gaussian as j increases. To do this we adapt to our particular setting a perturbative calculation of the kind that appears in [30].

For $X \subset \Lambda$ and V as defined in (43), define

$$I_{j,X}(V) = e^{-V(X)} \left(1 + \frac{1}{2} W_j(V, X)\right), \tag{64}$$

where

$$W_j(V, X) = (1 - \text{Loc}_X) F_{w_j}(V(X), V(\Lambda)) \tag{65}$$

with

$$w_j = \sum_{i=1}^j C_i, \tag{66}$$

$$F_{w_j}(V(X), V(\Lambda)) = \sum_{n \geq 1} \frac{1}{n!} (D_R^n V(X)) w_j^n (D_L^n V(\Lambda)), \quad X \subset \Lambda \setminus \{a, b\}; \tag{67}$$

the latter sum truncates at $n = 4$ due to our quartic interaction. The symbols D_R and D_L denotes right and left differentiation with respect to fields. The “left/right” is to specify signs, but this and the precise definition are not of immediate importance, so we just give an example. If X contains a or b there is an additional combinatorial factor of 2 multiplying terms in $F_{w_j}(V(X), V(\Lambda \setminus X))$ that are linear in $\sigma, \bar{\sigma}$.

Example 2. For $V = \psi\bar{\psi}$ and $X = \{x\}$, $(D_R^n V(X))w_j^n(D_L^n V(\Lambda))$ equals

$$\begin{cases} \sum_{y \in \Lambda} (\psi_x w_j(x, y)\bar{\psi}_y + \bar{\psi}_x w_j(x, y)\psi_y) & n = 1 \\ -\sum_{y \in \Lambda} w_j^2(x, y) & n = 2. \end{cases} \tag{68}$$

When $j = 0$, $I_{j,X}(V) = e^{-V(\Lambda)}$ because $w_0 = 0$. Therefore we can choose the coupling constants to make it equal to Z_0 . Furthermore, $I_{j,X}(V)$ has the martingale-like property exhibited in Proposition 7.1, which says that integrating out the fluctuation field ξ_{j+1} is approximately the same as changing the coupling constants in V to new coupling constants called $(g_{\text{pt}}, \nu_{\text{pt}}, z_{\text{pt}}, \lambda_{\text{pt}}, q_{\text{pt}})$. The formulas for the new coupling constants are called *perturbative flow equations*.

Proposition 7.1. *As a formal power series in (g, ν, z, λ, q) ,*

$$\mathbb{E}_{C_{j+1}} I_{j,\Lambda}(V) = I_{j+1,\Lambda}(V_{\text{pt}}) \pmod{(g, \nu, z, \lambda, q)^3}, \tag{69}$$

where

$$V_{\text{pt}} = V_{\text{pt}}(V) \tag{70}$$

has the same form (43) as V , with (g, ν, z, λ, q) replaced by

$$g_{\text{pt}} = g - c_g g^2 + r_{g,j}^{\text{pt}}, \tag{71}$$

$$\nu_{\text{pt}} = \nu_+ + r_{\nu,j}^{\text{pt}}, \tag{72}$$

$$z_{\text{pt}} = z + r_{z,j}^{\text{pt}}, \tag{73}$$

$$\lambda_{\text{pt}} = \left(1 + \sum_{y \in \Lambda} (\nu_+ w_{j+1}(0, y) - \nu w_j(0, y)) \right) \lambda, \tag{74}$$

$$q_{\text{pt}} = q + \lambda^2 C_{j+1}(a, b), \tag{75}$$

where $c_g > 0$, $\nu_+ = \nu + 2gC_{j+1}(0, 0)$, and $r_{g,j}^{\text{pt}}, L^{2j}r_{\nu,j}^{\text{pt}}, r_{z,j}^{\text{pt}}$ are computable uniformly bounded homogeneous polynomials of degree 2 in (g, ν, z) . There are g^2 terms in $r_{g,j}^{\text{pt}}$, but they are summable in j and therefore do not overpower $c_g g^2$.

The β function. The right hand side of (71) is known as the β function. The simpler recursion obtained by setting $r_{\nu,j}^{\text{pt}} = 0$, namely

$$\bar{g}_{j+1} = \bar{g}_j - c_g \bar{g}_j^2, \quad \bar{g}_0 = g_0, \tag{76}$$

creates a sequence \bar{g}_j that tends to zero like j^{-1} as $j \rightarrow \infty$. The sequence Z_j becomes more Gaussian due to the famous observation, known as *infra-red asymptotic freedom*, that (76) controls the behaviour of the more complex recursion of Proposition 7.1 and drives the τ^2 term to zero.

8. The Renormalisation Group Map

The problem with the second order perturbative calculation in Section 7 is that the error is not only of order 3 in the coupling constants, but it also fails to be uniform in the volume Λ . The remedy is not to work with $I_{j,\Lambda}$, but with $\prod_{B \subset \Lambda} I_{j,B}$ where B is a cube and the allowed cubes pave Λ . The idea is that by choosing the side of B to be bigger than the range of C_{j+1} , we can take advantage of independence of cubes that do not touch to more or less use our perturbation theory with Λ replaced by individual cubes. This idea requires a systematic organisation which we describe in this section.

8.1. Scales and the circle product. Let $L \geq 3$ be an integer. Let $R = L^N$, and let $\Lambda = \mathbb{Z}^d / (R\mathbb{Z}^d)$.

Definition 1. (a) *Blocks.* For each $j = 0, 1, \dots, N$, the torus Λ is paved in a natural way by L^{N-j} disjoint d -dimensional cubes of side L^j . The cube that contains the origin has the form (for L odd)

$$\left\{ x \in \Lambda : |x| \leq \frac{1}{2}(L^j - 1) \right\}, \tag{77}$$

and all the other cubes are translates of this one by vectors in $L^j\mathbb{Z}^d$. We call these cubes *j-blocks*, or *blocks* for short, and denote the set of *j*-blocks by $\mathcal{B}_j = \mathcal{B}_j(\Lambda)$.

(b) *Polymers.* A union of *j*-blocks is called a *polymer* or *j-polymer*, and the set of *j*-polymers is denoted $\mathcal{P}_j = \mathcal{P}_j(\Lambda)$. The size $|X|_j$ of $X \in \mathcal{P}_j$ is the number of *j*-blocks in X .

(c) *Connectivity.* A subset $X \subset \Lambda$ is said to be *connected* if for any two points $x_a, x_b \in X$ there exists a path $(x_i, i = 0, 1, \dots, n) \in X$ with $\|x_{i+1} - x_i\|_\infty = 1$, $x_0 = x_a$ and $x_n = x_b$. According to this definition, a polymer can be decomposed into connected components; we write $\mathcal{C}(X)$ for the set of connected components of X . We say that two polymers X, Y *do not touch* if $\min\{\|x - y\|_\infty : x \in X, y \in Y\} > 1$.

(d) *Small sets.* A polymer $X \in \mathcal{P}_j$ is said to be a *small set* if $|X|_j \leq 2^d$ and X is connected. Let \mathcal{S}_j be the set of all small sets in \mathcal{P}_j .

(e) *Small set neighbourhood.* For $X \subset \Lambda$ let

$$X^* = \bigcup_{Y \in \mathcal{S}_j : X \cap Y \neq \emptyset} Y. \tag{78}$$

The *polymers* of Definition 1 have nothing to do with long chain molecules. This concept has a long history in statistical mechanics going back to the important paper [17].

Proposition 8.1. *Suppose that $X_1, \dots, X_n \in \mathcal{P}_{j+1}$ do not touch each other and let $F_i(X_i) \in \mathcal{N}(X_i)$. The expectation $\mathbb{E}_{C_{j+1}}$ has the factorisation property:*

$$\mathbb{E}_{C_{j+1}} \prod_{i=1}^n F_i(X_i) = \prod_{m=1}^n \mathbb{E}_{C_{j+1}} F_m(X_m). \tag{79}$$

Proof. Gaussian random variables are independent if and only if the off-diagonal part of their covariance matrix vanishes. This generalises to our forms setting, and so the proposition follows from the finite range property of C_{j+1} . □

Given forms F, G defined on \mathcal{P}_j , let

$$(F \circ G)(\Lambda) = \sum_{X \in \mathcal{P}_j} F(X)G(\Lambda \setminus X). \tag{80}$$

This defines an associative product, which is also commutative provided F and G both have even degree.

8.2. The renormalisation group map. Recall that we have defined $I_{j,X}(V)$ in (64). Given a yet-to-be-constructed sequence V_j , for $X \in \mathcal{P}_j$, let

$$I_j(X) = \prod_{B \in \mathcal{B}_j} I_{j,B}(V_j). \tag{81}$$

We have defined V_0 in (31). Let $K_0(X) = \mathbb{1}_{X=\emptyset}$. Then the Z_0 defined in (61) is also given by

$$Z_0 = I_0(\Lambda) = (I_0 \circ K_0)(\Lambda), \tag{82}$$

because $I_{0,\Lambda}(\Lambda) = e^{-V_0(\Lambda)}$ since $w_0 = 0$.

Definition 2. We say that $K : \mathcal{P}_j \rightarrow \mathcal{N}$ has the *component factorisation property* if

$$K(X) = \prod_{Y \in \mathcal{C}(X)} K(Y). \tag{83}$$

Suppose, inductively, that we have constructed (V_j, K_j) where $K_j : \mathcal{P}_j \rightarrow \mathcal{N}$ is such that

- (i) $Z_j = (I_j \circ K_j)(\Lambda)$,
 - (ii) K_j has the component factorisation property,
 - (iii) For $X \in \mathcal{P}_j$, $K_j(X) \in \mathcal{N}(X^*)$.
- (84)

Our objective is to define (V_{j+1}, K_{j+1}) , where $K_{j+1} : \mathcal{P}_{j+1} \rightarrow \mathcal{N}$ has the same properties at scale $j + 1$. Then the action of $\mathbb{E}_{C_{j+1}}\theta$ on Z_j has been expressed as the map:

$$(V_j, K_j) \mapsto (V_{j+1}, K_{j+1}). \tag{85}$$

This map will be constructed next. We call it the *renormalisation group map*. Unlike $Z_j \mapsto \mathbb{E}\theta Z_j$ it is not linear, so this looks like a poor trade, but in fact it is a good trade because the data (V_j, K_j) is local, unlike creatures such as $\exp(-V_j(\Lambda))$ in Z_j . The component factorisation property and Proposition 8.1 allow us to work with K_j on the domain of all connected sets in \mathcal{P}_j . We can prove that $K_j(X)$ is very small when the number of blocks in X is large; in fact, only the restriction of K_j to the small sets \mathcal{S}_j plays an important role.

9. The Inductive Step: Construction of V_{j+1}

In accordance with the program set out in Section 8.2 we describe how V_{j+1} is constructed, given (V_j, K_j) . Our definition of V_{j+1} will be shown to have an additional property that there is an associated K_{j+1} , which, as a function of K_j , is contractive in norms described in Section 10.

Recall that the set \mathcal{S} of small sets was given in Definition 1. For $B \in \mathcal{B}_j$ not containing a, b define V_{j+1} to be the local interaction determined by:

$$\hat{V}_j(B) = V_j(B) + \text{Loc}_B \sum_{Y \in \mathcal{S}, Y \supset B} \frac{1}{|Y|} I_j(Y)^{-1} K_j(Y), \tag{86}$$

$$V_{j+1} = V_{\text{pt}}(\hat{V}_j),$$

where $V_{\text{pt}} = V_{\text{pt}}(V)$ with generic argument V is defined in (70). Recalling the discussion of “relevant terms” just after (58), in (86) V_{j+1} has been defined so that relevant and marginal terms inside K_j are absorbed into V_{j+1} in such a manner that they will not contribute to K_{j+1} . If B contains a or b the combinatorial factor $\frac{1}{|Y|}$ is modified for terms in $\text{Loc}_B K_j$ which are divisible by σ or $\bar{\sigma}$.

We have completed the V part of the inductive construction of the sequence (V_j, K_j) . Before discussing the K induction we have to define some norms so that we can state the contractive property.

10. Norms for K

Let $\mathfrak{h}_j > 0$ and $\mathfrak{s}_j > 0$. For a test function f as defined in Section 5.2 we introduce a norm

$$\|f\|_{\Phi_j} = \sup_{x,y \in \Lambda^*, z \in \{1,2\}^*} \sup_{|\alpha|_\infty \leq 3} \mathfrak{h}_j^{-p-q} \mathfrak{s}_j^{-r} L^{j|\alpha|_1} |\nabla^\alpha f_{x,y,z}|. \tag{87}$$

Multiple derivatives up to order 3 on each argument are specified by the multi-index α . The gradient ∇ represents the finite-difference gradient, and the supremum is taken componentwise over both the forward and backward gradients. A test function f is required to have the property that $f_{x,y,z} = 0$ whenever the sequence x has length $p > 9$ or the sequence z has length $r > 2$; there is no restriction on the length of y . By the definition of the norm, test functions satisfy

$$|\nabla^\alpha f_{x,y,z}| \leq \mathfrak{h}_j^{p+q} \mathfrak{s}_j^r L^{-j|\alpha|_1} \|f\|_{\Phi_j}. \tag{88}$$

We discuss the choice of \mathfrak{s}_j in Section 12 when it first plays a role, and here we focus on \mathfrak{h}_j . An important choice is

$$\mathfrak{h}_j = \ell_j = \ell_0 L^{-j[\phi]}, \tag{89}$$

for a given ℓ_0 . The $L^{-j[\phi]}$ is there because unit norm test functions of one variable should obey the same estimates as a typical field, and test functions of more than one variable should obey the estimates that a product of typical fields obeys.

Recall the pairing defined in (41) and, for $F \in \mathcal{N}$ and $\phi \in \mathbb{C}^\Lambda$, let

$$\|F\|_{T_{\phi,j}} = \sup_{g: \|g\|_{\Phi_j} \leq 1} |\langle F, g \rangle_\phi|. \tag{90}$$

The following proposition provides properties of this seminorm that are well adapted to the control of K .

Proposition 10.1. *Let $F, F_1, F_2 \in \mathcal{N}$. The T_ϕ norm obeys the product property*

$$\|F_1 F_2\|_{T_{\phi,j}} \leq \|F_1\|_{T_{\phi,j}} \|F_2\|_{T_{\phi,j}}, \tag{91}$$

and, if ℓ_0 is chosen large enough, the integration property

$$\|\mathbb{E}_{C_{j+1}} F\|_{T_{\phi,j}(\mathfrak{h}_j)} \leq \mathbb{E}_{C_{j+1}} \|F\|_{T_{\phi+\xi,j}(2\mathfrak{h}_j)}. \tag{92}$$

For further details, see [12]. The second conclusion shows that the norm controls the forms when a fluctuation field is integrated out: on the right hand side the norm is a zero degree form, and hence the expectation is a standard Gaussian expectation.

The most important case of the T_ϕ seminorm is the case $\phi = 0$, but knowing that $\|K(X)\|_{T_0} < \infty$ cannot tell us whether $K(X)$ is integrable. For this we

must limit the growth of $K(X)$ as $\phi \rightarrow \infty$, and the resolution of this issue will be obtained using Definition 3 below.

Our intuitive picture of $K_j(X)$, where $X \in \mathcal{P}_j$, is that it is dominated by a local version of the remainder $(g, \nu, z, \lambda, q)^3$ in (69). To estimate such remainders we must, in particular, estimate $I_{j,X}$ which contains $\exp(-g_j \sum_{x \in X} |\varphi_x|^4)$. By (57) the typical field φ at scale j is roughly constant on scale L^j , and X contains $O(L^{jd})$ points. Therefore this factor looks like $\exp(-g_j L^{dj} |\varphi|^4)$. This is a function of φ/h_j with $h_j \approx g_j^{-1/4} L^{-jd/4}$, which in four dimensions can be rewritten as $g_j^{-1/4} L^{-j[\phi]}$ because $[\phi] = 1$. We want to prove that g_j decays in the same way as does \bar{g}_j in (76), and with this in mind we replace g_j by the known sequence \bar{g}_j . This leads us to our second choice

$$h_j = h_j = k_0 \bar{g}_j^{-1/4} L^{-j[\phi]},$$

where the constant k_0 is determined so that $\exp(-V_j(B))$ will, uniformly in j , have a $T_\phi(h_j)$ norm close to one.

In the previous discussion we made the assumption that the typical φ at scale j is roughly constant on scales L^j . Our norm recognises this; it is a weighted L_∞ norm, where the weight permits growth as fields become atypical. The weight is called a large field regulator and is defined next.

Consider a test function f that is an ersatz field φ , namely a complex-valued function $f = f_x$ for $x \in \Lambda$. For $X \subset \Lambda$, we write $f \in \Pi(X)$ if f restricted to X is a polynomial of degree three or less. We define a seminorm on $\phi = (\varphi, \bar{\varphi})$ by

$$\|\phi\|_{\Phi_j(X)} = \inf\{\|\varphi - f\|_{\Phi_j(\ell_j)} : f \in \Pi(X)\}; \tag{93}$$

note that we are setting $h_j = \ell_j$ in the above equation.

Definition 3. Let $j \in \mathbb{N}_0$, $X \in \mathcal{P}_j$, and $\phi \in \mathbb{C}^\Lambda$. The *large-field regulator* is given by

$$\tilde{G}_j(X, \phi) = \prod_{B \in \mathcal{B}_j(X)} \exp \|\phi\|_{\Phi_j(B^*)}^2, \tag{94}$$

where B^* is the small set neighbourhood of B defined in (78). For each $X \in \mathcal{P}_j$, we define a seminorm on $\mathcal{N}(X^*)$ as follows. For $K(X) \in \mathcal{N}(X^*)$, we define $\|K(X)\|_{\tilde{G}_j, h_j}$ to be the best constant C in

$$\|K(X)\|_{T_{\phi, j}(h_j)} \leq C \tilde{G}_j(X, \phi), \tag{95}$$

where we have made explicit in the notation the fact that the norm on the left hand side is based on the choice $h_j = h_j$.

11. The Inductive Step Completed: Existence of K_{j+1}

We have already specified V_{j+1} in (86). Now we complete the inductive step by constructing K_{j+1} such that (84) holds. The following theorem is at the heart of our method [12]. It provides K_{j+1} and says that we can continue to prolong the sequence (V_j, K_j) for as long as the coupling constants (g_j, ν_j, z_j) remain small. Moreover, in this prolongation, the T_0 norm of K_{j+1} remains third order in the coupling constants and is therefore much smaller than the perturbative (K -independent) part of V_{j+1} .

For $a \geq 0$, set $f_j(a, \emptyset) = 0$, and define

$$f_j(a, X) = 3 + a(|X|_j - 2^d)_+, \quad X \in \mathcal{P}_j \text{ with } X \neq \emptyset. \tag{96}$$

Note that $f_j(a, X) = 3$ when $X \in \mathcal{S}_j$, but that $f_j(a, X)$ is larger than 3 and increases with the size of $|X|_j$ if $X \notin \mathcal{S}_j$. We fix a to have a sufficiently small positive value.

The following theorem is proved for two different choices of the norm pairs $\|\cdot\|_j$ and $\|\cdot\|_{j+1}$, in (97) and (98), and for two corresponding choices of the small parameter $\epsilon_{\delta I}$, as follows:

- $\|\cdot\|_j = \|\cdot\|_{\tilde{\mathcal{G}}_j, h_j}$ with $h_j = k_0 \bar{g}_j^{-1/4} L^{-j[\phi]}$, and $\|\cdot\|_{j+1} = \|\cdot\|_{\tilde{\mathcal{G}}_{j+1}, h_{j+1}}$ with $h_{j+1} = k_0 \bar{g}_{j+1}^{-1/4} L^{-(j+1)[\phi]}$. The small parameter $\epsilon_{\delta I}$ is proportional to $g_j^{1/4}$.
- $\|\cdot\|_j = \|\cdot\|_{T_0, \ell_j}$ with $\ell_j = \ell_0 L^{-j[\phi]}$, and $\|\cdot\|_{j+1} = \|\cdot\|_{T_0, \ell_{j+1}}$. The small parameter $\epsilon_{\delta I}$ is proportional to g_j .

Define a cone $C = \{(g_j, \nu_j, z_j) | g > 0, |\nu| \vee |z| \leq bg, g_j \leq c(b, L)\}$. The constant b is determined in Section 12, and $c(b, L)$ is a function of b, L constructed in the proof of the next theorem.

Theorem 11.1. *Let $(g_j, \nu_j, z_j) \in C$. Let a be sufficiently small, and let M be any (large) positive constant that is independent of d, L . There is a constant c_{pt} (depending on d, L) such that the following holds. Suppose that $K_j : \mathcal{P}_j \rightarrow \mathcal{N}_j$ has properties (84) and satisfies*

$$\|K_j(X)\|_j \leq M c_{\text{pt}} \epsilon_{\delta I}^{f_j(a, X)}, \quad X \in \mathcal{P}_j \text{ connected}, \tag{97}$$

Then, if L is sufficiently large (depending on M), there exists $K_{j+1} : \mathcal{P}_{j+1} \rightarrow \mathcal{N}_{j+1}$ with properties (84) at scale $j + 1$ and

$$\|K_{j+1}(U)\|_{j+1} \leq 2 c_{\text{pt}} \epsilon_{\delta I}^{f_{j+1}(a, U)}, \quad U \in \mathcal{P}_{j+1} \text{ connected}. \tag{98}$$

12. Decay of the Two-point Function

Finally, we combine the machinery we have developed, to outline the proof of Theorem 4.1. As we have already noted, Theorem 1.2 is a consequence of Theorem 4.1.

We must study the coupling constant flow. The linear map $\text{Loc}_B : \mathcal{N} \rightarrow \mathcal{V}$ is bounded in T_0 norm [12], so according to the inductive assumption (97) on the T_0 norm of K_j , the coupling constants in \hat{V}_j of (86) are small (third order) adjustments to the coupling constants in V_j . Theorem 11.1 ensures that this smallness is preserved as the scale advances.

We first consider the case $(\lambda_0, q_0) = (0, 0)$. In this case, $(\lambda_j, q_j) = (0, 0)$ for all j . The definition of V_{j+1} in (86) then gives rise to a non-perturbative version of the flow equations of Proposition 7.1, in which the effect of K is now taken into account. When $V_j \mapsto V_{j+1}$ is expressed as

$$(g_j, \nu_j, z_j) \mapsto (g_{j+1}, \nu_{j+1}, z_{j+1}) \tag{99}$$

we find that

$$g_{j+1} = g_j - c_g g_j^2 + r_{g,j}, \tag{100}$$

$$\nu_{j+1} = \nu_j + 2g C_{j+1}(0, 0) + r_{\nu,j}, \tag{101}$$

$$z_{j+1} = z_j + r_{z,j}, \tag{102}$$

$$K_{j+1} = r_{K,j}(g_j, \nu_j, z_j, K_j), \tag{103}$$

where the r 's now depend also on K_j , and where we have added the map $r_{K,j} : (g_j, \nu_j, z_j, K_j) \mapsto K_{j+1}$ defined by Theorem 11.1. Furthermore, we prove that the r 's are Lipschitz functions of (g_j, ν_j, z_j, K_j) , where K belongs to a Banach space normed by a combination of the norms in Section 11. These are the properties needed to prove that K only causes a small deformation of the perturbative flow $V \mapsto V_{\text{pt}}$.

The main theorem now reduces to an exercise in dynamical systems. We prove that with a suitable choice of the constant b defining the cone C there is a *Lipschitz stable manifold* of initial conditions $(z_0, \nu_0) = h(m^2, g_0)$ for which the sequence (V_j, K_j) , $j = 0, \dots, N$, has a limit as $N \rightarrow \infty$ and $m^2 \downarrow 0$. We call this the *global trajectory*. For $m^2 = 0$, the global trajectory tends to the fixed point $(V, K) = (0, 0)$. In particular, $g_j \rightarrow 0$, which is infra-red asymptotic freedom. Referring to (29), we have four unknown parameters g_0, ν_0, z_0, m^2 related by three equations, and now there is a fourth equation $(z_0, \nu_0) = h(m^2, g_0)$. By the implicit function theorem we solve for the unknowns as functions of (g, ν) . As $\nu \downarrow \nu_c(g)$, $m^2 \downarrow 0$ and vice-versa.

Now we consider the flow for (λ_j, q_j) . According to (60), $\lambda_0 = 1$ and $q_0 = 0$. Using (50), we prove that the terms $r_{g,j}, r_{\nu,j}, r_{z,j}$ do not depend on λ_j, q_j and thus the coupling constants g, ν, z have no dependence on λ, q . From (86) we

find

$$\lambda_{j+1} = \left(1 + \sum_{y \in \Lambda} (\nu_{j+1} w_{j+1}(0, y) - \nu_j w_j(0, y)) \right) \lambda_j + r_{\lambda, j}, \tag{104}$$

$$q_{j+1} = q_j + \lambda^2 C_{j+1}(a, b) + r_{q, j}, \tag{105}$$

where $r_{\lambda, j}, r_{q, j}$ are corrections that include contributions from K_j .

Recall that \mathcal{S}_j was defined in Definition 1. Let $s_{a, b}$ be the first scale j such that there exists a polymer in \mathcal{S}_j that contains $\{a, b\}$. The correction $r_{q, j}$ is zero for all scales $j < s_{a, b}$: according to (50) and the definition of \hat{V} in (86) there can be no $\sigma\bar{\sigma}$ contribution from K_j until the first scale where there is a set $X \in \mathcal{S}_j$ that covers $\{a, b\}$. Also, by the finite range property, $C_{j+1}(a, b) = 0$ for $j < s_{a, b}$. Thus (105) gives

$$q_N = \sum_{j=s_{a, b}}^N (\lambda_j^2 C_{j+1}(a, b) + r_{q, j}). \tag{106}$$

At scale N , Λ is a single block in \mathcal{B}_N , so by the definition of the circle product, Z_N is simply given by

$$Z_N = (I_N \circ K_N)(\Lambda) = I_N(\Lambda) + K_N(\Lambda). \tag{107}$$

The final renormalisation group map is the action of \mathbb{E}_{C_N} , not $\mathbb{E}_{C_N}\theta$. This means that the fields ϕ, ψ are to be set to zero in I_N, K_N , and only dependence on σ remains. By (64) we compute two σ derivatives of I_N and find

$$-\frac{\partial^2}{\partial\sigma\partial\bar{\sigma}} \Big|_0 Z_N = q_N - K_{\bar{\sigma}\sigma}, \quad \text{where } K_{\bar{\sigma}\sigma} = \frac{\partial^2 K_N(\Lambda)}{\partial\sigma\partial\bar{\sigma}} \Big|_0. \tag{108}$$

The $\bar{\sigma}\sigma$ derivative is a coefficient in the pairing (41), and the T_0 norm bounds this pairing, so Theorem 11.1 gives

$$|K_{\bar{\sigma}\sigma}| \leq \|K\|_{T_0, N} \mathfrak{s}_N^{-2} \leq O(g_N^3) \mathfrak{s}_N^{-2}. \tag{109}$$

We are able to prove Theorem 11.1 with

$$\mathfrak{s}_j = \mathfrak{s}_0 \ell_j^{-1} \approx O(L^{j \wedge s_{a, b}}), \tag{110}$$

where \mathfrak{s}_0 is a constant, so that, when $N > s_{a, b}$,

$$|K_{\bar{\sigma}\sigma}| \leq O(g_N^3) L^{-2(N \wedge s_{a, b})} = O(g_N^3 |a - b|^{-2}). \tag{111}$$

This tends to zero as $N \rightarrow \infty$.

By a similar estimate we can control the $r_{\lambda, j}, r_{q, j}$ terms in (104), (106). These contain σ derivatives of the K_j terms in (86). The conclusion is that

$\lambda_\infty = \lim_{N \rightarrow \infty} \lambda_N$ and $q_\infty = \lim_{N \rightarrow \infty} q_N$ exist and are bounded away from zero.

By (32), the left hand side of (30) is given by

$$\lim_{\nu \downarrow \nu_c} (1 + z_0) \lim_{\Lambda \uparrow \mathbb{Z}^d} \int_{\mathbb{C}^\Lambda} e^{-S(\Lambda) - \tilde{V}_0(\Lambda)} \bar{\varphi}_a \varphi_b = \lim_{m^2 \downarrow 0} (1 + z_0) q_\infty. \tag{112}$$

From (104) and (106) we find that

$$\lim_{m^2 \downarrow 0} q_\infty \sim \lambda_\infty^2 \sum_{j=s_{a,b}}^\infty C_{j+1}(a, b), \tag{113}$$

where $m^2 = 0$ in C_{j+1} , and \sim means that the ratio of the left hand side and the right hand side tends to one as $a - b \rightarrow \infty$. Next, we use the finite range property to restore the scales $j < s_{a,b}$ to the sum, which then becomes the complete finite range decomposition for the infinite volume simple random walk two-point function $(-\Delta)^{-1}(a, b)$,

$$\lim_{m^2 \downarrow 0} q_\infty \sim \lambda_\infty^2 (-\Delta)^{-1}(a, b). \tag{114}$$

The right hand side of (114), and hence of (112), is thus asymptotic to a multiple of $|a - b|^{-2}$ as $|a - b| \rightarrow \infty$, as desired, since the inverse Laplacian has this behaviour.

Acknowledgements

We thank Roland Bauerschmidt for contributions to the proof of Proposition 2.1. The work of both authors was supported in part by NSERC of Canada. DB gratefully acknowledges the support and hospitality of the Institute for Advanced Study and Eurandom, where part of this work was done, and dedicates this work to his wife Betty Lu. GS gratefully acknowledges the support and hospitality of the Institut Henri Poincaré, and of the Kyoto University Global COE Program in Mathematics, during stays in Paris and Kyoto where part of this work was done.

References

- [1] G. Benfatto, N. Cassandro, G. Gallavotti, F. Nicolo, E. Olivieri, E. Presutti, and E. Scacciatelli. On the ultraviolet-stability in the Euclidean scalar field theories. *Commun. Math. Phys.*, **71**:95–130, (1980).
- [2] E. Brézin, J.C. Le Guillou, and J. Zinn-Justin. Approach to scaling in renormalized perturbation theory. *Phys. Rev. D*, **8**:2418–2430, (1973).
- [3] D. Brydges, J. Dimock, and T. R. Hurd. A non-Gaussian fixed point for ϕ^4 in $4 - \epsilon$ dimensions. *Commun. Math. Phys.*, **198**(1):111–156, (1998).

-
- [4] D. Brydges, S.N. Evans, and J.Z. Imbrie. Self-avoiding walk on a hierarchical lattice in four dimensions. *Ann. Probab.*, **20**:82–124, (1992).
- [5] D. Brydges, A. Járai Jr., and A. Sakai. Self-interacting walk and functional integration.
<http://www.math.ubc.ca/db5d/Seminars/PIMSLectures2001/lectures.pdf>, (2001).
- [6] D.C. Brydges. Lectures on the renormalisation group. In S. Sheffield and T. Spencer, editors, *Statistical Mechanics*, pages 7–93. American Mathematical Society, Providence, (2009). IAS/Park City Mathematics Series, Volume 16.
- [7] D.C. Brydges, G. Guadagni, and P.K. Mitter. Finite range decomposition of Gaussian processes. *J. Stat. Phys.*, **115**:415–449, (2004).
- [8] D.C. Brydges and J.Z. Imbrie. End-to-end distance from the Green’s function for a hierarchical self-avoiding walk in four dimensions. *Commun. Math. Phys.*, **239**:523–547, (2003).
- [9] D.C. Brydges and J.Z. Imbrie. Green’s function for a hierarchical self-avoiding walk in four dimensions. *Commun. Math. Phys.*, **239**:549–584, (2003).
- [10] D.C. Brydges, J.Z. Imbrie, and G. Slade. Functional integral representations for self-avoiding walk. *Probab. Surveys*, **6**:34–61, (2009).
- [11] D.C. Brydges, P.K. Mitter, and B. Scoppola. Critical $(\Phi^4)_{3,\epsilon}$. *Commun. Math. Phys.*, **240**:281–327, (2003).
- [12] D.C. Brydges and G. Slade. Weakly self-avoiding walk in dimensions four and higher: a renormalisation group analysis. In preparation.
- [13] D.C. Brydges and H.-T. Yau. Grad ϕ perturbations of massless Gaussian fields. *Commun. Math. Phys.*, **129**:351–392, (1990).
- [14] J. Dimock and T.R. Hurd. A renormalization group analysis of correlation functions for the dipole gas. *J. Stat. Phys.*, **66**:1277–1318, (1992).
- [15] K. Gawedzki and A. Kupiainen. Block spin renormalization group for dipole gas and $(\nabla\phi)^4$. *Ann. Phys.*, **147**:198, (1983).
- [16] S.E. Golowich and J.Z. Imbrie. The broken supersymmetry phase of a self-avoiding random walk. *Commun. Math. Phys.*, **168**:265–319, (1995).
- [17] C. Gruber and H. Kunz. General properties of polymer systems. *Commun. Math. Phys.*, **22**:133–161, (1971).
- [18] T. Hara. Decay of correlations in nearest-neighbor self-avoiding walk, percolation, lattice trees and animals. *Ann. Probab.*, **36**:530–593, (2008).
- [19] T. Hara, R. van der Hofstad, and G. Slade. Critical two-point functions and the lace expansion for spread-out high-dimensional percolation and related models. *Ann. Probab.*, **31**:349–408, (2003).
- [20] T. Hara and M. Ohno. Renormalization group analysis of hierarchical weakly self-avoiding walk in four dimensions. In preparation.
- [21] T. Hara and G. Slade. The lace expansion for self-avoiding walk in five or more dimensions. *Reviews in Math. Phys.*, **4**:235–327, (1992).
- [22] T. Hara and G. Slade. Self-avoiding walk in five or more dimensions. I. The critical behaviour. *Commun. Math. Phys.*, **147**:101–136, (1992).

-
- [23] I. Horvath, B. Toth, and B. Veto. Diffusive limit for self-repelling Brownian polymers in three and more dimensions. <http://arxiv.org/abs/0912.5174>.
- [24] D. Iagolnitzer and J. Magnen. Polymers in a weak random potential in dimension four: rigorous renormalization group analysis. *Commun. Math. Phys.*, **162**:85–121, (1994).
- [25] G.F. Lawler, O. Schramm, and W. Werner. On the scaling limit of planar self-avoiding walk. *Proc. Symposia Pure Math.*, **72**:339–364, (2004).
- [26] N. Madras and G. Slade. *The Self-Avoiding Walk*. Birkhäuser, Boston, (1993).
- [27] P.K. Mitter and B. Scoppola. The global renormalization group trajectory in a critical supersymmetric field theory on the lattice \mathbf{Z}^3 . *J. Stat. Phys.*, **133**:921–1011, (2008).
- [28] K. G. Wilson. Renormalization group and critical phenomena I and II. *Phys. Rev.*, **B4**:3174–3183, 3184–3205, (1971).
- [29] K. G. Wilson. The renormalization group and critical phenomena. *Rev. Modern Phys.*, **55**(3):583–600, (1983).
- [30] K. G. Wilson and J. Kogut. The renormalization group and the ϵ expansion. *Phys. Rep. (Sect C of Phys Lett.)*, **12**:75–200, (1974).

A Key Large Deviation Principle for Interacting Stochastic Systems

Frank den Hollander*

Abstract

In this paper we describe two large deviation principles for the empirical process of words cut out from a random sequence of letters according to a random renewal process: one where the letters are frozen (“quenched”) and one where the letters are not frozen (“annealed”). We apply these large deviation principles to five classes of interacting stochastic systems: interacting diffusions, coupled branching processes, and three examples of a polymer chain in a random environment. In particular, we show how these large deviation principles can be used to derive variational formulas for the critical curves that are associated with the phase transitions occurring in these systems, and how these variational formulas can in turn be used to prove the existence of certain intermediate phases.

Mathematics Subject Classification (2010). Primary 60F10, 60G50, 60K35; Secondary 82C22, 82D60.

Keywords. Large deviation principle, quenched vs. annealed, interacting stochastic systems, variational formulas, phase transitions, intermediate phases.

1. Large Deviation Principles

In Section 1 we describe two large deviation principles that were derived in Birkner, Greven and den Hollander [3]. In Sections 2–4 we apply these large deviation principles to five classes of interacting stochastic systems that exhibit a phase transition. In Section 5 we argue why these applications open up a new window of research, with a variational view, and we make a few closing remarks.

*The research described in this paper is joint work with M. Birkner (Mainz), E. Bolthausen (Zürich), D. Cheliotis (Athens) and A. Greven (Erlangen).

Mathematical Institute, Leiden University, Leiden, The Netherlands.
E-mail: denholla@math.leidenuniv.nl.

1.1. Letters, words and sentences. Let E be a Polish space (e.g. $E = \mathbb{Z}^d$, $d \geq 1$, with the lattice norm or $E = \mathbb{R}$ with the Euclidean norm). Think of E as an alphabet, i.e., a set of *letters*. Let $\tilde{E} = \cup_{n \in \mathbb{N}} E^n$ be the set of finite *words* drawn from E , which can be metrised to become a Polish space.

For ν a probability measure on E , let $X = (X_k)_{k \in \mathbb{N}_0}$ (with $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$) be i.i.d. with law ν . For ρ a probability measure on \mathbb{N} , let $\tau = (\tau_i)_{i \in \mathbb{N}}$ be i.i.d. with law ρ . Assume that X and τ are independent and write Pr to denote their joint law.

Given X and τ , define $Y = (Y^{(i)})_{i \in \mathbb{N}}$ by putting

$$T_0 = 0 \quad \text{and} \quad T_i = T_{i-1} + \tau_i, \quad i \in \mathbb{N}, \tag{1.1}$$

and

$$Y^{(i)} = (X_{T_{i-1}}, X_{T_{i-1}+2}, \dots, X_{T_i-1}), \quad i \in \mathbb{N}. \tag{1.2}$$

In words, Y is the infinite sequence of words cut out from the infinite sequence of letters X according to the renewal times τ (see Fig. 1). Clearly, under the law Pr , Y is i.i.d. with law $q_{\rho, \nu}^{\otimes \mathbb{N}}$ on $\tilde{E}^{\mathbb{N}}$, the set of infinite *sentences*, where the marginal law $q_{\rho, \nu}$ on \tilde{E} is given by

$$q_{\rho, \nu}((x_1, \dots, x_n)) = \rho(n) \nu(x_1) \cdots \nu(x_n), \quad n \in \mathbb{N}, x_1, \dots, x_n \in E. \tag{1.3}$$

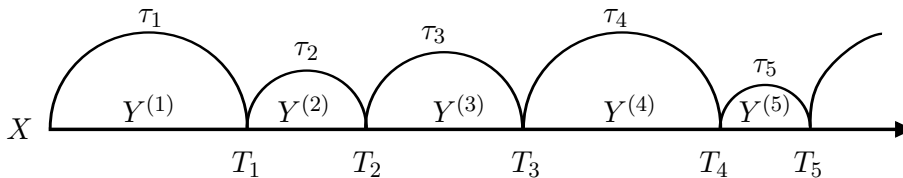


Figure 1. Cutting words out from a sequence of letters according to renewal times.

The reverse operation of *cutting* words out from a sequence of letters is *glueing* words together into a sequence of letters. Formally, this is done by defining a *concatenation* map κ from $\tilde{E}^{\mathbb{N}}$ to $E^{\mathbb{N}}$. This map induces in a natural way a map κ from $\mathcal{P}(\tilde{E}^{\mathbb{N}})$ to $\mathcal{P}(E^{\mathbb{N}})$, the sets of probability measures on $\tilde{E}^{\mathbb{N}}$ and $E^{\mathbb{N}}$ (endowed with the topology of weak convergence). The concatenation $q_{\rho, \nu}^{\otimes \mathbb{N}} \circ \kappa^{-1}$ of $q_{\rho, \nu}^{\otimes \mathbb{N}}$ equals $\nu^{\otimes \mathbb{N}}$, as is evident from (1.3).

Note that in the above set-up three objects can be freely chosen: E (alphabet), ν (letter law) and ρ (word length law). In what follows we will assume that ρ has infinite support and satisfies

$$\lim_{\substack{n \rightarrow \infty \\ \rho(n) > 0}} \frac{\log \rho(n)}{\log n} = -\alpha \quad \text{for some } \alpha \in [1, \infty). \tag{1.4}$$

1.2. Annealed LDP. Let $\mathcal{P}^{\text{inv}}(\tilde{E}^{\mathbb{N}})$ be the set of probability measures on $\tilde{E}^{\mathbb{N}}$ that are invariant under the left-shift $\tilde{\theta}$ acting on $\tilde{E}^{\mathbb{N}}$. For $N \in \mathbb{N}$, let

$(Y^{(1)}, \dots, Y^{(N)})^{\text{per}}$ be the periodic extension of the N -tuple $(Y^{(1)}, \dots, Y^{(N)}) \in \tilde{E}^N$ to an element of $\tilde{E}^{\mathbb{N}}$, and define

$$R_N = \frac{1}{N} \sum_{i=0}^{N-1} \delta_{\tilde{\theta}^i(Y^{(1)}, \dots, Y^{(N)})^{\text{per}}} \in \mathcal{P}^{\text{inv}}(\tilde{E}^{\mathbb{N}}). \tag{1.5}$$

This is the *empirical process of N -tuples of words* in Y . The following large deviation principle (LDP) is standard (see e.g. Dembo and Zeitouni [14], Corollaries 6.5.15 and 6.5.17). Let

$$H(Q \mid q_{\rho, \nu}^{\otimes \mathbb{N}}) = \lim_{N \rightarrow \infty} \frac{1}{N} h \left(Q_{|\mathcal{F}_N} \mid (q_{\rho, \nu}^{\otimes \mathbb{N}})_{|\mathcal{F}_N} \right) \in [0, \infty] \tag{1.6}$$

be the *specific relative entropy of Q w.r.t. $q_{\rho, \nu}^{\otimes \mathbb{N}}$* . Here, $\mathcal{F}_N = \sigma(Y^{(1)}, \dots, Y^{(N)})$ is the sigma-algebra generated by the first N words, $Q_{|\mathcal{F}_N}$ is the restriction of Q to \mathcal{F}_N , and $h(\cdot \mid \cdot)$ denotes relative entropy.

Theorem 1.1. [Annealed LDP] *The family of probability distributions $\Pr(R_N \in \cdot)$, $N \in \mathbb{N}$, satisfies the LDP on $\mathcal{P}^{\text{inv}}(\tilde{E}^{\mathbb{N}})$ with rate N and with rate function $I^{\text{ann}}: \mathcal{P}^{\text{inv}}(\tilde{E}^{\mathbb{N}}) \rightarrow [0, \infty]$ given by*

$$I^{\text{ann}}(Q) = H(Q \mid q_{\rho, \nu}^{\otimes \mathbb{N}}). \tag{1.7}$$

The rate function I^{ann} is lower semi-continuous, has compact level sets, has a unique zero at $Q = q_{\rho, \nu}^{\otimes \mathbb{N}}$, and is affine.

Informally, Theorem 1.1 says that $\Pr(R_N \approx Q) \approx e^{-NI^{\text{ann}}(Q)}$ as $N \rightarrow \infty$.

1.3. Quenched LDP. To formulate the quenched analogue of Theorem 1.1, which is the main result in Birkner, Greven and den Hollander [3], we need some further notation. Let $\mathcal{P}^{\text{inv}}(E^{\mathbb{N}})$ be the set of probability measures on $E^{\mathbb{N}}$ that are invariant under the left-shift θ acting on $E^{\mathbb{N}}$. For $Q \in \mathcal{P}^{\text{inv}}(\tilde{E}^{\mathbb{N}})$ such that $m_Q = E_Q[\tau_1] < \infty$ (where E_Q denotes expectation under the law Q and τ_1 is the length of the first word), define

$$\Psi_Q(\cdot) = \frac{1}{m_Q} E_Q \left[\sum_{k=0}^{\tau_1-1} \delta_{\theta^k \kappa(Y)}(\cdot) \right] \in \mathcal{P}^{\text{inv}}(E^{\mathbb{N}}). \tag{1.8}$$

Think of Ψ_Q as the shift-invariant version of $Q \circ \kappa^{-1}$ obtained after *randomising* the location of the origin. This randomisation is necessary because a shift-invariant Q in general does not (!) give rise to a shift-invariant $Q \circ \kappa^{-1}$.

For $\text{tr} \in \mathbb{N}$, let $[\cdot]_{\text{tr}}: \tilde{E} \rightarrow [\tilde{E}]_{\text{tr}} = \cup_{n=1}^{\text{tr}} E^n$ denote the *word length truncation* map defined by

$$y = (x_1, \dots, x_n) \mapsto [y]_{\text{tr}} = (x_1, \dots, x_{n \wedge \text{tr}}), \quad n \in \mathbb{N}, x_1, \dots, x_n \in E, \tag{1.9}$$

i.e., $[y]_{\text{tr}}$ is the word of length $\leq \text{tr}$ obtained from the word y by dropping all the letters with label $> \text{tr}$. This map induces in a natural way a map from $\tilde{E}^{\mathbb{N}}$ to $[\tilde{E}]_{\text{tr}}^{\mathbb{N}}$, and from $\mathcal{P}^{\text{inv}}(\tilde{E}^{\mathbb{N}})$ to $\mathcal{P}^{\text{inv}}([\tilde{E}]_{\text{tr}}^{\mathbb{N}})$. Note that if $Q \in \mathcal{P}^{\text{inv}}(\tilde{E}^{\mathbb{N}})$, then $[Q]_{\text{tr}}$ is an element of the set

$$\mathcal{P}^{\text{inv,fin}}(\tilde{E}^{\mathbb{N}}) = \{Q \in \mathcal{P}^{\text{inv}}(\tilde{E}^{\mathbb{N}}) : m_Q < \infty\}. \tag{1.10}$$

Theorem 1.2. [Quenched LDP] *For $\nu^{\otimes \mathbb{N}}$ -a.s. all X , the family of regular conditional probability distributions $\Pr(R_N \in \cdot \mid X)$, $N \in \mathbb{N}$, satisfies the LDP on $\mathcal{P}^{\text{inv}}(\tilde{E}^{\mathbb{N}})$ with rate N and with deterministic rate function $I^{\text{que}} : \mathcal{P}^{\text{inv}}(\tilde{E}^{\mathbb{N}}) \rightarrow [0, \infty]$ given by*

$$I^{\text{que}}(Q) = \begin{cases} I^{\text{fin}}(Q), & \text{if } Q \in \mathcal{P}^{\text{inv,fin}}(\tilde{E}^{\mathbb{N}}), \\ \lim_{\text{tr} \rightarrow \infty} I^{\text{fin}}([Q]_{\text{tr}}), & \text{otherwise,} \end{cases} \tag{1.11}$$

where

$$I^{\text{fin}}(Q) = H(Q \mid q_{\rho,\nu}^{\otimes \mathbb{N}}) + (\alpha - 1) m_Q H(\Psi_Q \mid \nu^{\otimes \mathbb{N}}). \tag{1.12}$$

The rate function I^{que} is lower semi-continuous, has compact level sets, has a unique zero at $Q = q_{\rho,\nu}^{\otimes \mathbb{N}}$, and is affine.

Informally, Theorem 1.2 says that $\Pr(R_N \approx Q \mid X) \approx e^{-NI^{\text{que}}(Q)}$ as $N \rightarrow \infty$ for $\nu^{\otimes \mathbb{N}}$ -a.s. all X .

Note from (1.7) and (1.11–1.12) that I^{que} equals I^{ann} plus an additional term that quantifies the deviation of Ψ_Q , the randomised concatenation of Q , from the reference law $\nu^{\otimes \mathbb{N}}$ of the letter sequence. This term, which also depends on the exponent α in (1.4), is explicit when $m_Q < \infty$, but requires a truncation approximation when $m_Q = \infty$. Further note that if $\alpha = 1$, then the additional term vanishes and $I^{\text{que}} = I^{\text{ann}}$.

2. Collision Local Time of Two Random Walks

In this section we apply Theorems 1.1–1.2 to study the collision local time of two random walks. The results are taken from Birkner, Greven and den Hollander [4]. In Section 3 we will use the outcome of this section to describe phase transitions in two interacting stochastic systems: interacting diffusions and coupled branching processes.

Let $S = (S_k)_{k \in \mathbb{N}_0}$ and $S' = (S'_k)_{k \in \mathbb{N}_0}$ be two independent random walks on \mathbb{Z}^d , $d \geq 1$, both starting at the origin and with an irreducible, symmetric and transient transition kernel $p(\cdot, \cdot)$. Write p^n for the n -th convolution power of p . Suppose that

$$\lim_{n \rightarrow \infty} \frac{\log p^{2n}(0, 0)}{\log n} = -\alpha \quad \text{for some } \alpha \in [1, \infty). \tag{2.1}$$

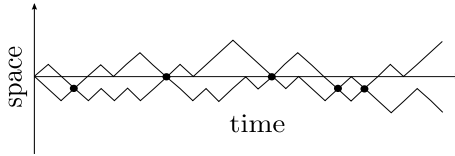


Figure 2. Two random walks that build up collision local time.

Write P to denote the joint law of S, S' . Let

$$V = V(S, S') = \sum_{k \in \mathbb{N}} 1_{\{S_k = S'_k\}} \tag{2.2}$$

be the *collision local time* of S, S' (see Fig. 2), which satisfies $P(V < \infty) = 1$ because $p(\cdot, \cdot)$ is transient. Define

$$z_1 = \sup \{z \geq 1 : E[z^V | S] < \infty \text{ } S\text{-a.s.}\}, \tag{2.3}$$

$$z_2 = \sup \{z \geq 1 : E[z^V] < \infty\}. \tag{2.4}$$

(The lower indices indicate the number of random walks being averaged over.) Note that, by the tail triviality of S , the range of z -values for which $E[z^V | S]$ converges is S -a.s. constant.

As shown in [4], Theorems 1.1–1.2 can be applied with the following choice of E, ν and ρ :

$$E = \mathbb{Z}^d, \quad \nu(x) = p(0, x), \quad \rho(n) = p^{2\lfloor n/2 \rfloor}(0, 0) / [2\bar{G}(0, 0) - 1], \tag{2.5}$$

where $\bar{G}(0, 0) = \sum_{n \in \mathbb{N}_0} p^{2n}(0, 0)$ is the Green function at the origin associated with $p^2(\cdot, \cdot)$, the transition kernel of $S - S'$. The following theorem provides variational formulas for z_1 and z_2 . This theorem requires additional assumptions on $p(\cdot, \cdot)$:

$$\begin{aligned} & \sum_{x \in \mathbb{Z}^d} \|x\|^\delta p(0, x) < \infty \text{ for some } \delta > 0, \\ & \liminf_{n \rightarrow \infty} \frac{\log[p^n(0, S_n) / p^{2\lfloor n/2 \rfloor}(0, 0)]}{\log n} \geq 0 \quad S - a.s., \\ & \inf_{n \in \mathbb{N}} E[\log[p^n(0, S_n) / p^{2\lfloor n/2 \rfloor}(0, 0)]] > -\infty. \end{aligned} \tag{2.6}$$

As shown in [4], the last two assumptions hold for a large class of random walks, including those that are in the domain of attraction of a normal law, respectively, a symmetric stable law. They potentially hold in full generality under a mild regularity condition on $p(\cdot, \cdot)$.¹

¹The symmetry of $p(\cdot, \cdot)$ implies that $p^{2n}(0, 0) > 0$ for all $n \in \mathbb{N}_0$ and $p^n(0, x) / p^{2\lfloor n/2 \rfloor}(0, 0) \leq 1$ for all $n \in \mathbb{N}_0$ and $x \in \mathbb{Z}^d$.

Theorem 2.1. Assume (2.1) and (2.6). Then $z_1 = 1 + e^{-r_1}$, $z_2 = 1 + e^{-r_2}$ with

$$r_1 = \sup_{Q \in \mathcal{P}^{\text{inv}}(\widetilde{\mathbb{Z}}^{\mathbb{N}})} \left\{ \int_{\widetilde{\mathbb{Z}}^d} (\pi_1 Q)(dy) \log f(y) - I^{\text{que}}(Q) \right\} \in \mathbb{R}, \quad (2.7)$$

$$r_2 = \sup_{Q \in \mathcal{P}^{\text{inv}}(\widetilde{\mathbb{Z}}^{\mathbb{N}})} \left\{ \int_{\widetilde{\mathbb{Z}}^d} (\pi_1 Q)(dy) \log f(y) - I^{\text{ann}}(Q) \right\} \in \mathbb{R}, \quad (2.8)$$

where $\pi_1 Q$ is the projection of Q onto $\widetilde{\mathbb{Z}}^d$, i.e., the law of the first word, and $f: \widetilde{\mathbb{Z}}^d \rightarrow [0, \infty)$ is given by

$$f((x_1, \dots, x_n)) = \frac{1}{\rho(n)} p^n(0, x_1 + \dots + x_n), \quad n \in \mathbb{N}, x_1, \dots, x_n \in \mathbb{Z}^d. \quad (2.9)$$

Remark: Since $P(V = k) = (1 - \bar{F})\bar{F}^k$, $k \in \mathbb{N}_0$, with $\bar{F} = P(\exists k \in \mathbb{N}: S_k = S'_k)$, an easy computation gives $z_2 = 1/\bar{F}$. Since $\bar{F} = 1 - [1/\bar{G}(0, 0)]$, we therefore have $z_2 = \bar{G}(0, 0)/[\bar{G}(0, 0) - 1]$. This simple formula reflects itself in the fact that the variational formula in (2.8) can be solved explicitly (see [4]). However, unlike for z_2 , no closed form expression is known for z_1 , because the variational formula in (2.7) cannot be solved explicitly.

Because $I^{\text{que}} \geq I^{\text{ann}}$, we have $r_1 \leq r_2$, and hence $z_2 \leq z_1$. The following corollary gives conditions under which strict inequality holds or not. Its proof in [4] relies on a comparison of the two variational formulas in (2.7–2.8).

Corollary 2.2. Assume (2.1) and (2.6).

- (a) If $p(\cdot, \cdot)$ is strongly transient, i.e., $\sum_{n \in \mathbb{N}} np^n(0, 0) < \infty$, then $z_2 < z_1$.
- (b) If $\alpha = 1$, then $z_1 = z_2$.

Analogous results hold when we turn the discrete-time random walks S and S' into continuous-time random walks $\tilde{S} = (S_t)_{t \geq 0}$ and $\tilde{S}' = (\tilde{S}'_t)_{t \geq 0}$ by allowing them to make steps at rate 1, while keeping the same transition kernel $p(\cdot, \cdot)$. Then the collision local time becomes

$$\tilde{V} = \int_0^\infty 1_{\{\tilde{S}_t = \tilde{S}'_t\}} dt. \quad (2.10)$$

For the analogous quantities \tilde{z}_1 and \tilde{z}_2 , variational formulas like in Theorem 2.1 can be derived, and a result similar to Corollary 2.2 holds:

Corollary 2.3. Assume (2.1) and (2.6).

- (a) If $p(\cdot, \cdot)$ is strongly transient, then $\tilde{z}_2 < \tilde{z}_1$.
- (b) If $\alpha = 1$, then $\tilde{z}_1 = \tilde{z}_2$.

An easy computation gives $\log \tilde{z}_2 = 2/G(0, 0)$, where $G(0, 0) = \sum_{n \in \mathbb{N}_0} p^n(0, 0)$ is the Green function at the origin associated with $p(\cdot, \cdot)$. There is again no closed form expression for \tilde{z}_1 .

Recent progress on extending the gaps in Corollaries 2.2(a) and 2.3(a) to transient random walks that are not strongly transient (like simple random walk in $d = 3, 4$) can be found in Birkner and Sun [5], [6], and in Berger and Toninelli [1]. These papers require assumptions on the tail of $p(0, \cdot)$ and use fractional moment estimates rather than variational formulas.

3. Two Applications Without Disorder

3.1. Interacting diffusions. Consider the following system of coupled stochastic differential equations:

$$dX_x(t) = \sum_{y \in \mathbb{Z}^d} p(x, y)[X_y(t) - X_x(t)] dt + \sqrt{qX_x(t)^2} dW_x(t), \quad x \in \mathbb{Z}^d, t \geq 0. \quad (3.1)$$

Here, $p(\cdot, \cdot)$ is a random walk transition kernel on \mathbb{Z}^d , $q \in (0, \infty)$ is a diffusion constant, and $W = (W(t))_{t \geq 0}$ with $W(t) = \{W_x(t)\}_{x \in \mathbb{Z}^d}$ is a collection of independent standard Brownian motions on \mathbb{R} . The initial condition is chosen such that $\{X_x(0)\}_{x \in \mathbb{Z}^d}$ is a shift-invariant and shift-ergodic random field taking values in $[0, \infty)$ with a positive and finite mean (the evolution in (3.1) preserves the mean).

It was shown in Greven and den Hollander [19] that if $p(\cdot, \cdot)$ is irreducible, symmetric and transient, then there exist $0 < q_2 \leq q_* < \infty$ such that the system in (3.1) locally dies out when $q > q_*$, but converges to a non-trivial equilibrium when $q < q_*$, and this equilibrium has an *infinite second moment* when $q \geq q_2$ and a *finite second moment* when $q < q_2$. It was conjectured in [19] that $q_2 < q_*$. Since it was shown in [19] that

$$q_* \geq \log \tilde{z}_1, \quad q_2 = \log \tilde{z}_2, \quad (3.2)$$

Corollary 2.3(a) settles this conjecture when $p(\cdot, \cdot)$ satisfies (2.1) and (2.6) and is strongly transient.

3.2. Coupled branching processes. Consider a spatial population model on \mathbb{Z}^d evolving as follows:

- (1) Each individual migrates at rate 1 according to $p(\cdot, \cdot)$.
 - (2) Each individual gives birth to a new individual at the same site at rate q .
 - (3) Each individual dies at rate $q(1 - r)$.
 - (4) All individuals at the same site die simultaneously at rate qr .
- (3.3)

Here, $p(\cdot, \cdot)$ is a random walk transition kernel on \mathbb{Z}^d , $q \in (0, \infty)$ is a birth-death rate, and $r \in [0, 1]$ is a coupling parameter. The case $r = 0$ corresponds to a critical branching random walk, for which the average number of individuals

per site is preserved. The case $r > 0$ is challenging because the individuals descending from different ancestors are no longer independent.

For the case $r = 0$, the following *dichotomy* holds (where for simplicity we restrict to an irreducible and symmetric $p(\cdot, \cdot)$): if the initial configuration is drawn from a shift-invariant and shift-ergodic random field taking values in \mathbb{N}_0 with a positive and finite mean, then the system in (3.3) locally dies out when $p(\cdot, \cdot)$ is *recurrent*, but converges to a non-trivial equilibrium when $p(\cdot, \cdot)$ is *transient*, both irrespective of the value of q . In the latter case, the equilibrium has the same mean as the initial distribution and has all moments finite.

For the case $r > 0$, the situation is more subtle. It was shown in Greven [17], [18] that there exist $0 < r_2 \leq r_* \leq 1$ such that the system in (3.3) locally dies out when $r > r_*$, but converges to a non-trivial equilibrium when $r < r_*$, and this equilibrium has an *infinite second moment* when $r \geq r_2$ and a *finite second moment* when $r < r_2$. It was conjectured in [18] that $r_2 < r_*$. Since it was shown in [18] that

$$r_* \geq 1 \wedge (q^{-1} \log \tilde{z}_1), \quad r_2 = 1 \wedge (q^{-1} \log \tilde{z}_2), \tag{3.4}$$

Corollary 2.3(a) settles this conjecture when $p(\cdot, \cdot)$ satisfies (2.1) and (2.6) and is strongly transient, and $q > \log \tilde{z}_2 = 2/G(0, 0)$.

4. Three Applications with Disorder

4.1. A polymer in a random potential.

Path measure. Let $S = (S_k)_{k \in \mathbb{N}_0}$ be a random walk on \mathbb{Z}^d , $d \geq 1$, starting at the origin and with transition kernel $p(\cdot, \cdot)$. Write P to denote the law of S . Let $\omega = \{\omega(k, x) : k \in \mathbb{N}_0, x \in \mathbb{Z}^d\}$ be an i.i.d. field of \mathbb{R} -valued non-degenerate random variables with marginal law μ_0 , playing the role of a *random environment*. Write $\mathbb{P} = (\mu_0)^{\otimes [\mathbb{N}_0 \times \mathbb{Z}^d]}$ to denote the law of ω . Assume that

$$M(\lambda) = \mathbb{E}(e^{\lambda \omega(0,0)}) < \infty \quad \forall \lambda \in \mathbb{R}. \tag{4.1}$$

For fixed ω and $n \in \mathbb{N}$, define

$$\frac{dP_n^{\beta, \omega}}{dP}((S_k)_{k=0}^n) = \frac{1}{Z_n^{\beta, \omega}} e^{-H_n^{\beta, \omega}((S_k)_{k=0}^n)} \tag{4.2}$$

with

$$H_n^{\beta, \omega}((S_k)_{k=0}^n) = -\beta \sum_{k=1}^n \omega(k, S_k), \tag{4.3}$$

i.e., $P_n^{\beta, \omega}$ is the Gibbs measure on the set of paths of length $n \in \mathbb{N}$ associated with the Hamiltonian $H_n^{\beta, \omega}$. Here, $\beta \in [0, \infty)$ plays the role of *environment strength* (or “inverse temperature”), while $Z_n^{\beta, \omega}$ is the normalising partition sum. In this model, ω represents a space-time medium of “random charges”

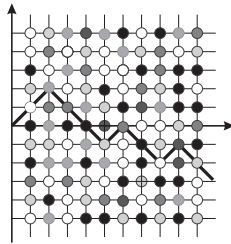


Figure 3. A directed polymer sampling random charges in a halfplane.

with which a directed polymer, described by the space-time path $(k, S_k)_{k=0}^n$, is interacting (see Fig. 3).

Weak vs. strong disorder. Let $\chi_n(\omega) = Z_n^{\beta, \omega} e^{-n \log M(\beta)}$, $n \in \mathbb{N}_0$. It is well known that $\chi(\omega) = (\chi_n(\omega))_{n \in \mathbb{N}_0}$ is a non-negative martingale with respect to the family of sigma-algebras $\mathcal{F}_n = \sigma(\omega(k, x), 0 \leq k \leq n, x \in \mathbb{Z}^d)$, $n \in \mathbb{N}_0$. Hence $\lim_{n \rightarrow \infty} \chi_n(\omega) = \chi_\infty(\omega) \geq 0$ ω -a.s., with $\mathbb{P}(\chi_\infty(\omega) = 0) = 0$ or 1 . This leads to two phases:

$$\begin{aligned} \mathcal{W} &= \{\beta \in [0, \infty) : \chi_\infty(\omega) > 0 \text{ } \omega\text{-a.s.}\}, \\ \mathcal{S} &= \{\beta \in [0, \infty) : \chi_\infty(\omega) = 0 \text{ } \omega\text{-a.s.}\}, \end{aligned} \tag{4.4}$$

which are referred to as the *weak disorder phase* and the *strong disorder phase*, respectively. It was shown in Comets and Yoshida [13] that there is a unique critical value $\beta_* \in [0, \infty]$ (depending on $d, p(\cdot, \cdot)$ and μ_0) such that weak disorder holds for $0 \leq \beta < \beta_*$ and strong disorder holds for $\beta > \beta_*$. Moreover, in the weak disorder phase the paths have a Gaussian scaling limit under the Gibbs measure, while this is not the case in the strong disorder phase. In the strong disorder phase the path tends to localise around the highest values of ω in a narrow space-time tube.

Suppose that $p(\cdot, \cdot)$ is irreducible, symmetric and transient. Abbreviate $\Delta(\beta) = \log M(2\beta) - 2 \log M(\beta)$. Bolthausen [9] observed that

$$\mathbb{E} [\chi_n(\omega)^2] = \mathbb{E} \left[e^{\Delta(\beta) V_n} \right] \quad \text{with} \quad V_n = \sum_{k=1}^n 1_{\{S_k = S'_k\}}, \tag{4.5}$$

where S and S' are two independent random walks with transition kernel $p(\cdot, \cdot)$, and concluded that $\chi(\omega)$ is L^2 -bounded if and only if $\beta < \beta_2$ with $\beta_2 \in (0, \infty]$ the unique solution of

$$\Delta(\beta_2) = \log z_2 \tag{4.6}$$

(with $\beta_2 = \infty$ whenever $\Delta(\infty) \leq \log z_2$). Since

$$\mathbb{P}(\chi_\infty(\omega) > 0) \geq \mathbb{E}[\chi_\infty(\omega)]^2 / \mathbb{E}[\chi_\infty(\omega)^2], \quad \mathbb{E}[\chi_\infty(\omega)] = \chi_0(\omega) = 1, \tag{4.7}$$

it follows that $\beta < \beta_2$ implies weak disorder, i.e., $\beta_* \geq \beta_2$. By a stochastic representation of the size-biased law of $\chi_n(\omega)$, it was shown in Birkner [2] that in fact weak disorder holds if $\beta < \beta_1$ with $\beta_1 \in (0, \infty]$ the unique solution of

$$\Delta(\beta_1) = \log z_1, \tag{4.8}$$

i.e., $\beta_* \geq \beta_1$. Since $\beta \mapsto \Delta(\beta)$ is strictly increasing for any non-degenerate μ_0 satisfying (4.1), it follows from (4.6–4.8) and Corollary 2.2(a) that $\beta_1 > \beta_2$ when $p(\cdot, \cdot)$ satisfies (2.1) and (2.6) and is strongly transient, provided μ_0 is such that $\beta_2 < \infty$. In that case the weak disorder phase contains a subphase for which $\chi(\omega)$ is *not* L^2 -bounded. This disproves a conjecture of Monthus and Garel [21], who argued that $\beta_2 = \beta_*$.

For further details, see den Hollander [20], Chapter 12. Main contributions in the mathematical literature towards understanding the two phases have come from M. Birkner, E. Bolthausen, A. Camanes, P. Carmona, F. Comets, B. Derrida, M.R. Evans, Y. Hu, J.Z. Imbrie, O. Mejane, M. Petermann, M.S.T. Piza, T. Shiga, Ya.G. Sinai, T. Spencer, V. Vargas and N. Yoshida.

4.2. A polymer pinned at an interface.

Path measure. Let $S = (S_k)_{k \in \mathbb{N}_0}$ be a recurrent Markov chain on a countable state space starting at a marked point 0. Write P to denote the law of S . Let K denote the law of the first return time of S to 0, which is assumed to satisfy

$$\lim_{n \rightarrow \infty} \frac{\log K(n)}{\log n} = -\alpha \quad \text{for some } \alpha \in [1, \infty). \tag{4.9}$$

Let $\omega = (\omega_k)_{k \in \mathbb{N}_0}$ be an i.i.d. sequence of \mathbb{R} -valued non-degenerate random variables with marginal law μ_0 , again playing the role of a *random environment*. Write $\mathbb{P} = \mu_0^{\otimes \mathbb{N}_0}$ to denote the law of ω . Assume that

$$M(\lambda) = \mathbb{E}(e^{\lambda \omega_0}) < \infty \quad \forall \lambda \in \mathbb{R}. \tag{4.10}$$

Without loss of generality we take: $\mathbb{E}(\omega_0) = 0, \mathbb{E}(\omega_0^2) = 1$.

For fixed ω and $n \in \mathbb{N}$, define, in analogy with (4.2–4.3),

$$\frac{dP_n^{\beta, h, \omega}}{dP}((S_k)_{k=0}^n) = \frac{1}{Z_n^{\beta, h, \omega}} e^{-H_n^{\beta, h, \omega}((S_k)_{k=0}^n)} \tag{4.11}$$

with

$$H_n^{\beta, h, \omega}((S_k)_{k=1}^n) = - \sum_{k=1}^n (\beta \omega_k - h) 1_{\{S_k=0\}}, \tag{4.12}$$

where $\beta \in [0, \infty)$ again plays the role of *environment strength*, and $h \in [0, \infty)$ the role of *environment bias*. This models a directed polymer interacting with “random charges” at an interface (see Fig. 4). A key example is when S is simple random walk on \mathbb{Z} , which corresponds to the case $\alpha = \frac{3}{2}$.

The *quenched free energy per monomer* $f^{\text{que}}(\beta, h) = \lim_{n \rightarrow \infty} \frac{1}{n} \log Z_n^{\beta, h, \omega}$ is constant ω -a.s. (a property called self-averaging), and has two phases

$$\begin{aligned} \mathcal{L} &= \{(\beta, h) : f^{\text{que}}(\beta, h) > 0\}, \\ \mathcal{D} &= \{(\beta, h) : f^{\text{que}}(\beta, h) = 0\}, \end{aligned} \tag{4.13}$$

which are referred to as the *localised phase* and the *delocalised phase*. These two phases are the result of a competition between entropy and energy: by staying close to the interface the polymer loses entropy, but at the same time it gains energy because it can more easily pick up large charges at the interface. The lower bound comes from the strategy where the path spends all its time above the interface, i.e., $S_k > 0$ for $1 \leq k \leq n$. Indeed, in that case $H_n^{\beta, h, \omega}((S_k)_{k=0}^n) = 0$, and since $\log[\sum_{m>n} K(m)] \sim -(\alpha - 1) \log n$ as $n \rightarrow \infty$, the cost of this strategy under P is negligible on an exponential scale.

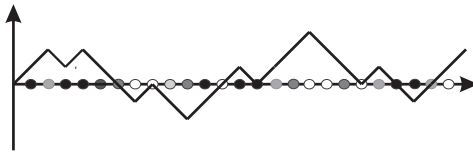


Figure 4. A directed polymer sampling random charges at an interface.

The associated *quenched critical curve* is

$$h_c^{\text{que}}(\beta) = \inf\{h : f^{\text{que}}(\beta, h) = 0\}, \quad \beta \in [0, \infty). \tag{4.14}$$

Both f^{que} and h_c^{que} are unknown. However, their *annealed* counterparts

$$f^{\text{ann}}(\beta, h) = \lim_{n \rightarrow \infty} \frac{1}{n} \log \mathbb{E}(Z_n^{\beta, h, \omega}), \quad h_c^{\text{ann}}(\beta) = \inf\{h : f^{\text{ann}}(\beta, h) = 0\}, \tag{4.15}$$

can be computed explicitly, because they correspond to the degenerate case where $\omega_k = (1/\beta) \log M(\beta)$, $k \in \mathbb{N}_0$. In particular, $h_c^{\text{ann}}(\beta) = \log M(\beta)$. Since $f^{\text{que}} \leq f^{\text{ann}}$, it follows that $h_c^{\text{que}} \leq h_c^{\text{ann}}$.

Disorder relevance vs. irrelevance. For a given choice of K , μ_0 and β , the disorder is said to be *relevant* when $h_c^{\text{que}}(\beta) < h_c^{\text{ann}}(\beta)$ and *irrelevant* when $h_c^{\text{que}}(\beta) = h_c^{\text{ann}}(\beta)$. Various papers have appeared in the literature containing various conditions under which relevant disorder, respectively, irrelevant disorder occurs, based on a variety of different estimation techniques. Main contributions in the mathematical literature have come from K. Alexander, B. Derrida, G. Giacomin, H. Lacoin, V. Sidoravicius, F.L. Toninelli and N. Zygouras. For overviews, see Giacomin [16], Chapter 5, and den Hollander [20], Chapter 11.

In work in progress with D. Cheliotis [12] a different view is taken. Namely, with the help of Theorems 1.1–1.2 for the choice

$$E = \mathbb{R}, \quad \nu = \mu_0, \quad \rho = K, \tag{4.16}$$

the following variational formulas are derived for h_c^{que} and h_c^{ann} .

Theorem 4.1. For all $\beta \in [0, \infty)$,

$$\begin{aligned} h_c^{\text{que}}(\beta) &= \sup_{Q \in \mathcal{C}} [\beta \Phi(Q) - I^{\text{que}}(Q)], \\ h_c^{\text{ann}}(\beta) &= \sup_{Q \in \mathcal{C}} [\beta \Phi(Q) - I^{\text{ann}}(Q)], \end{aligned} \tag{4.17}$$

where

$$\mathcal{C} = \left\{ Q \in \mathcal{P}^{\text{inv}}(\tilde{\mathbb{R}}^{\mathbb{N}}) : \int_{\mathbb{R}} |x| (\pi_{1,1}Q)(dx) < \infty \right\}, \quad \Phi(Q) = \int_{\mathbb{R}} x (\pi_{1,1}Q)(dx), \tag{4.18}$$

with $\pi_{1,1}Q$ the projection of Q onto \mathbb{R} , i.e., the law of the first letter of the first word.

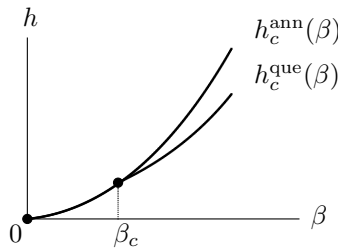


Figure 5. Critical curves for the pinned polymer

It is shown in [12] that a comparison of the two variational formulas in Theorem 4.1 yields the following *necessary and sufficient* condition for disorder relevance.

Corollary 4.2. For every $\beta \in [0, \infty)$,

$$h_c^{\text{que}}(\beta) < h_c^{\text{ann}}(\beta) \iff I^{\text{que}}(Q_\beta) > I^{\text{ann}}(Q_\beta), \tag{4.19}$$

where $Q_\beta = q_{K,\beta}^{\otimes \mathbb{N}}$ is the unique maximiser of the annealed variational formula in (4.17), given by

$$q_{K,\beta}((x_1, \dots, x_n)) = K(n) \mu_\beta(x_1) \cdots \mu_\beta(x_n), \quad n \in \mathbb{N}, x_1, \dots, x_n \in \mathbb{R}, \tag{4.20}$$

with μ_β the law obtained from μ_0 by tilting:

$$d\mu_\beta(x) = \frac{1}{M(\beta)} e^{\beta x} d\mu_0(x), \quad x \in \mathbb{R}. \tag{4.21}$$

As shown in [12], an immediate consequence of the variational characterisation in Corollary 4.2 is that there is a unique critical inverse temperature (see Fig. 5).

Corollary 4.3. *For all μ_0 and K there exists a $\beta_c = \beta_c(\mu_0, K) \in [0, \infty]$ such that*

$$h_c^{\text{que}}(\beta) \begin{cases} = h_c^{\text{ann}}(\beta) & \text{if } \beta \in [0, \beta_c], \\ < h_c^{\text{ann}}(\beta) & \text{if } \beta \in (\beta_c, \infty). \end{cases} \quad (4.22)$$

Moreover, necessary and sufficient conditions on μ_0 and K can be derived under which $\beta_c = 0$, $\beta_c \in (0, \infty)$, respectively, $\beta_c = \infty$, providing a *full classification of disorder relevance*.

4.3. A copolymer near a selective interface.

Path measure. Let S be a recurrent random walk on \mathbb{Z} . Keep (4.9–4.11), but change the Hamiltonian in (4.12) to

$$H_n^{\beta, h, \omega}((S_k)_{k=1}^n) = -\beta \sum_{k=1}^n (\omega_k + h) \text{sign}(S_k). \quad (4.23)$$

This model was introduced in Garel, Huse, Leibler and Orland [15]. For the special case where $\mu_0 = \frac{1}{2}(\delta_{-1} + \delta_{+1})$, it models a copolymer consisting of a random concatenation of hydrophobic and hydrophilic monomers (represented by ω), living in the vicinity of a linear interface that separates oil (above the interface) and water (below the interface) as solvents. The polymer is modelled as a two-dimensional directed path $(k, S_k)_{k \in \mathbb{N}_0}$. The Hamiltonian in (4.23) is such that hydrophobic monomers in oil ($\omega_k = +1$, $S_k > 0$) and hydrophilic monomers in water ($\omega_k = -1$, $S_k < 0$) receive a negative energy, while the other two combinations receive a positive energy.

The *quenched free energy per monomer*, $f^{\text{que}}(\beta, h) = \lim_{n \rightarrow \infty} \frac{1}{n} \log Z_n^{\beta, h, \omega}$ ω -a.s., again has two phases (see Fig. 6)

$$\begin{aligned} \mathcal{L} &= \{(\beta, h) : g^{\text{que}}(\beta, h) > 0\}, \\ \mathcal{D} &= \{(\beta, h) : g^{\text{que}}(\beta, h) = 0\}, \end{aligned} \quad (4.24)$$

where $g^{\text{que}}(\beta, h) = f^{\text{que}}(\beta, h) - \beta h$. These two phases are again the result of a competition between entropy and energy: by staying close to the interface the copolymer loses entropy, but it gains energy because it can more easily switch between the two sides of the interface in an attempt to place as many monomers as possible in their preferred solvent. The lower bound again comes from the strategy where the path spends all its time above the interface, i.e., $S_k > 0$ for $1 \leq k \leq n$. Indeed, in that case $\text{sign}(S_k) = +1$ for $1 \leq k \leq n$, resulting in $H_n^{\beta, h, \omega}((S_k)_{k=0}^n) = -\beta hn[1 + o(1)]$ ω -a.s. as $n \rightarrow \infty$ by the strong law of large numbers for ω . Since $\log[\sum_{m>n} K(m)] \sim -(\alpha - 1) \log n$ as $n \rightarrow \infty$, the cost of this strategy under P is again negligible on an exponential scale.

The associated *quenched critical curve* is

$$h_c^{\text{que}}(\beta) = \inf\{h : g^{\text{que}}(\beta, h) = 0\}, \quad \beta \in [0, \infty). \quad (4.25)$$

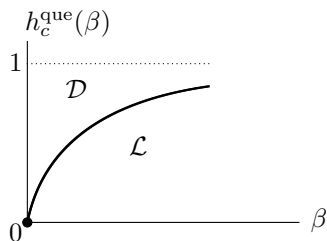


Figure 6. Quenched critical curve for the copolymer.

Both g^{que} and h_c^{que} are unknown. Their *annealed* counterparts $g^{\text{ann}}(\beta, h)$ and $h_c^{\text{ann}}(\beta) = \inf\{h : g^{\text{ann}}(\beta, h) = 0\}$ can again be computed explicitly.

The copolymer model is *much harder* than the pinning model described in Section 4.2, because the disorder ω is felt not just *at* the interface but along the *entire* polymer chain. The following bounds are known:

$$\left(\frac{2}{\alpha}\beta\right)^{-1} \log M\left(\frac{2}{\alpha}\beta\right) \leq h_c^{\text{que}}(\beta) \leq h_c^{\text{ann}}(\beta) = (2\beta)^{-1} \log M(2\beta) \quad \forall \beta > 0. \tag{4.26}$$

The upper bound was proved in Bolthausen and den Hollander [10], and comes from the observation that $f^{\text{que}} \leq f^{\text{ann}}$. The lower bound was proved in Bodineau and Giacomin [7], and comes from strategies where the copolymer dips below the interface (into the water) during rare stretches in ω where the empirical density is sufficiently biased downwards (i.e., where the polymer is sufficiently hydrophilic).

Main contributions in the mathematical literature towards understanding the two phases have come from M. Biskup, T. Bodineau, E. Bolthausen, F. Caravenna, G. Giacomin, M. Gubinelli, F. den Hollander, H. Lacoïn, N. Madras, E. Orlandini, A. Reznitser, Ya.G. Sinai, C. Soteros, C. Tesi, F.L. Toninelli, S.G. Whittington and L. Zambotti. For overviews, see Giacomin [16], Chapters 6–8, and den Hollander [20], Chapter 9.

Strict bounds. Toninelli [22] proved that the upper bound in (4.26) is strict for μ_0 with unbounded support and large β . This was later extended by Bodineau, Giacomin, Lacoïn and Toninelli [8] to arbitrary μ_0 and β . The latter paper also proves that the lower bound in (4.26) is strict for small β . The proofs are based on fractional moment estimates of the partition sum and on finding appropriate localisation strategies.

In work in progress with E. Bolthausen [11], Theorems 1.1–1.2 are used, for the same choice as in (4.16), to obtain the following characterisation of the critical curves.

Theorem 4.4. *For every $\beta \in [0, \infty)$,*

$$h = h_c^{\text{que}}(\beta) \iff S^{\text{que}}(\beta, h) = 0, \tag{4.27}$$

$$h = h_c^{\text{ann}}(\beta) \iff S^{\text{ann}}(\beta, h) = 0, \tag{4.28}$$

with

$$S^{\text{que}}(\beta, h) = \sup_{Q \in \mathcal{P}^{\text{inv, fin}}(\tilde{\mathbb{R}}^{\mathbb{N}})} [\Phi_{\beta, h}(Q) - I^{\text{que}}(Q)], \tag{4.29}$$

$$S^{\text{ann}}(\beta, h) = \sup_{Q \in \mathcal{P}^{\text{inv, fin}}(\tilde{\mathbb{R}}^{\mathbb{N}})} [\Phi_{\beta, h}(Q) - I^{\text{ann}}(Q)], \tag{4.30}$$

where

$$\Phi_{\beta, h}(Q) = \int_{\tilde{\mathbb{R}}} (\pi_1 Q)(dy) \log \phi_{\beta, h}(y), \quad \phi_{\beta, h}(y) = \frac{1}{2} \left(1 + e^{-2\beta h \tau(y) - 2\beta \sigma(y)} \right), \tag{4.31}$$

with $\tau(y)$ and $\sigma(y)$ the length, respectively, the sum of the letters in the word y .

The variational formulas in Theorem 4.4 are more involved than those in Theorem 4.1 for the pinning model. The annealed variational formula in (4.30) can again be solved explicitly, the quenched variational formula in (4.29) cannot.

In [11] the strict upper bound in (4.26), which was proved in [8], is deduced from Theorem 4.4 via a criterion analogous to Corollary 4.2.

Corollary 4.5. $h_c^{\text{que}}(\beta) < h_c^{\text{ann}}(\beta)$ for all μ_0 and $\beta > 0$.

We are presently trying to prove that also the lower bound in (4.26) holds in full generality.

Weak interaction limit. A point of heated debate has been the slope of the quenched critical curve at $\beta = 0$,

$$\lim_{\beta \downarrow 0} \frac{1}{\beta} h_c^{\text{que}}(\beta) = K_c, \tag{4.32}$$

which is believed to be *universal*, i.e., to only depend on α and to be robust against small perturbations of the interaction Hamiltonian in (4.23). The existence of the limit was proved in Bolthausen and den Hollander [10]. The bounds in (4.26) imply that $K_c \in [\alpha^{-1}, 1]$, and various claims were made in the literature arguing in favor of $K_c = \alpha^{-1}$, respectively, $K_c = 1$. In Bodineau, Giacomin, Lacoïn and Toninelli [8] it is shown that $K_c \in (\alpha^{-1}, 1)$ under some additional assumptions on the excursion length distribution $K(\cdot)$ satisfying (4.9). We are presently trying to extend this result to arbitrary $K(\cdot)$ with the help of a space-time continuous version of the large deviation principles in Theorems 1.1–1.2.

5. Closing Remarks

The large deviation principles in Theorems 1.1–1.2 are a powerful new tool to analyse the large space-time behaviour of interacting stochastic systems based on excursions of random walks and Markov chains. Indeed, they *open up a*

window with a variational view, since they lead to explicit variational formulas for the critical curves that are associated with the phase transitions occurring in these systems. They are flexible, but at the same time technically demanding.

A key open problem is to find a good formula for $I^{\text{que}}(Q)$ when $m_Q = \infty$ (recall (1.11–1.12)), e.g. when Q is Gibbsian.

References

- [1] Q. Berger and F.L. Toninelli, On the critical point of the random walk pinning model in dimension $d = 3$, preprint 2009.
- [2] M. Birkner, A condition for weak disorder for directed polymers in random environment, *Electron. Comm. Probab.* 9 (2004) 22–25.
- [3] M. Birkner, A. Greven, F. den Hollander, Quenched large deviation principle for words in a letter sequence, *Probab. Theory Relat. Fields* 147 (2010), available online.
- [4] M. Birkner, A. Greven, F. den Hollander, Collision local time of transient random walks and intermediate phases in interacting stochastic systems, preprint 2008.
- [5] M. Birkner and R. Sun, Annealed vs quenched critical points for a random walk pinning model, preprint 2008.
- [6] M. Birkner and R. Sun, Disorder relevance for the random walk pinning model in $d = 3$, preprint 2009.
- [7] T. Bodineau and G. Giacomin, On the localization transition of random copolymers near selective interfaces, *J. Stat. Phys.* 117 (2004) 17–34.
- [8] T. Bodineau, G. Giacomin, H. Lacoin and F.L. Toninelli, Copolymers at selective interfaces: new bounds on the phase diagram, *J. Stat. Phys.* 132 (2008) 603–626.
- [9] E. Bolthausen, A note on the diffusion of directed polymers in a random environment, *Commun. Math. Phys.* 123 (1989) 529–534.
- [10] E. Bolthausen and F. den Hollander, Localization transition for a polymer near an interface, *Ann. Probab.* 25 (1997) 1334–1366.
- [11] E. Bolthausen and F. den Hollander, A copolymer near a linear interface: improved bounds on the critical curve, preprint 2010.
- [12] D. Cheliotis and F. den Hollander, Variational characterization of the critical curve in random pinning models, preprint 2010.
- [13] F. Comets and N. Yoshida, Directed polymers in random environment are diffusive at weak disorder, *Ann. Probab.* 34 (2006) 1746–1770.
- [14] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications* (2nd. ed.), Springer, New York, 1998.
- [15] T. Garel, D.A. Huse, S. Leibler and H. Orland, Localization transition of random chains at interfaces, *Europhys. Lett.* 8 (1989) 9–13.
- [16] G. Giacomin, *Random Polymer Models*, Imperial College Press, World Scientific, London, 2007.

- [17] A. Greven, Phase transition for the coupled branching process, Part I: The ergodic theory in the range of second moments, *Probab. Theory Relat. Fields* 87 (1991) 417–458.
- [18] A. Greven, On phase transitions in spatial branching systems with interaction, in: *Stochastic Models* (L.G. Gorostiza and B.G. Ivanoff, eds.), CMS Conference Proceedings 26 (2000) 173–204.
- [19] A. Greven and F. den Hollander, Phase transitions for the long-time behaviour of interacting diffusions, *Ann. Probab.* 35 (2007) 1250–1306.
- [20] F. den Hollander, *Random Polymers*, Lecture Notes in Mathematics 1974, Springer, Berlin, 2009.
- [21] C. Monthus and T. Garel, Freezing transition of the directed polymer in a 1+d random medium: Location of the critical temperature and unusual critical properties, *Phys. Rev. E* 74 (2006) 011101.
- [22] F.L. Toninelli, Disordered pinning models and copolymers: beyond annealed bounds, *Ann. Appl. Probab.* 18 (2008) 1569–1587.

Time and Chance Happeneth to Them all: Mutation, Selection and Recombination

Steven N. Evans*

Abstract

Many multi-cellular organisms exhibit remarkably similar patterns of aging and mortality. Because this phenomenon appears to arise from the complex interaction of many genes, it has been a challenge to explain it quantitatively as a response to natural selection. We survey attempts by the author and his collaborators to build a framework for understanding how mutation, selection and recombination acting on many genes combine to shape the distribution of genotypes in a large population. A genotype drawn at random from the population at a given time is described by a Poisson random measure on the space of loci and its distribution is characterized by the associated intensity measure. The intensity measures evolve according to a continuous-time measure-valued dynamical system. We present general results on the existence and uniqueness of this dynamical system and how it arises as a limit of discrete generation systems. We also discuss existence of equilibria.

Mathematics Subject Classification (2010). Primary 60G57, 92D15; Secondary 37N25, 60G55, 92D10.

Keywords. Measure-valued, dynamical system, population genetics, Poisson random measure, Wasserstein metric, equilibrium

1. Introduction

One of the main goals of mathematical population genetics is to satisfactorily model the biological mechanisms of mutation, selection and recombination and understand how they interact over time to change the distribution of genotypes (and hence phenotypic traits) in a population.

*Research supported in part by grants DMS-04-05778 and DMS-09-07630 from the National Science Foundation (U.S.A.)

Departments of Statistics and Mathematics, University of California at Berkeley, 367 Evans Hall, Berkeley, CA 94720-3860, U.S.A. E-mail: evans@stat.berkeley.edu.

Many traits are believed to result from complex, non-additive interactions between large numbers of mildly deleterious alleles that are simultaneously being slowly forced out of the population by natural selection and reintroduced by recurring mutations. For example, the Medawar–Williams–Hamilton [Med52, Wil57, Ham66] explanation of the evolution of aging invokes this mechanism (see, also, [Cha94, Cha01] and the introductory discussions in [SEW05, ESW06, WES08, WSE10] — excellent references for mathematical population genetics in general and its role in evolutionary theory are [Bür00, Ewe04]).

As noted in [Hol95, FK00], a quantitative understanding of the how patterns of senescence and mortality have evolved requires a tractable quantitative description of the changes wrought through time by the competing pressures of mutation and selection acting on an ensemble of interconnected genes. Some attempts that have been made in this direction are amenable to analysis, but they are, as observed in the review [PC98], too stylized and simplistic. Other approaches, particularly [BT91, KJB02], are flexible enough to accommodate essentially arbitrary mechanisms of selection, mating, linkage, mutation and phenotypic effects, and hence are extremely useful for doing numerical computations; but they incorporate too much explicit detail to be usable for theoretical investigations.

This paper is an overview of research over the last several years by the author and his collaborators, Aubrey Clayton, David Steinsaltz and Ken Wachter, to develop a framework that occupies the middle ground between perspectives that are too synoptic and ones that are overly burdened with specifics. This work began in [SEW05] and has been continued in [ESW06, WES08, CE09, WSE10]. When detailed proofs are not given, they can, unless otherwise noted, be found in [ESW06].

2. Ingredients

The key assumptions behind our model are:

- the population is infinite,
- the genome may consist of infinitely many (even uncountably many) *loci*,
- each individual has two parents,
- mating is random,
- the genotype of an individual is a random mosaic of the genotypes of its parents produced by the process of recombination,
- an individual has one copy of each gene rather than copies from each of its two parents (individuals are *haploid*),

- starting with an ancestral *wild type* mutant alleles only accumulate down any lineage (there is no back-mutation),
- fitness is calculated for individuals rather than for mating pairs,
- a genotype becomes less fit when it accumulates additional mutant alleles,
- recombination acts on a faster time scale than mutation or selection.

A consequence of these assumptions is that if we denote by \mathcal{M} the collection of loci in the portion of the genome that is of interest to us, then the genotype of an individual may be identified with the collection of loci at which mutant alleles are present. We allow \mathcal{M} to be quite general (in particular, we do not necessarily think of it \mathcal{M} a finite collection of physical DNA base positions or a finite collection of genes) and it is mathematically convenient to assume that \mathcal{M} is an arbitrary complete, separable metric space. A genotype is then an element of the space \mathcal{G} of integer-valued finite Borel measures on \mathcal{M} : the genotype $\sum_i \delta_{m_i}$, where δ_m is the unit point mass at the locus $m \in \mathcal{M}$, has mutations away from the ancestral wild type at loci m_1, m_2, \dots . The wild genotype is thus the null measure.

A further consequence of the assumptions is that the composition of the population at some time t is completely described by a probability measure P_t on \mathcal{G} , where $P_t(G)$ for some subset $G \subseteq \mathcal{G}$ represents the proportion of individuals in the population at time t that have genotypes belonging to G .

Fitnesses of genotypes are defined via a *selective cost function* $S : \mathcal{G} \rightarrow \mathbb{R}$. The difference $S(g') - S(g'')$ for $g', g'' \in \mathcal{G}$ is the difference in the rate of sub-population growth between the sub-population of individuals with genotype g'' and the sub-population of individuals with genotype g' . We make the normalizing assumption $S(0) = 0$ and suppose that

$$S(g + h) \geq S(h), \quad g, h \in \mathcal{G}, \quad (2.1)$$

to reflect assumption that genotypes with more accumulated mutations are less fit.

Example 2.1. Selective cost functions of the following form are relevant to the study of aging and mortality.

Suppose that the space of loci \mathcal{M} is general. Write $\ell_x(g)$ for the probability that an individual with genotype $g \in \mathcal{G}$ lives beyond age $x \in \mathbb{R}_+$. At age x , the corresponding *cumulative hazard* and *hazard function* are thus $-\log \ell_x(g)$ and $d/dx(-\log \ell_x(g))$, respectively. Assume that the infinitesimal rate that an individual at age $x \in \mathbb{R}_+$ has offspring is $f(x)$, independently of the individual's genotype, where $f : \mathbb{R}_+ \rightarrow \mathbb{R}_+$ is bounded. For individuals with genotype g , the size of the next generation relative to the current one is thus $\int_0^\infty f(x) \ell_x(g) dx$. The corresponding selective cost of genotype g is thus

$$S(g) = \int_0^\infty f(x) \ell_x(0) dx - \int_0^\infty f(x) \ell_x(g) dx. \quad (2.2)$$

Assume further that there is a constant background hazard λ and that an ancestral mutation at locus $m \in \mathcal{M}$ contributes an increment $\theta(m, x)$ to the hazard function at age x so that

$$\ell_x(g) = \exp\left(-\lambda x - \int_{\mathcal{M}} \theta(m, x) g(dm)\right). \tag{2.3}$$

Observe that the resulting cost function S has the monotonicity property (2.1). Moreover, S is bounded

$$\sup_{g \in \mathcal{G}} S(g) < \infty \tag{2.4}$$

and *concave* in the sense that

$$S(g + h + k) - S(g + h) \leq S(g + k) - S(g) \text{ for all } g, h, k \in \mathcal{G}; \tag{2.5}$$

that is, the marginal cost of an additional mutation decreases as more mutations are added to the genotype.

We are most interested in a continuous time model, but in order to justify the form of such a model we first consider a setting with discrete, non-overlapping generations. We imagine that in going from one generation to the next, the population is transformed successively by the effects of selection, mutation and recombination. Recall that the population at any time corresponds to a probability measure on the genotype space \mathcal{G} and so the actions of these mechanisms are each described by a map from \mathcal{G} to itself.

The effect of selection in one generation transforms a probability measure P to $\mathfrak{S}P$, where $\mathfrak{S}P[F]$, the integral of a bounded measurable function F against $\mathfrak{S}P$, is given by

$$\mathfrak{S}P[F] := \frac{\int_{\mathcal{G}} e^{-S(g)} F(g) P(dg)}{\int_{\mathcal{G}} e^{-S(g)} P(dg)} = \frac{P[e^{-S}F]}{P[e^{-S}]}. \tag{2.6}$$

We assume that new mutations from the ancestral type appear in some subset A of the locus space \mathcal{M} at rate $\nu(A)$, where ν is a finite measure on \mathcal{M} . Thus, the additional load of mutations appearing in one generation is distributed as a Poisson random measure on \mathcal{M} with intensity measure ν . Denoting such a random measure by X^ν , the action of mutation transforms a probability measure P to $\mathfrak{M}P$, where

$$\mathfrak{M}P[F] := \int_{\mathcal{G}} \mathbb{E}[F(g + X^\nu)] P(dg). \tag{2.7}$$

A recombination event takes two genotypes $g', g'' \in \mathcal{G}$ from the population and replaces the genotype g' by the genotype g defined by $g(A) := g'(A \cap R) + g''(A \cap R^c)$, where the ‘‘segregating set’’ $R \subseteq \mathcal{M}$ for the recombination event is

chosen according to a probability measure \mathcal{R} on the set $\mathcal{B}(\mathcal{M})$ of Borel subsets of \mathcal{M} and R^c denotes the complement of R . We may suppose without loss of generality that \mathcal{R} is *symmetric* in the sense that

$$\mathcal{R}(A) = \mathcal{R}(\{R \in \mathcal{B}(\mathcal{M}) : R^c \in A\}). \tag{2.8}$$

According to our assumption of random mating, the action of recombination transforms a probability measure P to $\mathfrak{R}P$, where

$$\mathfrak{R}P[F] := \int_{\mathcal{B}(\mathcal{M})} \int_{\mathcal{G}} \int_{\mathcal{G}} F(g'(\cdot \cap R) + g''(\cdot \cap R^c)) P(dg') P(dg'') \mathcal{R}(dR). \tag{2.9}$$

To make the definition of \mathfrak{R} rigorous we need an appropriate theory of random Borel sets, but this is provided by [Ken74].

Thus, if the population in generation 0 is described by the probability measure P_0 , then the population in generation k is described by $(\mathfrak{R}\mathfrak{M}\mathfrak{S})^k P_0$.

Observe that the operators \mathfrak{S}^k and \mathfrak{M}^k are of the same form as \mathfrak{S} and \mathfrak{M} with S and ν replaced by kS and $k\nu$, respectively.

The operator \mathfrak{R}^k is also easy to understand. Let R_1, \dots, R_k be independent identically distributed random subsets of \mathcal{M} with common distribution \mathcal{R} . Construct the random partition $\{A_1, \dots, A_L\}$ of \mathcal{M} that consists of the non-empty sets of the form $\tilde{R}_1 \cap \dots \cap \tilde{R}_k$, where \tilde{R}_i is either R_i or R_i^c (so that $L \leq 2^k$). Then,

$$\mathfrak{R}^k P[F] = \mathbb{E} \left[\int_{\mathcal{G}^\kappa} F \left(\sum_{j=1}^L g_j(\cdot \cap A_j) \right) P^{\otimes k}(d\mathbf{g}) \right]. \tag{2.10}$$

Observe that if P is the distribution of a Poisson random measure, then $\mathfrak{R}^k P = P$. Moreover, it is reasonable for suitable more general P that if the shuffling together of genotypes induced by the recombination mechanism is suitably thorough, then $\mathfrak{R}^k P$ should converge in some sense as $k \rightarrow \infty$ to the distribution of a Poisson random measure with intensity measure μP . Of course, this won't hold for all probability measures P ; in particular, it fails if μP is diffuse but P puts positive mass on the set of elements of \mathcal{G} with atoms of size greater than one.

Unfortunately, the operators \mathfrak{S} , \mathfrak{M} and \mathfrak{R} do not commute, and so the above observations do not translate into a similarly simple description of $(\mathfrak{R}\mathfrak{M}\mathfrak{S})^k$.

We now incorporate our assumption that recombination acts on a faster time scale than mutation and selection by considering a sequence of models indexed by the positive integers in which the recombination operator stays fixed but in the n^{th} model the mutation intensity measure ν is replaced by ν/n and the selective cost S is replaced by S/n . Denote the corresponding selection and mutation operators by \mathfrak{M}_n and \mathfrak{S}_n .

Note for any probability measure P on \mathcal{G} that

$$\lim_{n \rightarrow \infty} n(\mathfrak{M}_n P[F] - P[F]) = \int_{\mathcal{G}} \left(\int_{\mathcal{M}} F(g + \delta_m) - F(g) \nu(dm) \right) P(dg) \quad (2.11)$$

and

$$\lim_{n \rightarrow \infty} n(\mathfrak{S}_n P[F] - P[F]) = P[S] P[F] - P[S \cdot F]. \quad (2.12)$$

for suitable functions F . In particular, if we consider a linear F of the form $F(g) := \int_{\mathcal{M}} \varphi(m) g(dm)$ for some function $\varphi : \mathcal{M} \rightarrow \mathbb{R}$, so that $P[F] = \mu P[\varphi]$, where μP is the intensity measure associated with P , then

$$\lim_{n \rightarrow \infty} n(\mu \mathfrak{M}_n P[\varphi] - \mu P[\varphi]) = \int_{\mathcal{M}} \varphi(m) \nu(dm). \quad (2.13)$$

Similarly, if P is the distribution of a Poisson random measure, then

$$\begin{aligned} \lim_{n \rightarrow \infty} n(\mu \mathfrak{S}_n P[\varphi] - \mu P[\varphi]) &= P[S] \mu P[\varphi] - \int_{\mathcal{M}} P[S(\cdot + \delta_m)] \varphi(m) \mu P(dm) \\ &= - \int_{\mathcal{M}} (P[S(\cdot + \delta_m)] - P[S]) \varphi(m) \mu P(dm), \end{aligned} \quad (2.14)$$

where we have used Campbell’s theorem, which says that if π is a finite measure on \mathcal{M} , then

$$\mathbb{E}[G(X^\pi) X^\pi[\psi]] = \int_{\mathcal{M}} \mathbb{E}[G(X^\pi + \delta_m)] \psi(m) \pi(dm) \quad (2.15)$$

for a Poisson random measure X^π with intensity π and bounded Borel functions $G : \mathcal{G} \rightarrow \mathbb{R}$ and $\psi : \mathcal{M} \rightarrow \mathbb{R}$. Note also, that if P is the distribution of Poisson random measure, then the same is true of $\mathfrak{M}_n P$ and $\mathfrak{R}P$, whereas $\mathfrak{S}_n P$ is typically not be Poisson unless $S(g + h) = S(g) + S(h)$ for $g, h \in \mathcal{G}$.

Given these observations, it appears reasonable that if P_0 is the distribution of a Poisson random measure on \mathcal{M} , then for $t > 0$ the probability measure $(\mathfrak{R}\mathfrak{M}_n \mathfrak{S}_n)^{\lfloor nt \rfloor} P_0$ should converge in a suitable sense to a probability measure P_t that is also the distribution of a Poisson random measure. Moreover, if we write $\rho_t := \mu P_t$ for the the intensity measure of P_t (so that ρ_t is a finite measure on the space \mathcal{M} of loci), then $(\rho_t)_{t \geq 0}$ should satisfy an evolution equation that we may write informally as

$$\frac{d}{dt} \rho_t(dm) = \nu(dm) - \mathbb{E}[S(X^{\rho_t} + \delta_m) - S(X^{\rho_t})] \rho_t(dm). \quad (2.16)$$

3. Rigorous Definition of the Model

In this section we give a precise meaning to the somewhat heuristic equation (2.16) that describes the evolution of the family $(P_t)_{t \geq 0}$ of distributions of Poisson random measures on \mathcal{M} via a dynamical system for the associated intensity measures $\rho_t = \mu P_t$.

Definition 3.1. Denote by \mathcal{H} the space of finite signed Borel measures on \mathcal{M} . Let \mathcal{H}^+ be the subset of \mathcal{H} consisting of non-negative measures.

Definition 3.2. Given a metric space (E, d) , let Lip be the space of functions $f : E \rightarrow \mathbb{R}$ such that

$$\|f\|_{\text{Lip}} := \sup_x |f(x)| + \sup_{x \neq y} \frac{|f(x) - f(y)|}{d(x, y)} < \infty. \quad (3.1)$$

Define the Wasserstein norm $\|\cdot\|_{\text{Was}}$ on the space of finite signed Borel measures on (E, d) by

$$\|\pi\|_{\text{Was}} := \sup \{|\pi[f]| : \|f\|_{\text{Lip}} \leq 1\}. \quad (3.2)$$

We note that there is a huge literature on the metric induced by the Wasserstein norm and related metrics on spaces of measures (see, for example, [EK86, Rac91, RR98, AGS05, Vil03, Vil09]).

Definition 3.3. Define $F : \mathcal{M} \times \mathcal{H}^+ \rightarrow \mathbb{R}_+$ by

$$F_\pi(m) := \mathbb{E}[S(X^\pi + \delta_m) - S(X^\pi)] \text{ for } m \in \mathcal{M} \text{ and } \pi \in \mathcal{H}^+, \quad (3.3)$$

and define the non-linear operator $D : \mathcal{H}^+ \rightarrow \mathcal{H}^+$ by setting

$$\frac{d(D\pi)}{d\pi}(m) = F_\pi(m). \quad (3.4)$$

Formally, a solution to (2.16) is an \mathcal{H}^+ -valued function ρ that is continuous with respect to the metric induced by the Wasserstein norm and satisfies

$$\rho_t = \rho_0 + t\nu - \int_0^t D\rho_s ds \quad (3.5)$$

for all $t \geq 0$.

Equation (3.5) involves the integration of a measure-valued function, and such an integral can have a number of different meanings (see, for example, [DU77]). We require only the following notion: If $\eta : \mathbb{R}_+ \rightarrow \mathcal{H}$ is a Borel function, then for $t \geq 0$ the integral $\mathcal{I}_t = \int_0^t \eta_s ds$ is the element of \mathcal{H} satisfying

$$\mathcal{I}_t(A) = \int_0^t \eta_s(A) ds \text{ for every Borel } A \subseteq \mathcal{M}. \quad (3.6)$$

This integral certainly exists (and is unique) if the function η is continuous in the Wasserstein metric.

Example 3.4. Note that if $\pi \in \mathcal{H}^+$, then, in the notation of Example 2.1,

$$\mathbb{E}[\ell_x(X^\pi)] = \exp\left(-\lambda x - \int_{\mathcal{M}} \left(1 - e^{-\theta(m,x)}\right) \pi(dm)\right) \quad (3.7)$$

by a standard fact about the Laplace functional of a Poisson random measure, and hence for the selective cost function of Example 2.1

$$\begin{aligned} F_\pi(m') &= \int_0^\infty \left(1 - e^{-\theta(m',x)}\right) f(x) \\ &\quad \times \exp\left(-\lambda x - \int_{\mathcal{M}} \left(1 - e^{-\theta(m'',x)}\right) \pi(dm'')\right) dx \end{aligned} \quad (3.8)$$

for $m' \in \mathcal{M}$.

Theorem 3.5. Fix a mutation measure $\nu \in \mathcal{H}^+$ and a selective cost $S : \mathcal{G} \rightarrow \mathbb{R}_+$ that, along with the standing conditions

- $S(0) = 0$,
- $S(g) \leq S(g+h)$ for all $g, h \in \mathcal{G}$,

also satisfies the Lipschitz condition

- for some constant K , $|S(g) - S(h)| \leq K \|g - h\|_{\text{Was}}$, for all $g, h \in \mathcal{G}$.

Then, equation (3.5) has a unique solution for any $\rho_0 \in \mathcal{H}^+$.

The proof of Theorem 3.5 is via a reasonably standard fixed point argument. The definition of the non-linear operator D must be appropriately extended to all of \mathcal{H} so that the extension inherits a suitable Lipschitz property from the Lipschitz property of S , and it must be shown that solutions produced by the fixed point argument, which *a priori* take values in \mathcal{H} , actually take values in \mathcal{H}^+ .

Note that the demographic selective cost of Example 2.1 satisfies the Lipschitz condition of the theorem under mild conditions on the function θ .

The following result, which can be proved using the arguments in Section 2 of [CE09], shows that if ρ_0 is absolutely continuous with respect to ν with bounded Radon-Nikodym derivative, then the integral equation (3.5) may be thought of as a (possibly uncountable) system of one-dimensional ordinary differential equations.

Corollary 3.6. Suppose in addition to the assumptions of Theorem 3.5 that ρ_0 is absolutely continuous with respect to ν with a bounded Radon-Nikodym derivative. Then, ρ_t is absolutely continuous with respect to ν for all $t \geq 0$ and there is a non-negative Borel function $(t, m) \mapsto x_t(m)$ on $\mathbb{R}_+ \times \mathcal{M}$ such that the function $m \mapsto x_t(m)$ is a Radon-Nikodym derivative of ρ_t with respect to ν

for all $t \geq 0$, and for ν -a.e. $m \in \mathcal{M}$ the function $t \mapsto x_t(m)$ is differentiable with

$$\dot{x}_t(m) = 1 - F_{\rho_t}(m)x_t(m).$$

4. Convergence of the Discrete Generation Model

Recall that recombination is defined in terms of a probability measure \mathcal{R} on the space of Borel subsets of \mathcal{M} that describes the distribution of the random set of loci that comes from one of the two parents in a mating. Recall, moreover, that for (2.16) to be a limit of a sequence of discrete generation models it is intuitively necessary for the resultant shuffling of genotypes to be thorough enough to break up the dependence between loci introduced by selection. The following condition is useful in quantifying how successful recombination is at performing this task.

Definition 4.1. Given a (symmetric) recombination measure \mathcal{R} and $\lambda \in \mathcal{H}^+$, we say that the pair (\mathcal{R}, λ) is *shattering* if there is a positive constant α such that for any Borel set $A \subseteq \mathcal{M}$,

$$\lambda(A)^3 \leq 2\alpha \int \lambda(A \cap R)\lambda(A \cap R^c) \mathcal{R}(dR). \tag{4.1}$$

Note that if λ is a probability measure, then the right-hand side of (4.1) is, without the constant α , the probability that two random loci drawn independently according to λ are both in the subset A and receive their contents from different parents.

It can be shown that if the pair $(\mathcal{R}, \mu P)$ is shattering for some probability measure P on \mathcal{G} and there is a constant β such that

$$\int g(A) \mathbf{1}_{\{g(A) \geq 2\}} P(dg) \leq \beta \mu P(A)^2 \tag{4.2}$$

for all Borel sets $A \subseteq \mathcal{M}$, then $\mathfrak{R}^k P$ converges to the distribution of a Poisson random measure with intensity μP as $k \rightarrow \infty$. Such a result may be thought of as a generalization of classical Poisson convergence results such as [LC60].

To get a feeling for Definition 4.1, suppose that \mathcal{M} is equipped with a metric δ , that λ is a probability measure, and for some constant $c > 0$

$$\begin{aligned} p(r) &:= \inf\{\mathcal{R}\{R : m' \in R, m'' \in R^c\} : m', m'' \in \mathcal{M}, \delta(m', m'') \geq r\} \\ &\geq c \sup\{\lambda\{m'' \in \mathcal{M} : \delta(m', m'') \leq r\} : m' \in \mathcal{M}\} =: c\varphi(r). \end{aligned} \tag{4.3}$$

for all $r \in \mathbb{R}_+$; that is, loosely speaking, the probability two loci inherit their contents from different parents dominates a multiple of the λ mass of a ball

with radius the distance between the two loci. By a change of variables,

$$\begin{aligned}
 \int \lambda(A \cap R) \lambda(A \cap R^c) \mathcal{R}(dR) &= \int_A \int_A \mathcal{R}\{R : m' \in R, m'' \in R^c\} \lambda(dm'') \lambda(dm') \\
 &\geq \int_A \int_A p(\delta(m', m'')) \lambda(dm'') \lambda(dm') \\
 &\geq c \int_A \int_{\mathcal{M}} \mathbf{1}\{\delta(m', m'') \leq \varphi^{-1}(\nu(A))\} \varphi(\delta(m', m'')) \lambda(dm'') \lambda(dm') \\
 &= c\lambda(A) \frac{1}{2} \lambda(A)^2,
 \end{aligned}
 \tag{4.4}$$

and so (\mathcal{R}, λ) is shattering with constant $\alpha = c^{-1}$.

Definition 4.2. Given $\pi \in \mathcal{H}^+$, denote by Π_π the probability measure on \mathcal{G} that is the distribution of a Poisson random measure with intensity measure π . That is, Π_π is the distribution of the random measure X^π .

Theorem 4.3. Let $(\rho_t)_{t \geq 0}$ be the measure-valued dynamical system of (3.5) whose existence is guaranteed by Theorem 3.5. Suppose in addition to the hypotheses of Theorem 3.5 that the selective cost S is bounded, the pair (\mathcal{R}, ν) (respectively, (\mathcal{R}, ρ_0)) consisting of the recombination measure and the mutation intensity measure (respectively, the recombination measure and the initial intensity measure) is shattering, and the initial measure P_0 is Poisson (with intensity ρ_0). Then, for each $T > 0$,

$$\lim_{n \rightarrow \infty} \sup_{0 \leq t \leq T} \left\| \Pi_{\rho_t} - (\mathfrak{R}\mathfrak{M}_n\mathfrak{S}_n)^{\lfloor nt \rfloor} P_0 \right\|_{\text{Was}} = 0.$$

The proof of Theorem 4.3 is quite long and complex. The first step involves establishing the following analogous result in which the recombination operator \mathfrak{R} that “partially Poissonizes” a probability measure on \mathcal{G} is replaced by the “complete Poissonization operator” \mathfrak{P} that transforms a probability measure P on \mathcal{G} into $\mathfrak{P}P := \Pi_{\mu P}$. That is, $\mathfrak{P}P$ is the distribution of a Poisson random measure with the same intensity measure μP as P .

Proposition 4.4. Suppose that the hypotheses of Theorem 4.3 hold. Then, for each $T > 0$,

$$\lim_{n \rightarrow \infty} \sup_{0 \leq t \leq T} \left\| \Pi_{\rho_t} - (\mathfrak{P}\mathfrak{M}_n\mathfrak{S}_n)^{\lfloor nt \rfloor} P_0 \right\|_{\text{Was}} = 0.$$

Because Proposition 4.4 involves a comparison of two Poisson distributions, it suffices to consider the associated intensity measures and establish for each $T > 0$ that

$$\lim_{n \rightarrow \infty} \sup_{0 \leq t \leq T} \left\| \rho_t - \mu(\mathfrak{P}\mathfrak{M}_n\mathfrak{S}_n)^{\lfloor nt \rfloor} P_0 \right\|_{\text{Was}} = 0.
 \tag{4.5}$$

However, it can be shown using (2.12) and (2.11) that there are constants a, b, c which do not depend on n or T such that

$$\begin{aligned} \|\rho_{(m+1)/n} - \mu(\mathfrak{P}\mathfrak{M}_n\mathfrak{S}_n)^{m+1}P_0\|_{\text{Was}} &\leq \frac{a}{n}\|\rho_{m/n} - \mu(\mathfrak{P}\mathfrak{M}_n\mathfrak{S}_n)^mP_0\|_{\text{Was}} \\ &\quad + \frac{bT + c}{n^2} \end{aligned} \tag{4.6}$$

for $0 \leq m \leq Tn$. We note that (4.5) may be thought of as a *shadowing theorem* about the convergence of a discrete-time dynamical system to an ODE, but standard results in that area (see, for example, [CKP95]) do not seem to apply.

The most difficult part of the proof involves estimating the Wasserstein distance between $(\mathfrak{R}\mathfrak{M}_n\mathfrak{S}_n)^mP_0$ and $(\mathfrak{P}\mathfrak{M}_n\mathfrak{S}_n)^mP_0$. Both probability measures are absolutely continuous with respect to $\mathfrak{M}_n^mP_0$, and so it suffices to estimate the $L^1(\mathfrak{M}_n^mP_0)$ distance between their Radon-Nikodym derivatives. The key idea in accomplishing this is to replace the original genotype space \mathcal{G} by the richer space

$$\mathcal{G}^* := \mathcal{G} \sqcup \mathcal{G}^2 \sqcup \dots, \tag{4.7}$$

where \sqcup denotes disjoint union. An element of \mathcal{G}^i records the mutations from the ancestral wild type that appear in each of i consecutive generations and a probability measure Q on \mathcal{G}^* may be thought of as the distribution of a finite sequence (Y_0, \dots, Y_I) of random measures. Each of the operators $\mathfrak{P}, \mathfrak{R}, \mathfrak{M}_n, \mathfrak{S}_n$ lift in a natural way to this richer setting, and the labeling of mutations by generations makes it easier to keep track of how successive applications of the analogues of $\mathfrak{R}\mathfrak{M}_n\mathfrak{S}_n$ and $\mathfrak{P}\mathfrak{M}_n\mathfrak{S}_n$ alter the composition of the population.

5. Equilibria in General

We assume throughout this section that the assumptions of Theorem 3.5 hold.

Definition 5.1. An *equilibrium* for the dynamical system (3.5) is a measure $\rho_* \in \mathcal{H}^+$ such that $\nu = F_{\rho_*} \cdot \rho_*$. That is, ρ_* is absolutely continuous with respect to ν with Radon-Nikodym derivative satisfying

$$F_{\rho_*} \frac{d\rho_*}{d\nu} = 1 \quad \nu\text{-a.e.} \tag{5.1}$$

Of course, if ρ_* is an equilibrium for (3.5) and $\rho_0 = \rho_*$, then $\rho_t = \rho_*$ for all $t > 0$.

The zero measure is clearly an equilibrium for (3.5) when $\nu = 0$. Note also that $F_\pi(m)$ is $S(\delta_m)$ for any $\pi \in \mathcal{H}^+$ when S is *additive* (that is, $S(g + h) = S(g) + S(h)$ for all $g, h \in \mathcal{G}$), and so $\rho_*(dm) := S(\delta_m)^{-1}\rho(dm)$ is an equilibrium for such a selective cost provided the measure ρ_* belongs to \mathcal{H}^+ ; that is, provided $\int_{\mathcal{M}} S(\delta_m)^{-1}\rho(dm) < \infty$. For any selective cost, $F_\pi(m)$

is approximately $S(\delta_m)$ when the total mass of π is small, and hence it is reasonable that (3.5) should have an equilibrium when the total mass of ν is sufficiently small.

In order to state such a result, we define a family of dynamical systems indexed by $u \in \mathbb{R}_+$ by

$$\rho_t^{(u)} = \rho_0^{(u)} + tu\nu - \int_0^t D\rho_s^{(u)} ds. \tag{5.2}$$

That is, we replace the mutation measure ν in equation (3.5) by the multiple $u\nu$.

Theorem 5.2. *Suppose the selective cost of a nonzero genotype is bounded below. That is,*

$$\inf\{S(\delta_m) : m \in \mathcal{M}\} = \inf\{S(g) : g \in \mathcal{G}, g \neq 0\} > 0.$$

Then, there exists $U > 0$ such that there is an equilibrium for the equation (5.2) for all $u \in [0, U]$. That is, there exist measures $\rho_^{(u)} \in \mathcal{H}^+$, $0 \leq u \leq U$, such that*

$$F_{\rho^{(u)}} \cdot \rho^{(u)} = u\nu.$$

The crux of the proof is to observe that if the measures $\rho_*^{(u)} \in \mathcal{H}^+$, $0 \leq u \leq U$, exist with corresponding Radon-Nikodym derivatives $p^{(u)}$ and ν , then the equilibrium condition is $F_{p^{(u)}}(m)p^{(u)}(m) = u$, where we adopt the convention $F_{p^{(u)}} := F_{\rho^{(u)}}$, and if we differentiate both sides of this equality with respect to u we get the relation

$$\left[\int_{\mathcal{M}} K_{p^{(u)}}(m', m'') \frac{dp^{(u)}}{du}(m'') \nu(dm'') \right] p^{(u)}(m') + F_{p^{(u)}}(m') \frac{dp^{(u)}}{du}(m') = 1, \tag{5.3}$$

where

$$K_\pi(m', m'') := \mathbb{E} \left[S(X^\pi + \delta_{m'} + \delta_{m''}) - S(X^\pi + \delta_{m''}) - S(X^\pi + \delta_{m'}) + S(X^\pi) \right] \tag{5.4}$$

for $\pi \in \mathcal{H}^+$ and $m', m'' \in \mathcal{M}$. It therefore suffices to check that the ODE (5.3) with the boundary condition $p^{(0)} = 0$ has a solution for $u \in [0, U]$.

Theorem 5.2 gives one approach to producing equilibria. Another, more obvious, approach is to start the dynamical system (3.5) with some initial conditions ρ_0 and hope that ρ_t converges to a limit $\rho_* \in \mathcal{H}^+$, in which case ρ_* is an equilibrium – of course, the existence of such limits is the primary reason for being interested in equilibria. If we can show that $t \mapsto \rho_t$ is non-decreasing in the usual partial order on \mathcal{H}^+ for a particular value of ρ_0 , then a limit exists provided that $\sup_{t \geq 0} \rho_t(\mathcal{M}) < \infty$. The following result establishes such a monotonicity property for concave selective costs such as the demographic selective cost of Example 2.1.

Theorem 5.3. *Suppose that the selective cost S is concave. If $\dot{\rho}_0 \geq 0$ (respectively, ≤ 0), then the solution of equation (3.5) guaranteed by Theorem 3.5 satisfies $\rho_s \leq \rho_t$ (resp. $\rho_s \geq \rho_t$) for all $0 \leq s \leq t < \infty$.*

It follows from Theorem 5.3 that if S is concave and $\rho_0 = 0$, so that $\dot{\rho}_0 = \nu \geq 0$, then $t \mapsto \rho_t$ is non-decreasing. Therefore, in this case either $\lim_{t \rightarrow \infty} \rho_t(\mathcal{M}) = \infty$ or $\lim_{t \rightarrow \infty} \rho_t = \rho_* \in \mathcal{H}^+$ exists and is an equilibrium. The following comparison result shows that the latter occurs if and only if there is some equilibrium ρ_{**} , in which case $\rho_* \leq \rho_{**}$. In particular, if there are any equilibria in the concave case, then there is a well-defined *minimal equilibrium*.

Theorem 5.4. *Consider two selective cost functions S' and S'' that satisfy the conditions of Theorem 5.3. Let ρ' and ρ'' be the corresponding solutions of (3.5). Suppose that $S'(g + \delta_m) - S'(g) \geq S''(g + \delta_m) - S''(g)$ for all $g \in \mathcal{G}$ and $m \in \mathcal{M}$ and that $\rho'_0 \leq \rho''_0$. Then, $\rho'_t \leq \rho''_t$ for all $t \geq 0$.*

It can be shown for a concave selective cost that the equilibria $\rho_*^{(u)} \in \mathcal{H}^+$, $0 \leq u \leq U$, for the dynamical systems (5.2) produced by the ODE technique in the proof Theorem 5.2 are minimal.

We note that Theorem 5.3 and Theorem 5.4 are very useful for establishing conditions under which equilibria are *stable* and *attractive* in appropriate senses. We refer the reader to [ESW06] for several results in this direction.

6. Equilibria for Demographic Selective Costs

As we remarked in Example 2.1, the demographic selective cost is concave, and so the general results of Section 5 for concave costs apply. However, we can obtain more refined results by finding another selective cost for which we can compute an equilibrium explicitly and then using Theorem 5.4 to compare the two dynamical systems.

Consider the selective cost

$$S(g) = 1 - \exp\left(-\int_{\mathcal{M}} \sigma(m) g(dm)\right) \tag{6.1}$$

for some $\sigma : \mathcal{M} \rightarrow \mathbb{R}_+$. For this cost,

$$F_\pi(m') = (1 - \exp(-\sigma(m')))\exp\left(-\int 1 - \exp(-\sigma(m'')) \pi(dm'')\right), \tag{6.2}$$

and hence a measure $\rho_* \in \mathcal{H}^+$ is an equilibrium if and only if

$$\frac{d\rho_*}{d\nu}(m') = \frac{\exp\left(\int 1 - \exp(-\sigma(m'')) \nu(dm'')\right)}{1 - \exp(-\sigma(m'))}. \tag{6.3}$$

Therefore, an equilibrium exists if and only if

$$\int \frac{1}{1 - \exp(-\sigma(m))} \nu(dm) < \infty \tag{6.4}$$

and there is a constant $c > 0$ such that

$$c = \int \frac{\exp(c)}{1 - \exp(-\sigma(m))} \nu(dm) - \int \exp(-\sigma(m)) \frac{\exp(c)}{1 - \exp(-\sigma(m))} \nu(dm) \tag{6.5}$$

$$= \exp(c) \nu(\mathcal{M}),$$

in which case the equilibrium ρ_* is given by

$$\frac{d\rho_*}{d\nu}(m') = \frac{\exp(c)}{1 - \exp(-\sigma(m))} = \frac{c}{\nu(\mathcal{M})(1 - \exp(-\sigma(m)))}. \tag{6.6}$$

Such a constant exists if and only if

$$\nu(\mathcal{M}) \leq \sup_{x \geq 0} x \exp(-x) = e^{-1}. \tag{6.7}$$

Now, for the demographic selective cost of Example 2.1,

$$S(g + \delta_{m'}) - S(g) = \int_0^\infty (1 - e^{-\theta(m',x)}) f(x) \times \exp\left(-\lambda x - \int_{\mathcal{M}} \theta(m'',x) g(dm'')\right) dx. \tag{6.8}$$

Suppose that $\sup_{m,x} \theta(m,x) < \infty$ and $\inf_m \inf_{x \in B} \theta(m,x) > 0$ for some set $B \subset \mathbb{R}_+$ such that $\int_B f_x dx > 0$. Then, for some constant $\xi > 0$ and function $\tau : \mathcal{M} \rightarrow \mathbb{R}_+$ such that

$$\int_{\mathcal{M}} \frac{1}{1 - \exp(-\tau(m))} \nu(dm) < \infty, \tag{6.9}$$

we have

$$S(g + \delta_{m'}) - S(g) \geq \xi [1 - \exp(-\tau(m'))] \exp\left(-\int_{\mathcal{M}} \tau(m'') g(dm'')\right) \tag{6.10}$$

for all $m' \in \mathcal{M}$. It follows from our observations above, Theorem 5.3 and Theorem 5.4 that an equilibrium exists provided $\nu(\mathcal{M}) \leq e^{-1}\xi$.

Conversely, if we simply assume that $\sup_{m,x} \theta(m,x) < \infty$, then there are constants v and ζ

$$S(g + \delta_{m'}) - S(g) \leq \zeta [1 - \exp(-v(m'))] \exp\left(-\int_{\mathcal{M}} v(m'') g(dm'')\right). \tag{6.11}$$

Hence, there is no equilibrium when either

$$\int_{\mathcal{M}} \frac{1}{1 - \exp(-v(m))} \nu(dm) = \infty \tag{6.12}$$

or $\nu(\mathcal{M}) > e^{-1}\zeta$.

7. Step Profiles and Demographic Selective Costs

Recall the demographic selective cost function of Example 2.1. Take $\mathcal{M} = \mathbb{R}_+$ and $\theta(m, x) = \eta(m)\mathbf{1}\{x \geq m\}$; that is, we imagine that each mutation imposes a single hazard increment at some age of onset and we “index” mutations by their age of onset. It follows from (3.8) that

$$F_\pi(m') = \left(1 - e^{-\eta(m')}\right) \int_{m'}^{\infty} f(x) \times \exp\left(-\lambda x - \int_0^x \left(1 - e^{-\eta(m'')}\right) \pi(dm'')\right) dx. \quad (7.1)$$

Suppose that the mutation intensity measure ν has a density q against Lebesgue measure and that ρ is an equilibrium intensity measure with density r against Lebesgue measure.

We can observe the distribution of lifetimes in an equilibrium population and this distribution is determined uniquely by its hazard

$$\begin{aligned} h(x) &= \lim_{\epsilon \downarrow 0} \mathbb{P}\{\ell_x(X^\rho \in (x, x + \epsilon) \mid X^\rho > x\} / \epsilon \\ &= -\frac{d}{dx} \log \mathbb{E}[\ell_x(X^\rho)] \\ &= \frac{d}{dx} \left(\lambda x + \int_0^x \left(1 - e^{-\eta(m)}\right) r(m) dm \right) = \lambda + \left(1 - e^{-\eta(x)}\right) r(x). \end{aligned} \quad (7.2)$$

The equilibrium condition for ρ is thus

$$\begin{aligned} q(x) &= \left[\left(1 - e^{-\eta(x)}\right) \int_x^{\infty} f(y) \exp\left(-\int_0^y h(z) dz\right) dy \right] r(x) \\ &= (h(x) - \lambda) \int_x^{\infty} f(y) \exp\left(-\int_0^y h(z) dz\right) dy. \end{aligned} \quad (7.3)$$

A remarkable conclusion from (7.3) is that the equilibrium hazard h is determined by the background hazard λ and the mutation intensity q . Varying the quantity $\eta(m)$ that gives the hazard increment introduced by a mutation with effect having age of onset m does not alter the equilibrium hazard. Rather, the density r changes so that $h(m) = \lambda + (1 - e^{-\eta(m)})r(m)$ stays constant. This outcome is reminiscent of *Haldane's principle* [Hal37], which says for models that may be thought of as linear approximations of the one considered here that the effect size $\eta(m)$ and the equilibrium density $r(m)$ should be such that the product $\eta(m)r(m)$ is constant. The fact that $(1 - e^{-\eta(m)})r(m) \approx \eta(m)r(m)$ when $\eta(m)$ is close to zero reflects that fact that the model considered here is, in some sense, close to its linear approximation in this regime.

It follows from (7.3) that

$$h(x) = -\frac{d}{dx} \log \left(-\frac{1}{f(x)} \frac{d}{dx} \frac{q(x)}{h(x) - \lambda} \right), \tag{7.4}$$

and hence the hazard h is the solution of a non-linear second-order ODE.

It is more convenient to study the function

$$T(x) := \int_x^\infty f(y) \exp \left(-\int_0^y h(z) dz \right) dy \tag{7.5}$$

from which h can be recovered. If we set

$$L := -T' \circ T^{-1}, \tag{7.6}$$

then

$$L' = -\frac{T'' \circ T^{-1}}{T' \circ T^{-1}} = h \circ T^{-1} - \frac{f' \circ T^{-1}}{f \circ T^{-1}} \tag{7.7}$$

and the equilibrium equation (7.3) becomes

$$L'(\tau) = \frac{q \circ T^{-1}(\tau)}{T \circ T^{-1}(\tau)} - \frac{f' \circ T^{-1}(\tau)}{f \circ T^{-1}(\tau)} + \lambda = \frac{q \circ T^{-1}(\tau)}{\tau} - \frac{f' \circ T^{-1}(\tau)}{f \circ T^{-1}(\tau)} + \lambda. \tag{7.8}$$

Suppose, for example, that q and f are constants, say \bar{q} and \bar{f} , on some finite interval $[\alpha, \beta]$ and zero elsewhere. We find that

$$L(\tau) = \lambda\tau + \bar{q} \log(\tau) + C \tag{7.9}$$

for some constant C provided $\tau \in [T(\alpha), T(\beta)]$, and so

$$\lambda T(x) + \bar{q} \log(T(x)) + C = L \circ T(x) = -T'(x) = \bar{f} \mathbb{E}[\ell_x(X^\rho)] \tag{7.10}$$

for $x \in [\alpha, \beta]$. Now $T(x) \rightarrow 0$ as $x \rightarrow \beta$, and so the left-hand side of (7.10) is negative for x sufficiently close to β , whereas the right-hand side is always non-negative. It follows that the equilibrium ρ does not exist in this case. This conclusion is also obvious from the observation that

$$\begin{aligned} S(g + \delta_{m'}) - S(g) &= \left(1 - e^{-\eta(m')} \right) \int_{m'}^\beta \bar{f} \exp \left(-\lambda x - \int_x^\beta \eta(m'') g(dm'') \right) dx \\ &\leq (\beta - m') \bar{f}, \end{aligned} \tag{7.11}$$

and so $F_\rho(m) \leq (\beta - m) \bar{f}$. However, the equilibrium condition $\bar{q} = F_\rho(m)r(m)$ for $m \in [\alpha, \beta]$ shows that $\int_\alpha^\beta r(m) dm = \infty$, and so ρ does not exist.

It is interesting to note that a linearization of this model was proposed in [Cha94] as an explanation of the phenomenon observed for many species that

the age-specific mortality has the celebrated Gompertz-Makeham form, which is equivalent to the equilibrium hazard h being of the form $h(x) = \lambda + \exp(a + bx)$ for suitable constants a and b . It thus appears that the simplification introduced by a linear approximation leads to misleading conclusions.

The arguments above for the non-existence of equilibria are extended to considerably more general selection costs and patterns of fertility and mutation in [WES08].

8. Polynomial Selective Costs

Suppose in this section that the selective cost S is *polynomial*, in the sense that

$$S(g) = \sum_{n=1}^N \int_{\mathcal{M}^n} a_n(\mathbf{m}) g^{\otimes n}(d\mathbf{m}) \tag{8.1}$$

for some positive integer N , where for each n the Borel function $a_n : \mathcal{M}^n \rightarrow \mathbb{R}_+$ is permutation-invariant (that is, $a_n(\pi\mathbf{m}) = a_n(\mathbf{m})$ for all permutations π) and has the property that $a_n(\mathbf{m}) = 0$ if there exist $i \neq j$ with $m_i = m_j$. Furthermore, assume that each function a_n is bounded. The number $n! a_n(m_1, \dots, m_n)$ represents the interactive effect of the n different mutations m_1, \dots, m_n over and above that of any subset of them, and this additional effect is independent of the order in which the mutations are written. Note that this selective cost is not concave unless $a_n = 0$ for $n \geq 2$.

Theorem 8.1. *Suppose that $\inf\{a_1(m) : m \in \mathcal{M}\} > 0$. Then the system (3.5) has a unique equilibrium $\rho_* \in \mathcal{H}^+$.*

This result is established in [CE09]. We sketch the proof and refer the reader to [CE09] for the missing details.

Note that

$$F_\pi(m) = \mathbb{E}[S(X^\pi + \delta_m) - S(X^\pi)] = \sum_{n=1}^N n \int_{\mathcal{M}^{(n-1)}} a_n(\mathbf{m}, m) \pi^{\otimes(n-1)}(d\mathbf{m}) \tag{8.2}$$

for $\pi \in \mathcal{H}^+$ by standard moment computations for Poisson random measures (see, for example, [DVJ88, Kal02]).

It follows from the the equilibrium condition (5.1) that if an equilibrium ρ_* exists, then it has Radon-Nikodym derivative x_* with respect ν that satisfies $\Gamma x_* = x_*$, where

$$\Gamma y(m) := \left[\sum_{n=1}^N n \int_{\mathcal{M}^{(n-1)}} a_n(\mathbf{m}, m) y(m_1) \cdots y(m_{n-1}) \nu^{\otimes(n-1)}(d\mathbf{m}) \right]^{-1} \tag{8.3}$$

for a non-negative Borel function $y : \mathcal{M} \rightarrow \mathbb{R}$.

Set $A := (\inf\{a_1(m) : m \in \mathcal{M}\})^{-1}$. Note that if $0 \leq y(m) \leq A$ for ν -a.e. $m \in \mathcal{M}$, then $B \leq \Gamma y(m) \leq A$ for ν -a.e. $m \in \mathcal{M}$, where

$$B := \left[\sum_{n=1}^N n \int_{\mathcal{M}^{(n-1)}} a_n(\mathbf{m}, m) A^{(n-1)} \nu^{\otimes(n-1)}(d\mathbf{m}) \right]^{-1}. \tag{8.4}$$

Therefore, Γ maps the convex set

$$R := \{y \in L_+^\infty(\mathcal{M}, \nu) : B \leq y(m) \leq A\} \tag{8.5}$$

into itself, and it suffices to show that Γ has a unique fixed point in this set.

It can be shown that if we regard $L^\infty(\mathcal{M}, \nu)$ as the dual of $L^1(\mathcal{M}, \nu)$, then the map Γ is weak*-continuous on the convex weak*-compact set R , and so the existence of at least one fixed point follows from the Schauder-Tychonoff Theorem (that is, the infinite-dimensional analogue of the Brouwer Fixed Point Theorem).

Suppose now that there are two fixed points in R . Then,

$$\begin{aligned} & \sum_{n=1}^N n \int_{\mathcal{M}^{(n-1)}} a_n(\mathbf{m}, m) x(m_1) \cdots x(m_{n-1}) x(m) \nu^{\otimes(n-1)}(d\mathbf{m}) \\ &= \sum_{n=1}^N n \int_{\mathcal{M}^{(n-1)}} a_n(\mathbf{m}, m) y(m_1) \cdots y(m_{n-1}) y(m) \nu^{\otimes(n-1)}(d\mathbf{m}). \end{aligned} \tag{8.6}$$

Therefore,

$$\begin{aligned} & \sum_{n=1}^N n \int_{\mathcal{M}^{(n-1)}} a_n(\mathbf{m}, m) x(m_1) \cdots x(m_{n-1}) x(m) \\ & \times \left(\frac{y(m_1) \cdots y(m_{n-1}) y(m)}{x(m_1) \cdots x(m_{n-1}) x(m)} - 1 \right) \nu^{\otimes(n-1)}(d\mathbf{m}) = 0 \end{aligned} \tag{8.7}$$

for ν -a.e. $m \in \mathcal{M}$. Setting $L_n(\mathbf{m}, m) := a_n(\mathbf{m}, m) x(m_1) \cdots x(m_{n-1}) x(m)$ and $\delta(m) := \log(y(m)/x(m))$, we obtain

$$\sum_{n=1}^N n \int_{\mathcal{M}^{(n-1)}} L_n(\mathbf{m}, m) \left(e^{\delta(m_1) + \cdots + \delta(m_{n-1}) + \delta(m)} - 1 \right) \nu^{\otimes(n-1)}(d\mathbf{m}) = 0. \tag{8.8}$$

Observe that $\delta(m)$ is bounded since x and y are in R . Thus, putting

$$\eta_n(\mathbf{m}, m) := L_n(\mathbf{m}, m) \frac{e^{\delta(m_1) + \cdots + \delta(m_{n-1}) + \delta(m)} - 1}{\delta(m_1) + \cdots + \delta(m_{n-1}) + \delta(m)}, \tag{8.9}$$

we get

$$\sum_{n=1}^N n \int_{\mathcal{M}^{(n-1)}} \eta_n(\mathbf{m}, m) (\delta(m_1) + \cdots + \delta(m_{n-1}) + \delta(m)) \nu^{\otimes(n-1)}(d\mathbf{m}) = 0 \tag{8.10}$$

for ν -a.e. $m \in \mathcal{M}$. Note that the function η_n is non-negative, since $\delta(m_1) + \dots + \delta(m_{n-1}) + \delta(m)$ and $e^{\delta(m_1) + \dots + \delta(m_{n-1}) + \delta(m)} - 1$ have the same sign. Also, η_n is permutation invariant and takes the value 0 whenever two of the coordinates of \mathbf{m} are equal.

Integrating the left-hand side of (8.10) against the function δ gives

$$\begin{aligned}
 0 &= \int_{\mathcal{M}} \sum_{n=1}^N n \int_{\mathcal{M}^{(n-1)}} \eta_n(\mathbf{m}, m) \\
 &\quad \times (\delta(m_1) + \dots + \delta(m_{n-1}) + \delta(m)) \nu^{\otimes(n-1)}(d\mathbf{m}) \delta(m) \nu(dm) \\
 &= \sum_{n=1}^N n \int_{\mathcal{M}^n} \eta_n(m_1, \dots, m_n) (\delta(m_1) + \dots + \delta(m_{n-1}) + \delta(m_n)) \\
 &\quad \times \delta(m_n) \nu^{\otimes n}(dm_1, \dots, dm_n) \tag{8.11} \\
 &= \sum_{n=1}^N \sum_{k=1}^n \int_{\mathcal{M}^n} \eta_n(m_1, \dots, m_n) (\delta(m_1) + \dots + \delta(m_n)) \\
 &\quad \times \delta(m_k) \nu^{\otimes n}(dm_1, \dots, dm_n),
 \end{aligned}$$

by the symmetry of η_n .

Therefore,

$$0 = \sum_{n=1}^N \int_{\mathcal{M}^n} \eta_n(m_1, \dots, m_n) [\delta(m_1) + \dots + \delta(m_n)]^2 \nu^{\otimes n}(dm_1, \dots, dm_n). \tag{8.12}$$

In particular,

$$0 = \int_{\mathcal{M}} \eta_1(m) \delta(m)^2 \nu(dm) \tag{8.13}$$

Since $\inf\{a_1(m) : m \in \mathcal{M}\} > 0$, it follows that $\eta_1(m) > 0$ for all $m \in \mathcal{M}$, and hence $\delta(m) = 0$ for ν -a.e. $m \in \mathcal{M}$, contradicting the assumption that $x \neq y$.

The preceding argument for uniqueness of the fixed point follows that used in Theorem 3.1 of [CF05] to establish a criterion for the uniqueness of equilibria in the finite dimensional *mass-action kinetics* systems of differential equations that arise in the study of *continuous flow stirred tank reactors*.

It is shown in [CE09] that the function $V : L_+^\infty(\mathcal{M}, \nu) \rightarrow \mathbb{R} \cup \{+\infty\}$ defined by

$$\begin{aligned}
 V(x) &:= - \int_{\mathcal{M}} \log(x(m)) \nu(dm) \\
 &\quad + \sum_{n=1}^N \int_{\mathcal{M}^n} a_n(\mathbf{m}) x(m_1) \dots x(m_n) \nu^{\otimes n}(d\mathbf{m})
 \end{aligned} \tag{8.14}$$

is bounded below and V is a *Lyapunov function* in the sense that if $(x_t)_{t \geq 0} = (\frac{d\rho_t}{d\nu})_{t \geq 0}$ with $x_0 \in L_+^\infty(\mathcal{M}, \nu)$ is the system of Radon-Nikodym derivatives

guaranteed by Corollary 3.6, then

$$\frac{d}{dt}V(x(t)) = - \int_{\mathcal{M}} (1 - F_{\rho_t}(m)x_t(m))^2 \frac{1}{x_t(m)} \nu(dm) \leq 0. \quad (8.15)$$

It follows by fairly standard arguments that x_t converges as $t \rightarrow \infty$ to $x_* := \frac{d\rho_*}{d\nu}$ in $L^\infty(\mathcal{M}, \nu)$ for any value of $x_0 = \frac{d\rho_0}{d\nu}$.

References

- [AGS05] Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré, *Gradient flows in metric spaces and in the space of probability measures*, Lectures in Mathematics ETH Zürich, Birkhäuser Verlag, Basel, 2005.
- [BT91] N. H. Barton and M. Turelli, *Natural and sexual selection on many loci*, *Genetics* **127** (1991), 229–55.
- [Bür00] Bürger, *The mathematical theory of selection, recombination, and mutation*, John Wiley & Sons, Chichester, New York, 2000.
- [CE09] Aubrey Clayton and Steven N. Evans, *Mutation-selection balance with recombination: convergence to equilibrium for polynomial selection costs*, *SIAM J. Appl. Math.* **69** (2009), no. 6, 1772–1792.
- [CF05] Gheorghe Craciun and Martin Feinberg, *Multiple equilibria in complex chemical reaction networks. I. The injectivity property*, *SIAM J. Appl. Math.* **65** (2005), no. 5, 1526–1546.
- [Cha94] Brian Charlesworth, *Evolution in age-structured populations*, Cambridge University Press, Cambridge, 1994.
- [Cha01] ———, *Patterns of age-specific means and genetic variances of mortality rates predicted by the mutation-accumulation theory of ageing*, *J. Theor. Bio.* **210** (2001), no. 1, 47–65.
- [CKP95] Brian A. Coomes, Hüseyin Koçak, and Kenneth J. Palmer, *A shadowing theorem for ordinary differential equations*, *Z. Angew. Math. Phys.* **46** (1995), no. 1, 85–106.
- [DU77] J. Diestel and J. J. Uhl, Jr., *Vector measures*, American Mathematical Society, Providence, R.I., 1977, With a foreword by B. J. Pettis, *Mathematical Surveys*, No. 15.
- [DVJ88] D. J. Daley and D. Vere-Jones, *An introduction to the theory of point processes*, Springer Series in Statistics, Springer-Verlag, New York, 1988.
- [EK86] Stewart Ethier and Thomas Kurtz, *Markov processes: Characterization and convergence*, John Wiley & Sons, 1986.
- [ESW06] Steven N. Evans, David Steinsaltz, and Kenneth W. Wachter, *A mutation-selection model for general genotypes with recombination*, Tech. Report 717, Department of Statistics, University of California at Berkeley, 2006, Available at <http://arxiv.org/abs/q-bio/0609046>.

- [Ewe04] Warren J. Ewens, *Mathematical population genetics. I*, second ed., Interdisciplinary Applied Mathematics, vol. 27, Springer-Verlag, New York, 2004.
- [FK00] Caleb E. Finch and Thomas B. L. Kirkwood, *Chance, development, and aging*, Oxford University Press, 2000.
- [Hal37] J. B. S. Haldane, *The effect of variation on fitness*, *American Naturalist* **71** (1937), 337–49.
- [Ham66] W. D. Hamilton, *The moulding of senescence by natural selection*, *J. Theor. Biol.* **12** (1966), 12–45.
- [Hol95] Robin Holliday, *Understanding ageing*, Cambridge University Press, 1995.
- [Kal02] Olav Kallenberg, *Foundations of modern probability*, second ed., Probability and its Applications (New York), Springer-Verlag, New York, 2002.
- [Ken74] D. G. Kendall, *Foundations of a theory of random sets*, Stochastic geometry (a tribute to the memory of Rollo Davidson), Wiley, London, 1974, pp. 322–376.
- [KJB02] Mark Kirkpatrick, Toby Johnson, and Nick Barton, *General models of multilocus evolution*, *Genetics* **161** (2002), 1727–50.
- [LC60] Lucien Le Cam, *An approximation theorem for the Poisson binomial distribution*, *Pacific Journal of Mathematics* **10** (1960), 1181–1197.
- [Med52] Peter Medawar, *An unsolved problem in biology: An inaugural lecture delivered at University College, London, 6 December, 1951*, H. K. Lewis and Co., London, 1952.
- [PC98] Scott D. Pletcher and James W. Curtsinger, *Mortality plateaus and the evolution of senescence: Why are old-age mortality rates so low?*, *Evolution* **52** (1998), no. 2, 454–64.
- [Rac91] Svetlozar T. Rachev, *Probability metrics and the stability of stochastic models*, Wiley Series in Probability and Mathematical Statistics: Applied Probability and Statistics, John Wiley & Sons Ltd., Chichester, 1991.
- [RR98] Svetlozar T. Rachev and Ludger Rüschemdorf, *Mass transportation problems. Vol. I*, Probability and its Applications (New York), Springer-Verlag, New York, 1998.
- [SEW05] David Steinsaltz, Steven N. Evans, and Kenneth W. Wachter, *A generalized model of mutation-selection balance with applications to aging*, *Adv. Appl. Math.* **35** (2005), no. 1, 16–33.
- [Vil03] Cédric Villani, *Topics in optimal transportation*, Graduate Studies in Mathematics, vol. 58, American Mathematical Society, Providence, RI, 2003.
- [Vil09] ———, *Optimal transport*, Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences], vol. 338, Springer-Verlag, Berlin, 2009.
- [WES08] Kenneth W. Wachter, Steven N. Evans, and David R. Steinsaltz, *The age-specific forces of natural selection and walls of death*, Tech. Report 757, Department of Statistics, University of California at Berkeley, 2008, Available at <http://arxiv.org/abs/0807.0483>.

- [Wil57] George C. Williams, *Pleiotropy, natural selection, and the evolution of senescence*, *Evolution* **11** (1957), 398–411.
- [WSE10] Kenneth W. Wachter, David R. Steinsaltz, and Steven N. Evans, *Vital rates from the action of mutation accumulation*, *Journal of Population Ageing* (2010), Currently only published online.

Coevolution in Spatial Habitats

Claudia Neuhauser*

Abstract

Empirical and theoretical studies have implicated habitat coarseness and coevolution as factors in driving the degree of specialization of mutualists and pathogens. We review recent advances in the development of a framework for host-symbiont interactions that considers both local and stochastic interactions in a spatially explicit habitat. These kinds of interactions result in models with large numbers of parameters due to the large number of potential interactions, making complete analysis difficult. Rigorous analysis of special cases is possible. We also point to the importance of combining experimental and theoretical studies to identify relevant parameter combinations.

Mathematics Subject Classification (2010). Primary 60K35; Secondary 82C22.

Keywords. Interacting particle systems, voter model, host-symbiont model, coevolution

1. Introduction

Naturalists in the 18th and first half of the 19th century started to catalogue the bewildering diversity of the natural world according to the system developed by Linnaeus. Darwin's work on pollination (1859, 1862) initiated a new line of research, namely that of species interaction. This quickly led to the realization that not only are there different kinds of interactions, such as predation, parasitism, mutualism, or competition, but also that the degree of specialization varies tremendously.

In parasitic, or pathogenic, interactions, the host is harmed and the parasite, or pathogen, benefits. In mutualistic interactions, both the host and the mutualist benefit. We will refer to mutualistic and pathogenic interactions collectively as symbiotic interactions.

*Partially supported by NSF Grants DMS-00-72262 and DMS-00-83468.

University of Minnesota Rochester, Biomedical Informatics and Computational Biology,
300 University Square, 111 S. Broadway, Rochester, MN 55904, USA.
E-mail: neuha001@umn.edu.

Species interactions affect specialization through coevolutionary processes and are far from static. Not only can the kinds of interactions change along the mutualism-parasitism continuum depending on the ecological context (Thompson 1988; Bronstein 1994; Herre et al. 1999; Hernandez 1998; Johnson et al. 2003), there is evidence that the degree of specialization can both increase and decrease along phylogenetic lineages. It appears though that specialization is the much more common lifestyle. Furthermore, as Thompson (1994, p. 122) pointed out, “[e]xtreme specialization extends to commensals and mutualistic symbionts that live on a single host individual (Thompson 1982), but it is in parasites that the pattern is most evident.”

The degree of specialization is also influenced by habitat heterogeneity and coarseness. Optimal foraging theory and habitat selection theory have given insights into the evolution of specialization (Rosenzweig 1987a). Rosenzweig (1981) and Pimm and Rosenzweig (1981) developed a theoretical framework, known as the isolog theory, that makes predictions about when species should show preferences for specific habitat types or be opportunistic based on other competitors and their own densities.

Brown (1990) expanded the work by Rosenzweig to predict the outcome of competition of specialist and generalist competitors in a heterogeneous environment using evolutionary game theory. He found that depending on the cost of habitat selection and fitness of the competitors, up to two competitors can share a habitat composed of two habitat types. When habitat selection is costly, a single generalist dominates, whereas when habitat selection is free, two specialist species dominate, each specialized on its respective habitat. When habitat availability is asymmetric or the specialist has higher fitness than the generalist, both the specialist and generalist can coexist, with the generalist exploiting the habitat that is underused by the specialist. Cost of habitat selection is closely related to coarseness of habitat, namely habitat selection in a fine-grained habitat tends to be more costly than in a coarse-grained habitat due to an increase in travel time. Since the models are non-spatial, habitat heterogeneity or coarseness is modeled indirectly. Brown (1990) incorporated coarseness of habitat indirectly through varying patch encounter rates. His results demonstrate that the coarseness of the habitat affects the evolutionary trajectory. Namely, selective forces are strongly stabilizing towards a single generalist strategy when the habitat is fine-grained; whereas if the habitat is coarse-grained and thus selection is relatively cost free, selection is disruptive and results in specialist strategies. It must be noted that a species' perception of habitat coarseness depends on its dispersal ability. A species that disperses only over small spatial ranges may perceive a habitat as coarse-grained, whereas a species that disperses over large spatial ranges may perceive the same habitat as fine-grained.

Habitat selection can promote coexistence of competitors (Levin 1974, Yodanis 1978, Hastings 1980), and has been implicated as a factor that increases the probability or rate of allopatric speciation. For instance, Thorpe (1945) based

on Peterson (1932) concluded that microlepidoptera genera that are mono or oligophagous are more species rich than polyphagous ones (see Rosenzweig 1987b). Vrba (1980), studying ungulates in Africa, came to a similar conclusion.

Observations in natural systems combined with spatially implicit, mathematical models thus allow us to conclude that both habitat coarseness and coevolution affect specialization: specialists are more likely to be associated with coarse-grained habitats, and extreme specialization is more likely to be found in parasitic interactions, implying that parasitic interactions are more likely to be found in coarse-grained habitats.

Factors that drive the dynamics in natural systems are difficult to tease apart, and models play an important role in studying the consequences of different factors in isolation. Species interactions are characterized by local interactions and chance encounters. Both factors are missing from the spatially implicit, mathematical models that were introduced by ecologists and evolutionary biologists to advance the theory of consequences of habitat heterogeneity and coevolution. Models have been introduced by mathematical ecologists to include explicit space and stochasticity to study the effects on ecological communities. These models range from minimal assumptions on the type of interactions, such as the neutral model advanced by Hubbell (2001), to large-scale, statistical models that rely on stochastic interactions and spatial heterogeneity, such as the macroecological models by Brown (1995) and Maurer (1999). A comprehensive theory of the consequences of host-symbiont interactions, however, has been hampered by the complexity of models. In particular, deterministic and spatially implicit models quickly lead to large systems of differential equations even if only a moderate number of hosts and symbionts are involved.

To advance the theory of host-symbiont interactions that takes into account both local and stochastic interactions, we pursued two main venues. To investigate the consequences of local and stochastic interactions between hosts and their symbionts, we developed an experimental system supported by a simulation model (Kerr et al. 2006). The experimental system of a bacterial host (*E. coli*) and its viral pathogen (T4) allowed us to study the effect of migration patterns on the evolution of this host-pathogen system. To understand the consequences of multiple hosts and symbionts interacting in a spatial environment, we introduced two mathematical models (Lanchier and Neuhauser 2006a, 2006b, 2010). These mathematical models idealized host-symbiont interactions and were amenable to mathematically rigorous treatment. In the first mathematical model, the *static host model*, the host population is fixed and we investigated how habitat coarseness affects the competitiveness of generalists versus specialists. In the second mathematical model, the *dynamic host model*, the host population changes dynamically, and we studied how feedback between hosts and their symbionts affects habitat coarseness.

In the static host model, the spatial pattern of the hosts is fixed and arranged in a checkerboard pattern. The size of the patches in relation to the dispersal

ranges of the symbionts determines whether the symbionts perceive the habitat as coarse-grained or fine-grained. The results of this investigation confirmed the previous observations that generalist symbionts are more competitive in fine-grained habitats and specialist symbionts are more competitive in coarse-grained habitats. In this model, we cannot distinguish between mutualistic or pathogenic interactions.

In the dynamic host model, we can distinguish between mutualists and pathogens. The dynamic host model allows us to study how the feedback between hosts and symbionts shapes the spatial patterns. While generalist symbionts do not affect qualitatively the spatial patterns of hosts, they tend to change the time scale of pattern formation, with pathogens speeding up spatial aggregation of the hosts. Specialist symbionts can profoundly alter the spatial patterns of hosts. Simulations indicate the pathogens promote coexistence, whereas mutualists lead to a coarse-grained habitat.

We will first describe the mathematical models and present some of the rigorous results from Lanchier and Neuhauser (2006a, 2006b, and 2010) interspersed with conjectures that warrant further investigations. We will conclude with a description of the experimental system to argue about the importance of combining theoretical investigations with experimental work.

2. Mathematical Models

As mentioned in the Introduction, we will describe two closely related, spatially explicit, stochastic models, one in which the host population is static (*static host model*), and the other in which hosts evolve dynamically in response to interactions with their symbionts (*dynamic host model*). The spatial models are continuous time Markov processes that evolve on the d -dimensional integer lattice \mathbf{Z}^d . We denote the configuration at time t by $\{\xi_t : x \in \mathbf{Z}^d\}$ where $\xi_t(x) = (i, j)$, $i = 1, 2, \dots, N_1$ and $j = 0, 1, 2, \dots, N_2$, means that site $x \in \mathbf{Z}^d$ is occupied by a host of type i , which is one of N_1 hosts, and a symbiont of type j , which is one of N_2 symbionts if $j \geq 1$ or not occupied by a symbiont if $j = 0$.

Each site is assigned a host type at time 0, which remains the same for all $t > 0$ in the static host model, but may change in the dynamic host model. The infection dynamics are the same in both models. We define two neighborhood sets, one for the dispersal of the symbiont (\mathcal{N}_2), and the other for the dispersal of the host (\mathcal{N}_1). If for $x \in \mathbf{Z}^d$, we set $\|x\| = \sup_{i=1,2,\dots,d} |x_i|$, then

$$\mathcal{N}_i = \{x \in \mathbf{Z}^d : 0 < \|x\| \leq R_i\}, \quad i = 1, 2$$

A healthy host of type i , $i \in \{1, 2, \dots, N_1\}$ and denoted by $(i, 0)$, at location x becomes infected by a symbiont of type j , $j \in \{1, 2, \dots, N_2\}$ at rate c_{ij} times the number of hosts in the neighborhood $x + \mathcal{N}_2$ that are infected with symbiont j :

$$(i, 0) \rightarrow (i, j) \quad \text{at rate} \quad c_{ij} \sum_{z \in x + \mathcal{N}_2} \sum_{l=1}^{N_1} I\{\xi(z) = (l, j)\}$$

The recovery dynamics of infected hosts depend on the dynamics of the host population. In the static-host model, recovery is spontaneous. That is, for $1 \leq i \leq N_1$ and $1 \leq j \leq N_2$,

$$(i, j) \rightarrow (i, 0) \quad \text{at rate} \quad 1$$

We consider two types of dynamic-host models. One where the symbiotic relationship affects fertility, called the *fertility model*; the other affects viability, called the *viability model*. In the fertility model, recovery is by replacement with healthy offspring of neighboring hosts. The fertility of an infected host may be higher (respectively, lower) than that of an uninfected host, in which case the symbiont is called a mutualist (respectively, pathogen). If $\xi_t(x) = (i, j)$, then

$$(i, j) \rightarrow (k, 0) \quad \text{at rate} \quad \sum_{z \in x + \mathcal{N}_1} \sum_{l=0}^{N_2} \gamma_l I\{\xi(z) = (k, l)\}$$

where we assume $\gamma_0 = 1$. The parameters γ_l , $l = 1, 2, \dots, N_2$, determine whether an interaction is mutualistic ($\gamma_l > 1$) or pathogenic ($0 \leq \gamma_l < 1$).

In the viability model, individuals die at a rate that depends on who they are associated with and are replaced with healthy offspring of neighboring hosts upon death. The death rate of an infected host may be lower (respectively, higher) than that of an uninfected host, in which case the symbiont is called a mutualist (respectively, pathogen). If $\xi_t(x) = (i, j)$, then

$$(i, j) \rightarrow (k, 0) \quad \text{at rate} \quad \delta_j \sum_{z \in x + \mathcal{N}_1} \sum_{l=0}^{N_2} I\{\xi(z) = (k, l)\}$$

where we assume $\delta_0 = 1$. The parameters δ_j , $j = 1, 2, \dots, N_2$, determine whether an interaction is mutualistic ($0 \leq \delta_j < 1$) or pathogenic ($\delta_j > 1$).

3. Results

3.1. Static Host Model. If the dispersal range is much larger than the spatial scale of the host patches, the dynamics of spatially explicit, stochastic models can be well approximated by a system of ordinary differential equations, which are called mean-field models (Durrett and Levin 1994). This corresponds to a fine-grained habitat. If the dispersal range is comparable to the spatial scale of the host patches, the dynamics of the spatially explicit, stochastic model can no longer be approximated by systems of differential equations. Instead, the full stochastic model must be analyzed.

We begin with the static host model and consider the case of two specialist symbionts and one generalist symbiont on two hosts. Since we cannot distinguish between mutualists and pathogens in this model, we refer to them

collectively as consumers. We assume that a fraction p of the habitat is occupied by host 1 and a fraction $1-p$ by host 2. In the fine-grained habitat, p is a parameter; in the coarse-grained habitat, we will consider the case where the habitat is a checkerboard of the two host types (for $p = 1/2$, see Lanchier and Neuhauser, 2006a).

Since there are two host types, 1 and 2, and three symbionts, two of which are specialists and the third one is a generalist, we set $c_{11} = c_{22} = \alpha > 0$, $c_{13} = c_{23} = \beta > 0$, and $c_{12} = c_{21} = 0$.

3.1.1. Mean-field Model. To define the dynamics in the fine-grained habitat we rescale the parameters, namely

$$a = \frac{\alpha}{|\mathcal{N}_2|} \quad \text{and} \quad b = \frac{\beta}{|\mathcal{N}_2|}$$

We denote by ν_{ij} the fraction of host of type i that is occupied by a symbiont of type j ; the fraction of unoccupied hosts of type i is denoted by u_i . The mean-field equations are given by (Lanchier and Neuhauser 2006a)

$$\begin{aligned} \frac{d\nu_{11}}{dt} &= -\nu_{11} + a u_1 \nu_{11} \\ \frac{d\nu_{22}}{dt} &= -\nu_{22} + a u_2 \nu_{22} \\ \frac{d\nu_{13}}{dt} &= -\nu_{13} + b u_1 (\nu_{13} + \nu_{23}) \\ \frac{d\nu_{23}}{dt} &= -\nu_{23} + b u_2 (\nu_{13} + \nu_{23}) \end{aligned}$$

It follows that

$$u_1 = p - \nu_{11} - \nu_{13} \quad \text{and} \quad u_2 = 1 - p - \nu_{22} - \nu_{23}$$

Furthermore, ν_{13} is positive if and only if ν_{23} is positive. We can therefore describe the possible equilibria in terms of presence or absence of the three competitors: generalist (G), specialist 1 (S_1), and specialist 2 (S_2). There are eight qualitatively different equilibria, namely the eight possible combinations of presence and absence of the three species. We will use the short-hand notation (G, S_1, S_2) to describe the equilibria, with “0” denoting the absence of the respective species. We refer to equilibria in which only one species is present as monoculture equilibria. The equilibrium in which all species are absent is the trivial equilibrium, denoted by $(0, 0, 0)$.

The monoculture equilibria are $\nu_{11} = p - 1/a$, $\nu_{22} = 1 - p - 1/a$, and $\nu_{13} + \nu_{23} = 1 - 1/b$, respectively. Hence, a necessary condition for the generalist to persist is $b > 1$. Each of the specialists requires $a > 1$ for survival. In addition, there is a minimum habitat requirement for each of the specialists in the absence of the other species. Namely, for specialist 1 to survive when

specialist 2 and the generalist are absent, the fraction p of habitat 1 must exceed $1/a$. (The behavior of specialist 2 is symmetric, i.e., $1 - p$ must exceed $1/a$.)

It follows that if $b \leq 1$ and $1 < a < 2$, specialist 2 can persist if $0 \leq p < 1 - 1/a$, and specialist 1 can persist if $1/a < p \leq 1$. Since the generalist cannot persist, the trivial equilibrium is the only possible equilibrium when $1 - 1/a < p < 1/a$.

The behavior of this model when $b > 1$ can be summarized as follows: If $a < b$, then $(G, 0, 0)$ is the only locally stable equilibrium. If $a > 2b$, then the specialists outcompete the generalist. For p small (respectively, p large), $(0, 0, S_2)$ (respectively, $(0, S_1, 0)$) is the only locally stable equilibrium. For intermediate values of p , the locally stable equilibrium is a coexistence equilibrium of specialists 1 and 2, $(0, S_1, S_2)$. Note that in the case of monoculture equilibria, the generalist cannot invade since there is not enough space available for it to persist. In this case, only one habitat type is occupied. When $b < a < 2b$, as p decreases from 1 to 0, the locally stable equilibria change from $(0, S_1, 0)$ to $(G, S_1, 0)$, $(G, 0, 0)$, $(G, 0, S_2)$, and $(0, 0, S_2)$. We find that for each parameter combination satisfying $b < a < 2b$, there is only one locally stable equilibrium, and all but the (G, S_1, S_2) equilibria are possible.

3.1.2. Spatially Explicit Model. Lanchier and Neuhauser (2006a) analyzed the static host model when two hosts lived on an alternating pattern of boxes on the d -dimensional integer lattice where each box is a translate of the box $H_L = [-L, L]^d$. To mimic specialist and generalist interactions, we stipulate that boxes centered at $Lx = (Lx_1, Lx_2, \dots, Lx_d)$ with $x_1 + x_2 + \dots + x_d$ is even (respectively, odd) are occupied by hosts of type 1 (respectively, 2). Specialists of type 1 live on hosts of type 1; specialists of type 2 live on hosts of type 2; and generalists of type 3 can live on either host. Recall the R_2 is the radius of the neighborhood \mathcal{N}_2 .

We denote by λ_c the critical value of the contact process. We found that if L and R_2 are fixed, $\beta > \lambda_c$ and $\beta \geq \alpha$, then if initially the generalists have positive density, the generalists win. In a coarse-grained habitat, however, specialists need only be marginally better to outcompete generalists. Assuming that R_2 is fixed, we found that in two or more dimensions, when $\alpha > \lambda_c$ and $\alpha > \beta$, then, provided that initially both 1's and 2's have positive densities, the two specialists can coexist and outcompete the generalist when L is large enough. As a corollary, under the same conditions on the parameters, if initially there is one specialist and one generalist and both have positive densities, the specialist and the generalist coexist since they are able to divide up the habitat in the same way as the two specialists, thus rendering the generalist effectively as a specialist.

3.1.3. Comparison. To compare the behavior of the static-host model in fine-grained and the coarse-grained habitats, we see that although competitive-

ness decreases with specialization, specialists are more competitive in a coarse-grained than in a fine-grained habitat. These results are consistent with the findings of Brown (1990). Furthermore, if we denote by α_c the smallest value of the infection rate of the specialist so that the specialist will outcompete a generalist (of course assuming that the habitat is such that the specialist in the absence of the generalist will survive), then the critical value will approach β in a coarse-grained habitat as the length of the habitat patches increases, but is equal to 2β in the fine-grained habitat.

3.2. Dynamic Host Model. Before we discuss the behavior of the model that includes both hosts and symbionts, let's assume that symbionts are absent. The resulting spatial stochastic model is then known as the voter model (Clifford and Sudbury 1973; Holley and Liggett 1975). The voter model is the simplest of all multi-species models. Its non-spatial version is equivalent to the Wright-Fisher model (Fisher 1930; Wright 1931), which does not allow for coexistence of multiple types if all types have the same dynamics, i.e., the neutral model, and mutations are absent.

In the voter model, each site is always occupied by one of the host types. The name of the model, voter model, comes from interpreting the dynamics as adopting opinions: at rate 1, an individual at site x chooses a neighbor at random and adopts his opinion. (Equivalently, at rate 1, an individual at x chooses a neighbor at random and imposes her opinion on that site.) The long-term behavior of this model exhibits a dichotomy, depending on the spatial dimension. In one or two spatial dimension, the model exhibits clustering, whereas in three or higher dimensions, coexistence is possible. (By clustering we mean that if any two sites are picked, the probability that these sites are occupied by different host types goes to 0 as time tends to infinity. This probability is positive when coexistence occurs.) The reason for this dichotomy in behavior lies in the fact that one and two dimensional, symmetric random walks are recurrent, that is, with probability 1, if starting at 0, the random walk will return to 0, whereas in three and higher dimensions, these random walks are transient and so there is a positive probability the random walk will never return to 0. Random walks enter into this discussion because the ancestral process (called dual process) of a particle in the voter model performs a random walk.

The dynamic host model allows us to distinguish between mutualist and pathogens. The feedback between the symbionts and the hosts has the potential to affect the spatial organization of the host in the presence of mutualists, and our interest will focus on whether symbionts qualitatively change the spatial organization of the host population.

3.2.1. Mean-field Model. In Lanchier and Neuhauser (2006b), the mean field model for the fertility model was analyzed under the assumption that the number of hosts and symbionts are the same and that for all hosts $c_{ii} = \alpha$ and

$c_{ij} = \beta$ for $i \neq j$ and $\alpha > \beta$, i.e., we assume some degree of specialization. Furthermore, γ_j was assumed not to depend on j for $j \geq 1$. A similar analysis can be carried out for the viability model where we assume that δ_j does not depend on j when $j \geq 1$. Numerical simulations indicate that for either the fertility or the viability model coexistence of multiple symbionts only occurs when the symbionts are pathogens.

3.2.2. Spatially Explicit Model. The dynamic-host fertility and viability models were investigated in Lanchier and Neuhauser (2006b, 2010). The qualitative behavior of the two models is the same—they appear like time-changes of each other (although the time change is non-trivial in the sense that there is no simple function that relates the two processes).

Rigorous results are available primarily for the viability model when either the symbionts are generalists or when the symbionts are specialist mutualists. There are some limited rigorous results available for other cases. We begin with stating results for the viability model when the mutualists are generalists. Generalist interactions result when the infection rate c_{ij} does not depend on the host type i . That is, we assume $c_{ij} = \alpha_j$. We assume that δ_j is positive. We find (Lanchier and Neuhauser 2010) that in $d \leq 2$, clustering of hosts occurs, that is, for any initial configuration, the probability that two sites will be occupied by different hosts goes to 0 as time goes to infinity. In $d \geq 3$, coexistence is possible. (This result was proved when dispersal was only among nearest neighbors but the proof can easily be extended to the neighborhoods considered here.) This is the same behavior as in the voter model (Clifford and Sudbury 1973; Holley and Liggett 1975). The reason is the same. In both the voter model and in our more complex host-symbiont model, we can follow the ancestry of each individual backwards in time. In both cases, the paths of ancestry perform random walks. When two different paths of ancestry collide, they coalesce, implying that the two starting sites will be of the same type. Since random walks are recurrent in $d \geq 2$ and transient in $d \geq 3$, the results follow. The random walk of the ancestry of a host individual is quite a bit more complicated in the case of the host-symbiont model. However, in the viability case, the dynamics of the symbiont can be treated separately from the dynamics of the host, in the sense that we can first run the symbiont dynamics forward in time to determine which sites will be infected. Because the death rate of an individual in the viability model only depends on whether it is infected or not but not on its neighborhood, we can then follow the ancestry of an individual host backwards in time on this graph where the times of infection are noted. This argument breaks down in the fertility model where an individual host dies if it is replaced by a birth of a neighboring host. A further analysis of the viability model on a complete graph indicates that ancestral paths coalesce faster when the symbiont is a pathogen than when it is a mutualist. This is confirmed in simulations in $d = 2$ where pathogens cause the hosts to cluster more quickly than mutualists.

Rigorous results for specialist interactions are available for the viability model when the specialist is a mutualist and nearest-neighbor dispersal is assumed. In this case, we were able to show that if host 1 is infected with a specialist mutualist 1 and host 2 is never infected (i.e., $c_{11} > 0$, $c_{12} = 0$ and no other symbionts are present), then starting from Bernoulli measure with a positive density of associated hosts of type 1, eventually all sites will be of host type 1 and the mutualist 1 will have positive density provided c_{11} is large enough. This result can be extended to multiple symbionts and we were able to show that if there is a preferred specialist mutualist, say mutualist 1 (i.e., $\delta_1 < \min\{\delta_0, \delta_2, \dots, \delta_{N_2}\}$), then starting from a Bernoulli measure with a positive density of infected hosts of type 1, eventually all sites will be occupied by host 1 and only mutualist 1 will survive provided its infection rate is large enough.

When the symbiont is a pathogen, we only have rigorous results in the one-dimensional, nearest neighbor case, which we conjecture behaves differently from more general neighborhoods or higher dimensions. In the general case, we conjecture that a specialist pathogen cannot survive and the system will eventually reduce to a host-only model in which hosts behave like a voter model.

3.2.3. Comparison. Simulations show that the feedback between hosts and their symbionts in the dynamic model can significantly alter the spatial patterns when the symbionts are specialists. When the interaction is mutualistic, the spatially explicit and stochastic dynamic-host model shows clustering that is very similar to the behavior of the voter model, whereas when the interaction is pathogenic, coexistence is possible (just as in the mean-field model).

The models we introduced above can mimic a wide variety of interactions. We have studied symmetric interactions. Durrett and Lanchier (2008) investigated the dynamic host model when the birth rates of unassociated hosts differ. They studied the system under long-range dispersal with two hosts, one of which may be associated with a symbiont. They identified cases when a host with a specialist pathogen can coexist with a second species, and conjectured that coexistence of two pathogens is possible but coexistence of a specialist mutualist with a second species is not.

4. Experiments

Mathematical models allow investigations of a wide range of parameters. While rigorous analysis may be difficult, simulations can at least yield conjectures that may lead to a fairly complete picture of the behavior of the model under investigation. However, such analysis does not provide insights into which combinations of parameters are realized in nature and are thus relevant for ecological studies. To link mathematical models to natural systems is often quite difficult due to the inherent complexity. Natural systems are complex webs of

interaction, few of which are well enough understood to relate a subset of interactions to a mathematical model. An alternative is provided by experimental laboratory systems that are composed of relatively simple communities and where the environment and species interactions can be tightly controlled.

Kerr et al. 2006 developed one such system that pointed to the importance of including evolutionary aspects into ecological models. We demonstrated experimentally that the pattern of migration (local versus global) affected the evolution of a viral pathogen (T4) that infected a bacterial host (*E. coli*). The environment mimicked a two-dimensional integer lattice and consisted of two microtitre plates, each with 96 wells that were filled with a nutrient solution in which the bacterial host could live. A high-throughput liquid-handling robot was programmed to execute a migration scheme where content from one well was transferred to a different well. In addition to the bacterial host, we introduced a pathogen that was able to live on the host. Different migration patterns resulted in different outcomes. Specifically, we found that when migration was only among neighboring wells, so-called “prudent” pathogens were selected, whereas when migration occurred randomly across the entire microtitre habitat, so-called “rapacious” pathogens were selected. The two pathogens differed in their competitiveness and productivity, namely the prudent pathogen was less competitive but more productive than the rapacious pathogen. The different migration schemes revealed a trade-off that restricts the parameter space to a feasible subset.

The experiment was accompanied by a stochastic simulation model that was parameterized by the experiment and mimicked the essential features of the experiment. The first model only included a single pathogen for both migration schemes and resulted in predictions that were not consistent with the experiment. A careful study of the host and the pathogen revealed that different strains of the virus evolved under the two different migration schemes. Once the different types were incorporated in the model, the model predictions agreed with the experimental outcome.

Such insights can only be gained from experiments and point towards the importance of mathematicians collaborating with biologists if mathematical models aim to have an impact on gaining an increased understanding of biological systems.

References

- [1] Bronstein, J.L. 1994. Conditional outcomes in mutualistic interactions. *Trends Ecol. Evol.* 9: 214–217.
- [2] Brown, J.S. 1990. Habitat selection as an evolutionary game. *Evolution* 44: 732–746.
- [3] Brown, J.H. 1995. *Macroecology*. University of Chicago Press, Chicago.
- [4] Clifford and Sudbury. 1973. A model for spatial conflict. *Biometrika* 60: 581–588.

- [5] Darwin, C. 1859. On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. Reprinted in Penguin Classics 1985.
- [6] Darwin, C. 1862. On the various contrivances by which British and foreign orchids are fertilized by insects, and on the good effects of intercrossing. Facsimile ed. Earl M. Coleman, Standfordville, N.Y.
- [7] Durrett, R. and N. Lanchier. 2008. Coexistence in host-pathogen systems. *Stoch. Proc. Appl.* 118: 1004–1021.
- [8] Durrett, R. and S.A. Levin. 1994. The importance of being discrete (and spatial). *Theor. Popul. Biol.* 46: 363–394.
- [9] Fisher, R.A. 1930. *The Genetic Theory of Natural Selection*. Clarendon Press, Oxford.
- [10] Hastings, A. 1980. Disturbance, coexistence, history, and competition for space. *Theor. Popul. Biol.* 18: 363–373.
- [11] Hernandez, J.J. 1998. Dynamics of transitions between population interactions: a nonlinear interaction alpha-function defined. *Proc. R. Soc. Lond. B* 265: 1433–1440.
- [12] Herre, E.A., N. Knowlton, U.G. Mueller, and S.A. Rehner. 1999. The evolution of mutualisms: exploring the paths between conflict and cooperation. *Trends Ecol. Evol.* 14: 49–53.
- [13] Holley and Liggett. 1975. Ergodic theorems for weakly interacting systems and the voter model. *Ann. Probab.* 6: 198–206.
- [14] Hubbell, S.P. 2001. *The unified neutral theory of biodiversity and biogeography*. Princeton University Press, Princeton.
- [15] Johnson, N.C., D.L. Rowland, L. Corkidi, L.M. Egerton-Warburton, and E.B. Allen. 2003. Nitrogen enrichment alters mucorrhizal allocation at five mesic to semiarid grasslands. *Ecology* 84: 1895–1908.
- [16] Kerr, B., C. Neuhauser, B.J.M. Bohannan, and A.M. Dean. 2006. Local migration promotes competitive restraint in a host-pathogen 'tragedy of the commons.' *Nature* 442: 75–78.
- [17] Lanchier, N. and C. Neuhauser 2006a. A spatially explicit model for competition among specialists and generalists in a heterogeneous environment. *Ann. Appl. Probab.* 16: 1385–1410.
- [18] Lanchier, N. and C. Neuhauser 2006b. Stochastic spatial models of host-pathogen and host-mutualist interactions I. *Ann. Appl. Probab.* 16: 448–474.
- [19] Lanchier, N. and C. Neuhauser 2010. Stochastic spatial models of host-pathogen and host-mutualist interactions II. To appear in *Stoch. Models*.
- [20] Levin, S.A. 1974. Dispersion and population interaction. *Am. Nat.* 108: 207–228.
- [21] Maurer, 1999. *Untangling Ecological Complexity? The macroscopic perspective*. University Chicago Press, Chicago.
- [22] Petersen, W. 1932. Die Arten der Gattung *Swammerdamia* Hb. (Lep) mit Bemerkungen zur Mutationslehre. *Arch. Natur.* 1: 197–224.

-
- [23] Pimm, S.L. and M.L. Rosenzweig 1981. Competitors and habitat use. *Oikos* 37: 1–6.
- [24] Rosenzweig, M.L. 1981. A theory of habitat selection. *Ecology* 62: 327–335.
- [25] Rosenzweig, M.L. 1987a. Habitat selection as a source of biological diversity. *Evol. Ecol.* 1: 315–330.
- [26] Rosenzweig, M.L. 1987b. Editor’s coda: central themes of the symposium. *Evol. Ecol.* 1: 401–407.
- [27] Thompson, J.N. 1982. *Interaction and Coevolution*. Wiley, New York.
- [28] Thompson, J.N. 1988. Variation in interspecific interactions. *Annu. Rev. Ecol. Syst.* 19: 65–87.
- [29] Thompson, J.N. 1994. *The Coevolutionary Process*. Chicago University Press.
- [30] Thorpe, W.H. 1945. The evolutionary significance of habitat selection. *J. Anim. Ecol.* 14: 67–70.
- [31] Vreba, E.S. 1980. Evolution, species and fossils: How does life evolve. *South African Journal of Science* 76: 61–84.
- [32] Wright, S. 1931. Evolution in Mendelian populations. *Genetics* 16: 97–159.
- [33] Yodzis, P. 1978. *Competition for space and the structure of ecological communities*. Springer, New York.

Weakly Asymmetric Exclusion and KPZ

Jeremy Quastel*

Abstract

We review recent results on the anomalous fluctuation theory of stochastic Burgers, KPZ and the continuum directed polymer in one space dimension, obtained through the weakly asymmetric limit of the simple exclusion process.

Mathematics Subject Classification (2010). Primary 82C22; Secondary 60H15.

Keywords. Kardar-Parisi-Zhang equation, stochastic Burgers equation, stochastic heat equation, random growth, asymmetric exclusion process, anomalous fluctuations, directed polymers.

1. KPZ and Asymmetric Exclusion

We report on some progress on the behaviour of the Kardar-Parisi-Zhang equation (KPZ),

$$\partial_t h = -\frac{1}{2}(\partial_x h)^2 + \frac{1}{2}\partial_x^2 h + \dot{W} \quad (1)$$

where $\dot{W}(t, x)$ is Gaussian space-time white noise,

$$E[\dot{W}(t, x)\dot{W}(s, y)] = \delta_0(t - s)\delta_0(y - x). \quad (2)$$

Like many stochastic partial differential equations, it was introduced in the hope that it would reflect the behaviour of a large class of discrete models, but also lead to some simplifications. In reality, the well-posedness became a stumbling block, and the recent advances have actually come through an improved understanding of one particular discretization, the asymmetric simple exclusion process (ASEP). Since the behaviour is also easier to understand in the discrete model, we start there.

ASEP is a continuous time Markov process on the discrete lattice \mathbb{Z} with state space $\{0, 1\}^{\mathbb{Z}}$: $\eta(x) = 1$ if there is a particle at $x \in \mathbb{Z}$ and $\eta(x) = 0$ if there

*Research supported by the Natural Sciences and Engineering Council of Canada
Departments of Mathematics and Statistics, University of Toronto, 40 St. George St.,
Toronto, ON M5S 1L2. E-mail: quastel@math.toronto.edu.

is no particle at x . Each particle has an independent alarm clock which rings at exponentially distributed times, with rate one. When the alarm goes off the particle flips a biased coin to decide which way to jump. With probability p it chooses to attempt a jump one step to the right and with probability $q = 1 - p$ it chooses to attempt a jump one step to the left. However, the jump is achieved only if there is no particle in the way; otherwise, the jump is suppressed. We will always assume that $p < q$ so that the model is really asymmetric.

Besides the straightforwardness of its description, ASEP enjoys several convenient special properties: There is a simple family of invariant measures, the Bernoulli product measures, parametrized by $\rho \in [0, 1]$: If one chooses initially $\eta(0, x)$, $x \in \mathbb{Z}$ to be independent, with $P(\eta(0, x) = 1) = \rho = 1 - P(\eta(0, x) = 0)$ then one will see exactly the same distribution at a later time. One can also describe a special *second class particle* with a different rule than the others: It jumps to unoccupied sites as the others do, but if a particle wants to jump to where the second class particle is, the two particles exchange positions. If one watches the resulting process without distinguishing the second class particle from the others, it is the simple exclusion process with one extra particle. On the other hand, if one watches the resulting process without distinguishing the second class particle from a hole, it is the simple exclusion process without the extra particle.

The process of occupation variables $\eta(t, x)$ can be thought of as a discretization of the stochastic Burgers equation,

$$\partial_t u = -\frac{1}{2} \partial_x (u^2) + \frac{1}{2} \partial_x^2 u + \partial_x \dot{W} \quad (3)$$

formally satisfied by the derivative

$$u = \partial_x h \quad (4)$$

of (1). The invariant measure is supposed to be white noise. The object which is a discretization of KPZ itself is the associated height function,

$$h(t, x) = \begin{cases} 2N(t) + \sum_{0 < y \leq x} \hat{\eta}(t, y), & x > 0, \\ 2N(t), & x = 0, \\ 2N(t) - \sum_{x < y \leq 0} \hat{\eta}(t, y), & x < 0, \end{cases} \quad (5)$$

where $\hat{\eta}(x) = 2\eta(x) - 1$ and $N(t) = \{\# \text{ of particles which crossed } 1 \mapsto 0\} - \{\# \text{ of particles which crossed } 0 \mapsto 1\}$ up to time t . This just means that the height function takes a jump up wherever there is a particle, and a jump down wherever there is a hole. If we linearly interpolate the function $h(x)$, then a configuration of particles 01 is represented by a \vee in the height function, and a configuration 10 is represented by a \wedge , and the entire dynamics (including the $N(t)$), is that \wedge 's become \vee 's at rate p and \vee 's become \wedge 's at rate q .

We will be interested in two special initial conditions for the process. The *corner growth model* corresponds to having initially sites $\{0, 1, 2, \dots\}$ occupied

and sites $\{\dots, -2, -1\}$ unoccupied. The initial height function is $|x|$ and as we watch, the corner starts to fill in as a little random parabola, still keeping $h(t, x) = |x|$ for large x . At time t the particle initially at m will be at $\mathbf{x}(t, m)$ and the height function can be read off from

$$h(t, x) \geq 2m - x \quad \Leftrightarrow \quad \mathbf{x}(t, m) \leq x \quad (6)$$

On the other hand, if we start ASEP in *equilibrium*, by which we will always mean here with the Bernoulli product measure with density $1/2$, then, modulo a global height shift, the height function will be at each time t a symmetric random walk in x . For different t , the walks are *not* independent of each other.

The equilibrium initial data corresponds in the continuum to starting KPZ (1) with a two-sided Brownian motion; $h(0, x) = B(x)$. At a later time, one expects to see, besides the global height shift, a new, but not independent two-sided Brownian motion. Note that even if we start with a smooth initial data, what we expect to see at a small positive time is a version of the initial data which looks locally like a Brownian motion, and herein lies the problem of well-posedness for KPZ: The non-linear term $(\partial_x h)^2$ is clearly divergent since h as a function of x is supposed to have non-trivial quadratic variation. Naturally, one expects that an appropriate Wick ordering of the non-linearity can lead to well defined solutions, however, numerous attempts have led only to non-physical answers [9].

The correct interpretation is that of L. Bertini and G. Giacomin [8] where h is simply *defined* through the Hopf-Cole transformation

$$h(t, x) \stackrel{\text{def}}{=} -\log Z(t, x) \quad (7)$$

where $Z(t, x)$ is the well-defined [34] solution of the stochastic heat equation,

$$\partial_t Z = \frac{1}{2} \partial_x^2 Z - Z \dot{W}. \quad (8)$$

The key fact is that (8) is well-posed [34]. It is not known how to show directly that h defined through (7) satisfies (1), or, for that matter what it would mean to be a solution of (1), so all our results will really be about (8). What is known [8], is that the solution of (1) with the noise smoothed out in space, converges to (7) as the smoothing is removed, after a subtraction of a large global height shift. Note that one expects in such problems to have to make such shifts in the reference frame in order to observe the universal nontrivial fluctuation behaviour.

A large class of one dimension random growth models are governed by (1) and we have chosen to concentrate on ASEP here only because its tractability has led to progress. Another special model which is to some extent solvable is the polynuclear growth model [22]. A model which is simple to describe and intriguing to watch on a computer, but not particularly tractable, is *ballistic aggregation*. Here one has a stack of particles $n(x) \in \{0, 1, 2, \dots\}$ at each $x \in \mathbb{Z}$. Each site now has an independent exponential alarmclock, rate one, and when

the alarm rings the stack at x is increased to $\max\{n(x-1), n(x)+1, n(x+1)\}$. Unlike ASEP, the nonlinearity is not already quadratic, and the invariant measures are not well understood. The idea in the derivation of (1) is that if one writes very roughly $F(\nabla h)$ for the net macroscopic effect of the nonlinearity, and expands $F(\nabla h) = F^{(0)} + F^{(1)}\nabla h + \frac{1}{2}F^{(2)}(\nabla h)^2 + \dots$, the first two terms can be absorbed in global shifts, and generically the quadratic term gives the main nontrivial macroscopic effect.

2. Directed Random Polymers

The problem of directed random polymers is closely related to (1). The free energy of the discrete random polymer is

$$f_\beta(n, x) = \log E_{x,0} \left[e^{-\beta \sum_{i=1}^n W(i, b_i)} \right] \quad (9)$$

where $W(i, j)$ are independent identically distributed random variables, and $E_{x,0}$ is the expectation over a nearest neighbour random walk b_i starting at x and conditioned to hit 0 at time n . Note that the directed polymer (9) also makes sense in higher dimensions. In $d \geq 3$ there is a phase transition, with standard Gaussian behaviour in the *weak coupling regime* $0 < \beta < \beta_c < \infty$ [17]. This has led to a lot of work in probability (see [10] for a survey). Above β_c we are in the poorly understood *strong coupling regime*, where the main effect is that the probability in (9) is concentrated or *localized* on favorite paths. In dimensions $d = 1$ and 2 the two sources of randomness, the random path, and the random environment, are strongly coupled for all β , ie. $\beta_c = 0$.

In one dimension, there is an associated continuum model. Let

$$F_\beta(t, x) = \log E_{x,0} \left[: \exp \left\{ -\beta \int_0^t \dot{W}(s, b(s)) ds \right\} \right] \quad (10)$$

where $E_{x,0}$ is the expectation over the Brownian bridge $b(t)$ with $b(0) = x$ and $b(t) = 0$. $F_\beta(t, x)$ is the free energy of the *continuum directed random polymer*. The Wick ordered exponential just means that one must order the times in the series expansion, ie. the expectation in (10) is defined by

$$\sum_{n=0}^{\infty} (-\beta)^n \int_{0 \leq t_1 \leq \dots \leq t_n \leq t} \int_{\mathbb{R}^n} p_{t_1, \dots, t_n}(x_1, \dots, x_n) W(dt_1 dx_1) \cdots W(dt_n dx_n), \quad (11)$$

where $p_{t_1, \dots, t_n}(x_1, \dots, x_n)$ are the transition probabilities of the bridge and $W(dt dx)$ refer to Wiener integrals with respect to the space-time white noise.

In fact, if we let

$$Z_\beta(t, x) = \frac{e^{-x^2/2t}}{\sqrt{2\pi t}} \exp\{F_\beta(t, x)\} \quad (12)$$

then $Z_\beta(t, x)$ satisfies

$$\partial_t Z_\beta = \frac{1}{2} \partial_x^2 Z_\beta - \beta Z_\beta \dot{W}. \tag{13}$$

with delta function initial data

$$Z_\beta(0, x) = \delta_0(x). \tag{14}$$

To see this, note that (8) is really shorthand for the integral equation

$$Z_\beta(t, x) = \frac{e^{-x^2/2t}}{\sqrt{2\pi t}} + \int_0^t \int_{-\infty}^\infty \frac{e^{-(x-y)^2/2(t-s)}}{\sqrt{2\pi(t-s)}} Z_\beta(s, y) W(dy, ds). \tag{15}$$

Iterating the integral equation, we arrive at (11).

It is not hard to see that at the continuum level we can rescale

$$Z_\beta(t, x) \stackrel{\text{distr.}}{=} \beta^2 Z(\beta^4 t, \beta^2 x) \tag{16}$$

so there it is enough to consider $\beta = 1$.

3. The $t^{1/3}$ Law

We now turn to some of the physical predictions. The most important one is that the fluctuations at time t are supposed to be of nonstandard order $t^{1/3}$. This was originally predicted for (8) by the dynamic renormalization group [13], and later for the equivalent (1) by [21]. For ASEP it was predicted by mode coupling [33], and for directed polymers by [16].

About ten years ago, there was sudden, significant progress on two models, the totally asymmetric simple exclusion process (TASEP) where $q = 1, p = 0$, and the related polynuclear growth model, where one has determinantal formulas [27]. Very precise results showed that the fluctuations are in some cases related to the universal distributions of eigenvalues of random matrices, and fit into various universality classes depending only on the type of initial data. For the corner growth initial conditions K. Johansson [18] (see also [4], [22],[19],[7]) proved for TASEP that

$$-h(t, t^{2/3}x) \sim c(t) + t^{1/3}(\zeta(x) - \frac{1}{2}x^2) \tag{17}$$

where $\zeta(x)$ is the Airy_2 process. In particular, it is stationary in x and has as its one-dimensional marginals the GUE Tracy-Widom distribution, i.e., the limiting distribution of the scaled and centered largest eigenvalue in the Gaussian unitary ensemble. The cumulative distribution function is

$$F_{GUE}(s) = e^{-\int_s^\infty (y-s)u^2(y)dy} = \det(I - K_{\text{Ai}})_{L^2(s, \infty)}. \tag{18}$$

Here u is the unique solution of Painleve II equation, $u'' = (y + 2u^2)u$ satisfying $u(y) \sim \text{Ai}(y)$ as $y \rightarrow \infty$, and K_{Ai} is the operator with kernel

$$K_{\text{Ai}}(x, y) = \int_{-\infty}^\infty \sigma(v) \text{Ai}(x+v) \text{Ai}(y+v) dv \tag{19}$$

where $\text{Ai}(x)$ is the Airy function and $\sigma(v) = 1$ if $v \geq 0$ and 0 if $v < 0$ [29]. TASEP can also be mapped into a version of the discrete random polymer corresponding to $\beta = \infty$, with geometrically distributed $W(i, j)$ [18].

In equilibrium ($\rho = 1/2$), P. Ferrari and H. Spohn [12] proved that the space-time correlation functions of TASEP satisfy

$$\lim_{\varepsilon \searrow 0} \varepsilon^{-1} E[\hat{\eta}(\varepsilon^{-3/2}t, \varepsilon^{-1}x)\hat{\eta}(0, 0)] = t^{-2/3}\Phi(xt^{-2/3}) \quad (20)$$

for a special universal Φ (see [12] for the very complicated formula). The limit process was studied in [3].

These computations represented genuine breakthroughs, but their applicability was restricted to a few models where there is a determinantal structure. The main goal now is to prove that these behaviours are universal, ie. find proofs that work for a broad class of models. At a more modest level, one can hope to show that (17) and (20) extend to ASEP or KPZ/Stochastic Burgers itself. In terms of the directed random polymer, the universality conjecture is that

$$f_\beta(n, n^{2/3}x) \sim a_\beta n + b_\beta n^{1/3}\zeta(x) \quad (21)$$

where a_β, b_β are non-universal and $\zeta(x)$ is as above in (17). Of course one has to assume some reasonable decay on the tails of the common distribution of the $W(i, j)$'s.

4. Weakly Asymmetric Limit of Simple Exclusion

There are various scaling regimes for ASEP. Consider the process on $\varepsilon\mathbb{Z}$ with scaling parameter $0 < \varepsilon \ll 1$. If one fixes the asymmetry and starts from a slowly varying initial profile, with $\hat{\eta}(0, x)$ independent with $E[\hat{\eta}(0, x)] = u(0, x)$, $x \in \varepsilon\mathbb{Z}$, then at a later time $\varepsilon^{-1}t$, the density profile of $\hat{\eta}$ will have evolved according to the inviscid Burgers equation,

$$\partial_t u = -\frac{1}{2}\partial_x(u^2). \quad (22)$$

The initial fluctuations are transported along the characteristic lines. On the other hand, one can let the asymmetry depend on ε as $\varepsilon \searrow 0$. If one takes $p = \frac{1}{2}(1 - \varepsilon)$, $q = \frac{1}{2}(1 + \varepsilon)$ then one has to wait until time $\varepsilon^{-2}t$, and one sees the viscous Burgers equation,

$$\partial_t u = -\frac{1}{2}\partial_x(u^2) + \frac{1}{2}\partial_x^2 u. \quad (23)$$

The fluctuations associated to (23) are Gaussian, because the process has been steered too close to the symmetric case.

In order to understand the intermediate scaling at which there are non-trivial fluctuations we use the stochastic Burgers (3) as a proxy for its discretization ASEP, and rescale it in equilibrium,

$$u_\varepsilon(t, x) = \varepsilon^{-\sigma} u(\varepsilon^{-\alpha}t, \varepsilon^{-1}x) \tag{24}$$

with the ε^{-1} corresponding to our rescaled lattice. To preserve the invariant white noise, we have to have $\sigma = 1/2$. The stochastic Burgers equation (3) becomes

$$\partial_t u_\varepsilon = -\frac{1}{2}\varepsilon^{-\alpha+3/2}\partial_x(u_\varepsilon^2) + \frac{1}{2}\varepsilon^{-\alpha+2}\partial_x^2 u_\varepsilon + \varepsilon^{-\frac{\alpha+2}{2}}\partial_x \dot{W}. \tag{25}$$

A first guess is to take $\alpha = 3/2$. The viscous and noise terms together represent a small Ornstein-Uhlenbeck perturbation, and the process appears to go to the renormalization fixed point $\partial_t u = -\frac{1}{2}\partial_x(u^2)$. Whatever this limit is – and we do *not* understand it – it is *not* the inviscid limit $\varepsilon \searrow 0$ of $\partial_t u_\varepsilon = -\frac{1}{2}\partial_x(u_\varepsilon^2) + \frac{1}{2}\varepsilon\partial_x^2 u_\varepsilon$ since the latter does not preserve the initial white noise, as can be easily checked from the Lax-Oleinik formula for the solution. It is interesting to note that appropriate dispersive perturbations, such as $\partial_t u_\varepsilon = -\frac{1}{2}\partial_x(u_\varepsilon^2) + \frac{1}{2}\varepsilon\partial_x^3 u_\varepsilon$, do preserve white noise [23].

A more moderate approach is to take $\alpha = 2$, and rescale the nonlinearity by $\varepsilon^{1/2}$ to compensate. KPZ/Stochastic Burgers is invariant under this rescaling, and hence one can anticipate an invariance principle under which it is the limit of processes like ASEP. Since the size of the nonlinearity is roughly $q - p$, it corresponds to taking

$$p = \frac{1}{2}(1 - \varepsilon^{1/2}) \quad q = \frac{1}{2}(1 + \varepsilon^{1/2}). \tag{26}$$

This is the *weakly asymmetric limit*.

J. Gärtner [14] discovered that there is an exact discrete Hopf-Cole transformation for ASEP. Let

$$\rho = \frac{1}{2} \log(q/p), \quad \nu = \sqrt{pq}, \quad \lambda = p + q - 2\sqrt{pq}, \tag{27}$$

and

$$z(t, x) = A \exp\{-\rho h(t, x) + \lambda t\}. \tag{28}$$

Then

$$\partial_t z = \nu \Delta z + z \dot{W} \tag{29}$$

where Δ is the lattice Laplacian $\Delta f(x) = \frac{1}{2}\{f(x+1) - 2f(x) + f(x-1)\}$, and \dot{W} refers to the derivative of certain jump martingales, which should be thought of as a messy version of space-time white noise.

Now we can make the connection between the discrete model and the continuum equations precise. Consider ASEP with the weak asymmetry (26). Corresponding to this process we have a $z_\varepsilon(t, x)$ as in (28). Observe it at time $\varepsilon^{-2}t$ and on space scale $\varepsilon^{-1}x$, so that rescaled

$$z_\varepsilon(t, x) = z_\varepsilon(\varepsilon^{-2}t, \lfloor \varepsilon^{-1}x \rfloor). \tag{30}$$

Assume that A and the initial data are chosen so that

$$z_\varepsilon(0, x) \rightarrow Z(0, x) \quad (31)$$

in the sense of convergence in distribution.

The first result, due to L. Bertini and G. Giacomin [8], is for initial data not far from equilibrium. Take $A = 1$ and assume that for each $p = 2, 4, \dots$ there is $C = C(p) < \infty$ such that

$$E[z_\varepsilon^p(0, x)] \leq C e^{C|x|}. \quad (32)$$

In the equilibrium situation $\log z_\varepsilon(0, x)$ will be roughly Gaussian mean 0 and variance $|x|$, in which case we have (32). So (32) is a way of saying that the initial data scales diffusively. Under these conditions, it is proved in [8] that

$$z_\varepsilon(t, x) \rightarrow Z(t, x) \quad (33)$$

in the sense of distributional convergence of stochastic processes, where $Z(t, x)$ is the solution of the stochastic heat equation (8) with initial data $Z(0, x)$.

The corner growth model does not satisfy (32) and the scaling is different. Since initially $h(x) = |x|$ one has to take

$$A = \varepsilon^{-1/2} \quad (34)$$

and one ends up with delta function initial data,

$$Z(0, x) = \delta_0(x). \quad (35)$$

It is shown in [1] that the method of [8] can be extended to prove that (33) holds in this case as well.

5. KPZ/Stochastic Burgers in Equilibrium: The Method of Second Class Particles

The space-time correlation functions of the occupation variable for ASEP in equilibrium turn out to be equal to both the transition probabilities of the second class particle $\mathbf{y}(t)$ and the discrete Laplacian of the variance of the height function [22]. After the weakly asymmetric rescaling, the identity for $x \in \varepsilon\mathbb{Z}$ reads

$$E_\varepsilon[u_\varepsilon(t, x)u_\varepsilon(0, 0)] = \varepsilon^{-1}P_\varepsilon(\mathbf{y}_\varepsilon(t) = x) = \Delta_\varepsilon \text{Var}_\varepsilon(h_\varepsilon(t, x)) \quad (36)$$

where E_ε , P_ε , Var_ε refer to the expectation, probability, and variance with respect to the weakly asymmetric process, ie. with p, q as in (26), $\Delta_\varepsilon f(x) = \frac{1}{2}\varepsilon^{-2}(f(x + \varepsilon) - 2f(x) + f(x - \varepsilon))$, and

$$u_\varepsilon(t, x) = \varepsilon^{-1/2}\hat{\eta}(\varepsilon^{-2}t, [\varepsilon^{-1}x]) \quad \mathbf{y}_\varepsilon(t) = \varepsilon\mathbf{y}(\varepsilon^{-2}t). \quad (37)$$

M. Balázs, T. Seppäläinen and the author [6] then showed that for each $1 \leq m < 3$, there is a $C = C(m) < \infty$ such that for all $t \geq 1$,

$$C^{-1}t^{1/3} \leq E_\varepsilon[|\mathbf{y}_\varepsilon(t)|^m]^{1/m} \leq Ct^{1/3}. \tag{38}$$

With some work we can pass to the limit $\varepsilon \searrow 0$ to conclude that the correlation functions of stochastic Burgers make sense, at least as a probability measure in the space variable, and satisfies

$$E[u(t, x)u(0, 0)] = \frac{1}{2}\partial_x^2 \text{Var}(h(t, x)) \tag{39}$$

and the bounds obtained from the limit of (38),

$$C^{-1}t^{1/3} \leq \left(\int |x|^m E[u(t, x)u(0, 0)] dx \right)^{1/m} \leq Ct^{1/3} \tag{40}$$

which identifies the correct order of fluctuations.

One also might guess that there is a limiting $\mathbf{y}_0(t) = \lim_{\varepsilon \searrow 0} \mathbf{y}_\varepsilon(t)$ whose transition probabilities give the correlation functions $E[u(t, x)u(0, 0)]$. $\mathbf{y}_0(t)$ would satisfy a stochastic differential equation,

$$d\mathbf{y}_0(t) = u(t, \mathbf{y}_0(t))dt + db \tag{41}$$

where b is yet another Brownian motion. For each t , $u(t, x)$ is a white noise in x . So $\mathbf{y}_0(t)$ is a kind of dynamical version of the Sinai diffusion. Note that (41) does not really make sense: The field $u(t, x)$ is just too wild, and $\mathbf{y}_0(t)$ will not even be absolutely continuous with respect to $b(t)$. But it does give a hint that many of the miracles of ASEP are actually inherited by KPZ/Stochastic Burgers, if only one could get the calculus right.

We now give a brief hint at the proof of the key estimate (38). It is adapted from earlier work of [5], which in turn goes back to [11].

The main problem is to estimate the probability of an event like

$$\mathcal{A}_\varepsilon = \{\mathbf{y}(t) \leq -\varepsilon^{1/3}t^{2/3}y\}$$

in the weakly asymmetric simple exclusion process. We couple two copies of the process, one with density $\frac{1}{2}$ with another with density $\frac{1}{2} - \varepsilon^{-1/6}t^{-1/3}y$, in such a way that all the extra particles in the first copy are second class particles. Let $\text{Curr}_{t,\varepsilon}$ be the net current of second class particles crossing the space-time line between $(0, 0)$ and $(t, -\varepsilon^{1/3}t^{2/3})$. Let

$$\mathcal{A}'_\varepsilon = \{\text{Curr}_{t,\varepsilon} \geq \varepsilon^{1/6}t^{1/3}y^2\}.$$

Then one can show that because of the ordering, $P(\mathcal{A}_\varepsilon) \sim P(\mathcal{A}'_\varepsilon)$. The expectation of this current can be computed without much difficulty: $E[\text{Curr}_{t,\varepsilon}] \sim \frac{1}{2}\varepsilon^{1/6}t^{1/3}y^2$. The key point is that there is a general fact [5] which relates the variance of the current back to the first moment of the second class particle,

$$\text{Var}(\text{Curr}_{t,\varepsilon}) = cE[|\mathbf{x}(t) + \varepsilon^{1/3}t^{2/3}|].$$

This means that we can use Chebyshev's inequality to estimate

$$P\{\mathbf{y}(t) \leq -\varepsilon^{1/3}t^{2/3}y\} \leq C \frac{\text{Var}(\text{Curr}_{t,\varepsilon})}{(\varepsilon^{1/6}t^{1/3}y^2)^2} = C' \frac{E[|\mathbf{x}(t) + \varepsilon^{1/3}t^{2/3}|]}{(\varepsilon^{1/6}t^{1/3}y^2)^2},$$

which miraculously gives (38).

The reason we can only get moments $m < 3$ is the y^4 in the last denominator, which cannot be improved with these methods. It should be emphasized that many things are happening on the same scale $t^{1/3}$ in this problem, and the second class particle method is not necessarily identifying the exact source of the anomalous fluctuations.

Using related ideas, T. Seppäläinen [28] has recently shown that for the polymer where the $W(i, j)$ have log-Gamma distribution, for $1 \leq p < 3/2$

$$E[|f_1(n, 0) - cn|^p] \leq Cn^{p/3},$$

which again identifies the correct order, but with fewer moments.

6. Tracy-Widom Formula for ASEP

C. Tracy and H. Widom [30], [31] discovered a formula for the probability distribution of the position at time t of the m th particle in the corner growth model. Let $\gamma = q - p$ and $\tau = p/q$.

$$P(\mathbf{x}(\gamma^{-1}t, m) \leq x) = \int_{S_{\tau+}} \frac{d\mu}{\mu} \prod_{k=0}^{\infty} (1 - \mu\tau^k) \det(I + \mu J_{t,m,x,\mu})_{L^2(\Gamma_\eta)} \quad (42)$$

where $S_{\tau+}$ is a circle centered at zero of radius between τ and 1, and where the kernel of the determinant is given by

$$J_{t,m,x,\mu}(\eta, \eta') = \int_{\Gamma_\zeta} \exp\{\Lambda_{t,m,x}(\zeta) - \Lambda_{t,m,x}(\eta')\} \frac{f(\mu, \zeta/\eta')}{\eta'(\zeta - \eta)} d\zeta. \quad (43)$$

The variables η and η' are on a circle Γ_η centered at zero and of radius between τ and 1, and the ζ integral is on a circle Γ_ζ centered at zero and of radius between 1 and τ^{-1} , and

$$f(\mu, z) = \sum_{k=-\infty}^{\infty} \frac{\tau^k}{1 - \tau^k \mu} z^k, \quad \Lambda_{t,m,x}(\zeta) = -x \log(1 - \zeta) + \frac{t\zeta}{1 - \zeta} + m \log \zeta. \quad (44)$$

There are also formulas for other initial data and for transition probabilities. However, they are extremely unwieldy, involving large numbers of contour integrations. Although the formula (42)-(44) looks pretty complicated, it is actually simple enough to start asymptotic analysis. In particular, in [32], Tracy and Widom used the method of steepest descent on (42) to prove that (17) holds for ASEP, in the sense of one dimensional marginals.

7. The Crossover Distributions

From the weakly asymmetric limit, we learn that the distribution functions of $F(t, x) = F_1(t, x)$ from (10),

$$F_t(s) = P(F(t, x) + \frac{t}{4} \leq s), \tag{45}$$

the *crossover distributions*, can be obtained from the limit

$$F_t(s) = \lim_{\varepsilon \searrow 0} P_\varepsilon(\mathbf{x}(\gamma^{-1}t, m) \leq x),$$

where $\gamma = q - p = \varepsilon^{1/2}$ and

$$m = \frac{1}{2}\varepsilon^{-1/2} \left(-s + \log \sqrt{2\pi t} + \log(\varepsilon^{-1/2}/2) + \frac{x^2}{2t} \right) + t/4 + x/2. \tag{46}$$

The limit was computed independently by T. Sasamoto and H. Spohn [24] and by G. Amir, I. Corwin, and the author [1] using the method of steepest descent on the Tracy-Widom formula (43). There are however serious new complications because the poles of the function f from (44) are becoming dense about the saddle point (see [1]).

The exact formula for the distribution functions F_t reads

$$F_t(s) = 1 - \int_{-\infty}^{\infty} G(r)f(a - r)dr \tag{47}$$

where $G(r) = e^{-e^{-r}}$ is the Gumbel distribution and

$$f(r) = 2^{1/3}t^{-1/3} \det(I - K_{\sigma_t}) \text{tr} \left((I - K_{\sigma_t})^{-1} P_{\text{Ai}} \right)_{L^2(2^{1/3}t^{-1/3}r, \infty)}. \tag{48}$$

The operators are defined through their kernels: $P_{\text{Ai}}(x, y) = \text{Ai}(x)\text{Ai}(y)$ and

$$K_{\sigma_t}(x, y) = \int_{-\infty}^{\infty} \tilde{\sigma}_t(v)\text{Ai}(x + v)\text{Ai}(y + v)dv + 2^{1/3}t^{-1/3}\pi G_{\frac{x-y}{2}} \left(\frac{x+y}{2} \right) \tag{49}$$

where

$$\tilde{\sigma}_t(v) = \frac{1}{1 - e^{-2^{-1/3}t^{1/3}v}} - \frac{1}{2^{-1/3}t^{1/3}v}, \tag{50}$$

is a smooth non-decreasing function with $\lim_{v \searrow -\infty} \tilde{\sigma}_t(v) = 0$ and $\lim_{v \searrow \infty} \tilde{\sigma}_t(v) = 1$. $G_a(x)$ is a Hilbert transform of the product of Airy functions, which can be partially computed,

$$G_a(x) = \frac{1}{2\pi^{3/2}} \int_0^{\infty} \xi^{-1/2} \sin \left(x\xi + \frac{\xi^3}{12} - \frac{a^2}{\xi} + \frac{\pi}{4} \right) d\xi.$$

The formula can be variously interpreted as giving an exact formula for the one point distributions of the stochastic heat equation (8) with delta initial data,

KPZ with narrow wedge initial conditions, or the distribution of the continuum random polymer free energy.

It is not hard to check from the formula that

$$\lim_{t \nearrow \infty} F_t(t^{1/3}s) \rightarrow F_{\text{GUE}}(2^{1/3}s) \quad (51)$$

and from the series (11) that

$$\lim_{t \searrow 0} F_t(2^{-1/2}\pi^{1/4}t^{1/4}s) = \int_{-\infty}^s \frac{e^{-x^2/2}}{\sqrt{2\pi}} dx. \quad (52)$$

Thus the family of distributions *crosses over* from Gaussian behaviour for small time, to GUE Tracy-Widom for large time.

For each t , $F(t, x)$ is a stationary process in x with one dimensional distributions given by the crossover formula. As $t \searrow 0$, one can obtain the full process level limit: It just comes from the first term (the Gaussian term) in the chaos expansion (11). A natural conjecture is that as $t \nearrow \infty$, $t^{-1/3}F(t, x)$, converges to the process $\text{Airy}_2(x)$.

Formula (49) is naturally compared to the determinantal formula (18) for F_{GUE} . In [1], there is also a version of the crossover formula generalizing the Painlevé II representation for F_{GUE} , reminiscent of inverse scattering theory:

$$\det(I - K_{\sigma_t})_{L^2(r, \infty)} = e^{-\int_r^\infty (x-r)V_t(x)dx} \quad (53)$$

where $V_t(x)$ satisfies both

$$V_t(x) = \int_{-\infty}^{\infty} \tilde{\sigma}'_t(v)q_v^2(x)dv + 2^{1/3}t^{-1/3}\partial_v Hq_v^2(x) \quad \text{and} \quad L_{2V_t}q_v = vq_v \quad (54)$$

where $q_v(x) \sim \text{Ai}(v+x)$ as $x \rightarrow \infty$, $L_V = \frac{d^2}{dx^2} - x - V$ is the Stark operator, and Hq_v^2 is the Hilbert transform in v of q_v^2 .

At this point we review what has been accomplished. Actually, all that we have done is show is that one more model, KPZ itself, belongs to the KPZ universality class. Despite the appearance of generality, KPZ is really no more fundamental than ASEP. So now KPZ and the continuum directed polymer (10) are added to the short list of models (TASEP, polynuclear growth, ASEP) for which one is able to rigorously establish some of the behaviour one expects to be universal. For generic models like ballistic aggregation one can still say essentially nothing.

8. The Intermediate Coupling Regime for Random Polymers

On the other hand, for the discrete polymers (9) we *can* say something universal [2]. Here there is also a weakly asymmetric limit. Setting $n \sim \varepsilon^{-2}$ we take as

asymmetry $n^{-1/4} \sim \varepsilon^{1/2}$. In other words we consider

$$f_{\beta n^{-1/4}}(n, x) = \log E_{x,0} \left[e^{-\frac{\beta}{n^{1/4}} \sum_{i=1}^n W(i, b_i)} \right]. \quad (55)$$

Assume the common distribution of the $W(i, j)$ has bounded moment generating function so that the discrete polymer makes sense and, for normalization, mean zero and variance one. Then it is not hard to check that as $n \nearrow \infty$, $f_{\beta n^{-1/4}}(n, x) - c_\beta n^{1/2}$ converges in distribution to that of $F_\beta(t, x)$ given by (10). This is proved by expanding the exponential on the left hand side of (55) and showing that the discrete chaos expansion converges term by term to the continuous chaos expansion (11). Hence we can show for *any* such distribution on the $W(i, j)$,

$$\lim_{n \nearrow \infty} P(f_{\beta n^{-1/4}}(n, x) - c_\beta n^{1/2} \leq s) = F_{\beta^4}(s) \quad (56)$$

So there is a region intermediate between the weak coupling ($\beta = 0$) and the strong coupling ($\beta > 0$) regimes where we can *universally* observe rigorously the transition in behaviour from Gaussian to F_{GUE} fluctuations, given by the crossover distributions (47). In particular, for any reasonable distribution of $W(i, j)$ we have the following weak version of *universality*,

$$\lim_{\beta \nearrow \infty} \lim_{n \nearrow \infty} P(\beta^{-4/3} (f_{\beta n^{-1/4}}(n, x) - c_\beta n^{1/2}) \leq s) = F_{GUE}(s). \quad (57)$$

Acknowledgement

The author would like to take this opportunity to acknowledge the very substantial contributions of T. Alberts, G. Amir, M. Balázs, I. Corwin, K. Khanin and T. Seppäläinen to the research described in this report. Additional thanks to T. Alberts, I. Corwin and K. Khanin for critical reading of the manuscript.

References

- [1] G. Amir, I. Corwin, J. Quastel, *Probability Distribution of the Free Energy of the Continuum Directed Random Polymer in 1 + 1 dimensions*, 2010, arXiv:1003.0443v2.
- [2] T. Alberts, K. Khanin, J. Quastel, *The Intermediate Disorder Regime for Directed Polymers in Dimension 1 + 1*, 2010, arXiv:1003.1885
- [3] J. Baik, P.L. Ferrari, S. Peche, *Limit process of stationary TASEP near the characteristic line*.2009, arXiv:0907.0226
- [4] J. Baik,, E. Rains, *Limiting distributions for a polynuclear growth model with external sources*. J. Statist. Phys. 100 523–541, 2000.
- [5] M. Balázs, T. Seppäläinen. *Order of current variance and diffusivity in the asymmetric simple exclusion process* math/0608400, to appear in Ann. Math.

- [6] M. Balázs, J. Quastel, T. Seppäläinen. *Scaling exponent for the Hopf-Cole solution of KPZ/stochastic Burgers*. arXiv:0909.4816v1
- [7] A. Borodin and P.L. Ferrari. *Large time asymptotics of growth models on space-like paths I: PushASEP*. Electron. J Probab., 13:1380–1418, 2008.
- [8] L. Bertini, G. Giacomin. *Stochastic Burgers and KPZ equations from particle systems*. Comm. Math. Phys. 183:571–607, 1997.
- [9] T. Chan, *Scaling limits of Wick ordered KPZ equation*, Comm. Math. Phys. 209:671–690,2000.
- [10] F. Comets, T. Shiga, N. Yoshida *Probabilistic analysis of directed polymers in a random environment: a review*. Stochastic analysis on large scale interacting systems, 115–142, Adv. Stud. Pure Math., 39, Math. Soc. Japan, Tokyo, 2004.
- [11] E. Cator, P. Groeneboom. *Second class particles and cube root asymptotics for Hammersleys process*. Ann. Probab., 34, 2006.
- [12] P.L. Ferrari, H. Spohn *Scaling limit for the space-time covariance of the stationary totally asymmetric simple exclusion process*. Comm. Math. Phys. 265, 1–44 2006.
- [13] D. Forster, D. R. Nelson, and M. J. Stephen, *Large-distance and long-time properties of a randomly stirred fluid* Phys. Rev. A 16, 732 (1977).
- [14] J. Gärtner, *Convergence towards Burgers equation and propagation of chaos for weakly asymmetric exclusion process*. Stoch. Proc. Appl.27:233–260, 1988.
- [15] T. Halpin-Healy and Y.-C. Zhang, *Kinetic roughening phenomena, stochastic growth, directed polymers and all that*, Physics Reports 254:215–414,1995.
- [16] D.A. Huse and C.L. Henley, *Pinning and Roughening of Domain Walls in Ising Systems Due to Random Impurities*, Phys. Rev. Lett. 54:2708–2711,1985.
- [17] J.Z. Imbrie and T. Spencer, *Diffusion of directed polymers in a random environment*, J. Statist. Phys. 52:609–626,1988.
- [18] K. Johansson, *Shape fluctuations and random matrices*. Comm. Math. Phys., 209:437–476, 2000.
- [19] K. Johansson. *Discrete polynuclear growth and determinantal processes*. Comm. Math. Phys., 242:277–329, 2003.
- [20] J. Krug and H. Spohn. *Kinetic roughening of growing surfaces*. In C. Godrèche (Ed.), Solids far from equilibrium pp. 117130. Cambridge University Press. 1992.
- [21] K. Kardar, G. Parisi, Y.Z. Zhang. *Dynamic scaling of growing interfaces*. Phys. Rev. Lett. 56:889–892, 1986.
- [22] M. Prähofer and H. Spohn. *Scale invariance of the PNG droplet and the Airy process*. J. Stat. Phys., 108:1071–1106, 2002.
- [23] J. Quastel, B. Valkó, *KdV preserves white noise*. Comm. Math. Phys. 277;707–714, 2008.
- [24] T. Sasamoto, H. Spohn. *Exact height distributions for the KPZ equation with narrow wedge initial condition*. arXiv:1002.1879,
- [25] T. Sasamoto, H. Spohn. *Universality of the one-dimensional KPZ equation*. arXiv:1002.1883,

-
- [26] T. Sasamoto, H. Spohn. *The crossover regime for the weakly asymmetric simple exclusion process*. arXiv:1002.1873.
 - [27] G.M. Schütz *Exact solution of the master equation for the asymmetric exclusion process*. J. Stat. Physics 88 (1997), 427445.
 - [28] T. Seppäläinen. *Scaling for a one-dimensional directed polymer with boundary conditions*. arXiv:0911.2446
 - [29] C. Tracy and H. Widom. *Level-spacing distributions and the Airy kernel*. Comm. Math. Phys., 159:151–174, 1994.
 - [30] C. Tracy and H. Widom. *Integral formulas for the asymmetric simple exclusion process*. Comm. Math. Phys., 279:815–844, 2008.
 - [31] C. Tracy and H. Widom. *A Fredholm determinant representation in ASEP*. J. Stat. Phys., 132:291–300, 2008.
 - [32] C. Tracy and H. Widom. *Asymptotics in ASEP with step initial condition*. Comm. Math. Phys., 290:129–154, 2009.
 - [33] H. van Beijeren, R. Kutner, H. Spohn. *Excess noise for driven diffusive systems*. Phys. Rev. Lett. 54:2026–2029, 1985.
 - [34] J. Walsh. *An introduction to stochastic partial differential equations*. In: Ecole d’Ete de Probabilites de Saint Flour XIV, Lecture Notes in Mathematics n. 1180. Berlin: Springer, 1986.

Stein's Method, Self-normalized Limit Theory and Applications

Qi-Man Shao*

Abstract

Stein's method is a powerful tool in estimating accuracy of various probability approximations. It works for both independent and dependent random variables. It works for normal approximation and also for non-normal approximation. The method has been successfully applied to study the absolute error of approximations and the relative error as well. In contrast to the classical limit theorems, the self-normalized limit theorems require no moment assumptions or much less moment assumptions. This paper is devoted to the latest developments on Stein's method and self-normalized limit theory. Starting with a brief introduction on Stein's method, recent results are summarized on normal approximation for smooth functions and Berry-Esseen type bounds, Cramér type moderate deviations under a general framework of the Stein identity, non-normal approximation via exchangeable pairs, and a randomized exponential concentration inequality. For self-normalized limit theory, the focus will be on a general self-normalized moderate deviation, the self-normalized saddlepoint approximation without any moment assumption, Cramér type moderate deviations for maximum of self-normalized sums and for Studentized U-statistics. Applications to the false discovery rate in simultaneous tests as well as some open questions will also be discussed.

Mathematics Subject Classification (2010). Primary 60F10, 60F05, 60G50; Secondary 60F15, 62E20, 62F03, 62F05, 00B10.

Keywords. Stein method, normal approximation, non-normal approximation, self-normalized sum, Studentized statistics, limit theory, large deviation, moderate deviation, concentration inequality, Berry-Esseen inequality, false discovery rate, simultaneous tests

*The author would like to thank Louis H.Y. Chen, Bing-Yi Jing, Weidong Liu, Qiyang Wang, Wang Zhou, Sourav Chatterjee, Vicotr de la Pena, Xiao Fang, Larry Goldstein, Xuming He, Tze Leung Lai for working with him on related projects and for their helpful comments. The research is partially supported by Hong Kong Research Grants Council (CERG-602206 and 602608).

Department of Mathematics, Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong, China. E-mail: maqmshao@ust.hk.

1. Introduction

Let $W := W_n$ be a random variable of interest. Assume that the limiting distribution of W_n is standard normal. A natural question is the accuracy of the approximation. There are mainly two approaches for estimating the error of the normal approximation. One approach is to study the absolute error $\sup_z |P(W \leq z) - \Phi(z)|$ via Berry-Esseen type bounds, where Φ is the standard normal distribution function. Another approach is to estimate the relative error of $P(W \geq z)$ to $1 - \Phi(z)$ through the Chernoff large deviation or the Cramér type moderate deviation. In this paper we mainly focus on two types of W : W satisfies a general framework of Stein's identity, and W is a self-normalized sum or a Studentized statistic.

When W is a standardized sum of independent random variables, Berry-Esseen bounds, Chernoff large deviations, and Cramér moderate deviations are extensively studied under certain moment conditions which are also necessary for these results. A standard approach to these classical results is Fourier methods and/or conjugate methods. However, these methods are much more difficult to apply for W under dependence structure. In a paper in the Proceedings of the Sixth Berkeley Symposium, Stein (1972) introduced a totally different method to determine the accuracy of the normal approximation to the distribution of a sum of dependent random variables satisfying a mixing condition. Since then many developments have taken place, both in extending the method beyond normal approximation and in applying the method to problems in other areas. The first focus in this paper is on the latest development on Stein's method especially when W satisfies a general framework of Stein identity. Starting with a brief introduction of Stein's method, Section 2 will be devoted to normal approximation for smooth functions and Berry-Esseen bounds, Cramér type moderate deviations, non-normal approximation via exchangeable pairs approach, and a randomized exponential concentration inequality.

In contrast to the standardized sums, it is now well-understood that self-normalized sums usually preserve much better properties and self-normalized limit theorems (namely, limit theorems for self-normalized sums) require no moment assumptions or much less moment assumptions than the classical limit theorems do. The second focus in this paper is on the latest development on self-normalized limit theory. Section 3 will summarize results on a general self-normalized moderate deviation, the self-normalized saddlepoint approximation without any moment assumption, Cramér type moderate deviations for maximum of self-normalized sums and for Studentized U-statistics. Applications to the false discovery rate in simultaneous tests will be discussed in Section 4.

2. Stein's Method

The classical approach to the central limit theorem and the accuracy of approximations for independent random variables relied heavily on Fourier methods.

Without independence, however, Fourier methods are much more difficult to apply, and bounds for the accuracy of approximations become even more difficult to find. It was Charles Stein in 1972 who introduced a startling technique, now known simply as Stein’s method, for normal approximation. The method works not only for independent random variables but also for dependent variables. It can also give bounds for accuracy of approximations. Extensive applications of Stein’s method to obtain uniform and non-uniform Berry-Esseen-type bounds for independent and dependent random variables can be found in, for example, Diaconis (1977), Baldi, Rinott and Stein (1989), Barbour (1990), Dembo and Rinott (1996), Goldstein and Reinert (1997), Chen and Shao (2001, 2004, 2007), Chatterjee (2008), and Nourdin and Peccati (2009). Stein’s ideas have been applied to many other probability approximations, notably to Poisson, Poisson process, compound Poisson and binomial approximations. Stein’s method has also found diverse applications in a wide range of fields, see for example, Arratia, Goldstein and Gordon (1990), Barbour, Holst and Janson (1992), and Chen (1993). Expositions of Stein’s method and its applications in normal and other distributional approximations can be found in Diaconis and Holmes (2004), Barbour and Chen (2005) and Chen, Goldstein and Shao (2010). In this section starting with a brief introduction to Stein’s method, we summarize some latest developments in this area.

2.1. Stein’s equation. Let Z be a standard normally distributed random variable and let \mathcal{C}_{bd} be the set of continuous and piecewise continuously differentiable functions $f: R \rightarrow R$ with $E|f'(Z)| < \infty$. Stein’s method rests on the following characterization.

Lemma 2.1. Let W be a real valued random variable. Then W has a standard normal distribution if and only if

$$E f'(W) = E\{W f(W)\}, \tag{2.1}$$

for all $f \in \mathcal{C}_{bd}$.

The proof of necessity is essentially a direct consequence of integration by parts. For the sufficiency, let $f(w) := f_z(w)$ denote the solution of the equation

$$f'(w) - w f(w) = I(w \leq z) - \Phi(z), \tag{2.2}$$

where z is a fixed number. It is easy to see that f_z is given by

$$f_z(w) = \begin{cases} \sqrt{2\pi}e^{w^2/2}\Phi(w)[1 - \Phi(z)] & \text{if } w \leq z, \\ \sqrt{2\pi}e^{w^2/2}\Phi(z)[1 - \Phi(w)] & \text{if } w \geq z \end{cases} \tag{2.3}$$

and that f_z is a bounded continuous and piecewise continuously differentiable function. Moreover, f_z has the following properties (see, e.g., Chen and Shao

(2005)): for all real w , u and v ,

$$0 < f_z(w) \leq \min(\sqrt{2\pi}/4, 1/|z|), \quad (2.4)$$

and

$$|f'_z(w)| \leq 1, \quad |f'_z(w) - f'_z(v)| \leq 1. \quad (2.5)$$

Equation (2.2) is a particular case of the more general Stein equation

$$f'(w) - wf(w) = h(w) - Eh(Z), \quad (2.6)$$

to be solved for f given a real valued measurable function h with $E|h(Z)| < \infty$. Similarly to (2.3), the solution $f = f_h$ is given by

$$\begin{aligned} f_h(w) &= e^{w^2/2} \int_{-\infty}^w [h(x) - Eh(Z)] e^{-x^2/2} dx \\ &= -e^{w^2/2} \int_w^{\infty} [h(x) - Eh(Z)] e^{-x^2/2} dx. \end{aligned} \quad (2.7)$$

If h is bounded, then

$$\|f_h\| \leq \sqrt{\pi/2} \|h(\cdot) - Eh(Z)\| \leq 2\|h\|, \quad (2.8)$$

and

$$\|f'_h\| \leq 2\|h(\cdot) - Eh(Z)\| \leq 4\|h\|, \quad (2.9)$$

where $\|\cdot\|$ denotes the sup-norm. If h is absolutely continuous, then

$$\|f_h\| \leq 2\|h'\|, \quad \|f'_h\| \leq \|h'\|, \quad \|f''_h\| \leq 2\|h'\|. \quad (2.10)$$

2.2. Normal approximation for smooth functions and Berry-Esseen bounds. Let $W := W_n$ be the random variable of interest. Our goal is to estimate $Eh(W) - Eh(Z)$. By (2.6), it is equivalent to estimate $Ef'_h(W) - EWf_h(W)$, which is often much easier to deal with than the original one. When W is the standardized sum of independent random variables or locally dependent random variables, the Stein method has been successfully applied to prove the uniform and non-uniform Berry-Esseen bounds (Chen and Shao, 2001, 2004). Here we focus on Berry-Esseen bounds for general W . Following Chen, Goldstein and Shao (2010), W is said to satisfy a general framework of Stein's identity if there exist a random function $\hat{K}(t)$ and a random variable R such that

$$EWf(W) = E \int_{-\infty}^{\infty} f'(W+t) \hat{K}(t) dt + ERf(W) \quad (2.11)$$

for all absolutely continuous functions f for which expectations exist. The following theorem provides the normal approximation for smooth functions.

Theorem 2.1. *Let h be absolutely continuous with $\|h'\| < \infty$ and \mathcal{F} any σ -algebra containing $\sigma(W)$. If (2.11) holds, then*

$$|Eh(W) - Eh(Z)| \leq \|h'\| (E|1 - \hat{K}_1| + 2E(\hat{K}_2) + 2E|R|), \tag{2.12}$$

where

$$\hat{K}_1 = E \left\{ \int_{-\infty}^{\infty} \hat{K}(t) dt \mid \mathcal{F} \right\} \quad \text{and} \quad \hat{K}_2 = \int_{-\infty}^{\infty} |t\hat{K}(t)| dt. \tag{2.13}$$

Proof. Let f_h be the Stein solution in (2.6). Then, by (2.11)

$$\begin{aligned} Eh(W) - Eh(Z) &= Ef'_h(W) - E \int_{-\infty}^{\infty} f'_h(W + t)\hat{K}(t)dt - E\{Rf_h(W)\} \\ &= Ef'_h(W)(1 - \hat{K}_1) - E\{Rf_h(W)\} \\ &\quad + E \int_{-\infty}^{\infty} \{f'_h(W) - f'_h(W + t)\}\hat{K}(t)dt. \end{aligned}$$

By the basic properties of the Stein solution (2.10) and the mean value theorem, we have

$$|Ef'_h(W)(1 - \hat{K}_1)| \leq \|h'\|E|1 - \hat{K}_1|, \quad |E\{Rf_h(W)\}| \leq 2\|h'\|E|R|$$

and

$$\left| E \int_{-\infty}^{\infty} \{f'_h(W) - f'_h(W + t)\}\hat{K}(t)dt \right| \leq E \int_{-\infty}^{\infty} 2\|h'\||t\hat{K}(t)|dt = 2\|h'\|E\hat{K}_2.$$

This proves (2.12). \square

Chen et al (2010) give four different approaches to construct \hat{K} in (2.11). In particular, for the exchangeable pairs approach (Stein (1986)), one constructs W' such that (W, W') is exchangeable. Suppose that there exist a constant λ ($0 < \lambda < 1$) and a random variable R such that

$$E(W - W' \mid W) = \lambda(W - R). \tag{2.14}$$

Then for all f

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

provided the expectation exists. This gives

$$\begin{aligned} EWf(W) &= \frac{1}{2\lambda}E\{(W - W')(f(W') - f(W))\} + E(Rf(W)) \\ &= E \int_{-\infty}^{\infty} f'(W + t)\hat{K}(t)dt + E(Rf(W)) \end{aligned} \tag{2.15}$$

for all absolutely continuous functions f for which expectations exist, where $\hat{K}(t) = \frac{1}{2\lambda}\Delta(I(-\Delta < t < 0) - I(0 \leq t < -\Delta))$ and $\Delta = W - W'$. Therefore with

$$\hat{K}_1 = E(\Delta^2|\mathcal{F})/(2\lambda), \quad \hat{K}_2 = |\Delta|^3/(4\lambda) \tag{2.16}$$

Theorem 2.1 leads to

Theorem 2.2. *Let h be absolutely continuous with $\|h'\| < \infty$ and \mathcal{F} any σ -algebra containing $\sigma(W)$, and let (W, W') be an exchangeable pair satisfying (2.14). Then*

$$|Eh(W) - Eh(Z)| \leq \|h'\|(E|1 - \hat{K}_1| + E|\Delta|^3/(2\lambda) + 2E|R|). \tag{2.17}$$

From the L_1 bound one can derive a Berry-Esseen bound, as highlighted below

$$\sup_z |P(W \leq z) - \Phi(z)| \leq 2 \left(\sup_{\|h'\| \leq 1} |Eh(W) - Eh(Z)| \right)^{1/2}. \tag{2.18}$$

However, the Berry-Esseen bound in (2.18) is usually not sharp.

When $\hat{K}(t)$ in (2.11) has a bounded support, Chen et al. (2010) give the following Berry-Esseen bound under the framework of (2.11).

Theorem 2.3. *Let W be any random variable and let f_z be the solution of the Stein equation (2.2) for $z \in \mathbb{R}$. Suppose that there exist random variables R_1 and $\hat{K}(t) \geq 0, t \in \mathbb{R}$, and constants δ and δ_1 not depending on z , such that $|E(R_1)| \leq \delta_1$ and*

$$EWf_z(W) = E \int_{|t| \leq \delta} f'_z(W + t)\hat{K}(t) dt + E(R_1). \tag{2.19}$$

Then

$$\sup_{z \in \mathbb{R}} |P(W \leq z) - \Phi(z)| \leq \delta(1.1 + E|W\hat{K}_1|) + 2.7E|1 - \hat{K}_1| + \delta_1, \tag{2.20}$$

where $\hat{K}_1 = E(\int_{|t| \leq \delta_0} \hat{K}(t)dt|W)$. In particular, if W, W' are mean zero, variance 1 exchangeable random variables satisfying (2.14) for some $0 < \lambda < 1$ and some random variable R , and if $|\Delta| \leq \delta$ for some constant δ , then

$$\sup_z |P(W \leq z) - \Phi(z)| \leq \delta(1.1 + E|W\hat{K}_1|) + 2.7E|1 - \hat{K}_1| + E|R|, \tag{2.21}$$

where $\hat{K}_1 = E(\Delta^2|W)/(2\lambda)$ and $\Delta = W - W'$.

Open Question 1. *Let (W, W') be an exchangeable pair satisfying (2.14). Is it true or under what additional assumptions*

$$\sup_z |P(W \leq z) - \Phi(z)| \leq A(E|1 - \hat{K}_1| + E|\Delta|^3/\lambda + E|R|)$$

for a universal constant A ?

2.3. Cramér type moderate deviations. Moderate deviations date back to Cramér (1938) who obtained expansions for tail probabilities for sums of independent random variables about the normal distribution. For i.i.d. random variables X_1, \dots, X_n with $EX_i = 0$ and $\text{Var}(X_i) = \sigma^2$ such that $Ee^{t_0|X_1|^{1/2}} < \infty$ for some $t_0 > 0$, Cramér’s result implies that

$$P(W_n > x)/(1 - \Phi(x)) \rightarrow 1$$

uniformly in $0 \leq x \leq o(n^{1/6})$ as $n \rightarrow \infty$, where $W_n = (X_1 + \dots + X_n)/(\sigma\sqrt{n})$. For random variable W satisfying (2.19), Chen, Fang and Shao (2009) apply Stein’s method to obtain the following Cramér type moderate deviation result.

Theorem 2.4. *Assume that there exist a non-negative random function $\hat{K}(t)$, a random variable R and a constant δ such that*

$$EWf(W) = E \int_{|t| \leq \delta} f'(W + t)\hat{K}(t)dt + E(Rf(W)) \tag{2.22}$$

for all absolutely continuous function f for which the expectation of either side exists. Put $\hat{K}_1 = E(\int_{|t| \leq \delta} \hat{K}(t)dt | W)$. Suppose that there exist constants δ_1, δ_2 and $\theta \geq 1$ such that

$$|\hat{K}_1 - 1| \leq \delta_1(1 + W^2), |E(R | W)| \leq \delta_2(1 + |W|), \text{ and } \hat{K}_1 \leq \theta. \tag{2.23}$$

Then

$$\frac{P(W > x)}{1 - \Phi(x)} = 1 + O(1)(\theta(1 + x^3)\delta + (1 + x^4)\delta_1 + (1 + x^2)\delta_2) \tag{2.24}$$

for $0 \leq x \leq \theta^{-1} \min(\delta^{-1/3}, \delta_1^{-1/4}, \delta_2^{-1/3})$, where $O(1)$ denotes a quantity whose absolute value is bounded by a universal constant.

Applying Theorem 2.4, Chen et al. (2009) established optimal Cramér type moderate deviations for the combinatorial central limit theorem, the anti-voter model on a complete graph, the binary expansion of a random integer, and the Curie-Weiss model. It is noted that Raic (2007) also used Stein’s method to obtain moderate deviation results for dependent random variables. However, the dependence structure considered by him is related to local dependence and is of a different nature from assumption (2.22).

2.4. Non-normal approximation via exchangeable pairs approach. Let $W := W_n$ be the random variable of interest. Since the exact distribution of W is not available for most cases, it is natural to seek the asymptotic distribution of W with a Berry-Esseen type error. Let (W, W') be an exchangeable pair satisfying

$$E(W - W'|W) = g(W) + r(W), \tag{2.25}$$

where $g(W)$ is a dominated term while $r(W)$ is a negligible term. When $g(W) = \lambda W$, and $E((W' - W)^2|W)/(2\lambda) \rightarrow 1$ in probability, Theorem 2.2 shows that the limiting distribution of W is normal under certain regularity conditions. Following the idea of the Stein's method of exchangeable pairs for normal approximation, Chatterjee and Shao (2008) are able to identify the limiting distribution of W and obtain the rate of convergence for general g . Let

$$G(t) = \int_0^t g(s)ds \text{ and } p(t) = c_1 e^{-c_0 G(t)}, \tag{2.26}$$

where $c_0 > 0$ is a constant and $c_1 = 1/\int_{-\infty}^{\infty} e^{-c_0 G(t)} dt$ is the normalizing constant so that $p(t)$ is a density function. Let Y be a random variable with the probability density function p . Assume that

- (H1) $g(t)$ is non-decreasing, and $g(t) \geq 0$ for $t > 0$ and $g(t) \leq 0$ for $t \leq 0$;
- (H2) there exists $c_2 < \infty$ such that for all x ,

$$\min(1/c_1, 1/|c_0 g(x)|)(|x| + 3/c_1) \max(1, c_0 |g'(x)|) \leq c_2.$$

Let $\Delta = W - W'$. Next result shows that W converges to Y in distribution as long as $c_0 E(\Delta^2|W)$ satisfies a law of large numbers.

Theorem 2.5. *Let h be absolutely continuous with $\|h'\| < \infty$. If (H1) and (H2) are satisfied, then*

$$\begin{aligned} & |Eh(W) - Eh(Y)| \tag{2.27} \\ & \leq \frac{(1 + c_2)\|h'\|}{c_1} \left\{ E|1 - (c_0/2)E(\Delta^2|W)| + c_0 c_1 E|\Delta|^3 + c_0 E|r(W)| \right\}. \end{aligned}$$

When Δ is bounded, Theorem 2.6 below gives a Berry-Esseen type inequality.

Theorem 2.6. *Assume that $|W - W'| \leq \delta$, where δ is a constant. Then*

$$\begin{aligned} & |P(W \leq z) - P(Y \leq z)| \\ & \leq \frac{2(1 + 2c_1)}{c_1} E|1 - (c_0/2)E(\Delta^2|W)| \\ & \quad + c_0(1 + c_1 + 1/c_1)\delta^3 E|c_0 g(W)| + c_0 c_1 c_2 \delta^3 / 2 \\ & \quad + c_1(1 + 2c_2)\delta/2 + \frac{2c_0}{c_1} E|r(W)|. \tag{2.28} \end{aligned}$$

The constant c_0 could be chosen so that $c_0 \sim 2/E(\Delta^2)$. As an application, Chatterjee and Shao (2008) obtain a Berry-Esseen type bound of order $1/\sqrt{n}$ in the non-central limit theorem for the magnetization in the Curie-Weiss ferromagnet at the critical temperature.

Open Question 2. *It would be interesting to see if a Cramér type moderate deviation holds for W satisfying (2.25) under conditions of Theorem 2.6.*

2.5. Randomized concentration inequalities. Concentration inequality approach is one of the powerful techniques for normal approximation by Stein’s method. By developing uniform and non-uniform concentration inequalities, Chen and Shao (2001, 2004, 2007) obtain optimal uniform and non-uniform Berry-Essen bounds for independent random variables, for dependent random variables under local dependence, and for non-linear statistics. Here we develop an exponential type randomized concentration inequality.

Let $\xi_i, 1 \leq i \leq n$ be independent random variables with zero means and finite second moments. Let $W = \sum_{i=1}^n \xi_i$, and $\Delta, \Delta_1, \Delta_2$ be measurable functions of $\{\xi_i, 1 \leq i \leq n\}$.

Theorem 2.7. *Assume that there exist $c_1 > c_2 > 0, \delta > 0$ such that*

$$\sum_{i=1}^n E\xi_i^2 \leq c_1 \tag{2.29}$$

and

$$\sum_{i=1}^n E|\xi_i| \min(\delta, |\xi_i|/2) \geq c_2. \tag{2.30}$$

Then for $\lambda \geq 0$

$$\begin{aligned} & Ee^{\lambda(W+\Delta)} I(\Delta_1 \leq W + \Delta \leq \Delta_2) \tag{2.31} \\ & \leq (Ee^{2\lambda(W+\Delta)})^{1/2} \exp\left(-\frac{c_2^2}{16c_1\delta^2}\right) \\ & \quad + \frac{2e^{\lambda\delta}}{c_2} \left\{ Ee^{\lambda(W+\Delta)} |W| (|\Delta_2 - \Delta_1| + 2\delta) \right. \\ & \quad + 2 \sum_{i=1}^n Ee^{\lambda(W^{(i)}+\Delta^{(i)})} |\xi_i| (|\Delta_1 - \Delta_1^{(i)}| + |\Delta_2 - \Delta_2^{(i)}|) \\ & \quad + \sum_{i=1}^n E|\Delta - \Delta^{(i)}| \min(|\xi_i|, |\Delta - \Delta^{(i)}|) (3 + \lambda(|\Delta_2 - \Delta_1| \\ & \quad \left. + 2\delta)) \max(e^{\lambda(W+\Delta)}, e^{\lambda(W^{(i)}+\Delta^{(i)})}) \right\} \end{aligned}$$

for any measurable functions $\Delta^{(i)}, \Delta_1^{(i)}, \Delta_2^{(i)}$ such that ξ_i is independent of $(W^{(i)}, \Delta^{(i)}, \Delta_1^{(i)}, \Delta_2^{(i)})$, where $W^{(i)} = W - \xi_i$.

Concentration inequality (2.31) will make it possible to obtain Cramér type moderate deviations for self-normalized sums and for Studentized statistics. Details will be given in a forthcoming paper.

Proof of Theorem 2.7. First, we assume

$$\Delta_1 \leq \Delta_2 \quad \text{and} \quad \Delta_1^{(i)} \leq \Delta_2^{(i)} \tag{2.32}$$

and show that

$$\begin{aligned} & Ee^{\lambda(W+\Delta)}I(\Delta_1 \leq W + \Delta \leq \Delta_2) \tag{2.33} \\ & \leq (Ee^{2\lambda(W+\Delta)})^{1/2} \exp\left(-\frac{c_2^2}{16c_1\delta^2}\right) \\ & \quad + \frac{2e^{\lambda\delta}}{c_2} \left\{ Ee^{\lambda(W+\Delta)}|W|(|\Delta_2 - \Delta_1| + 2\delta) \right. \\ & \quad + \sum_{i=1}^n Ee^{\lambda(W^{(i)}+\Delta^{(i)})}|\xi_i|(|\Delta_1 - \Delta_1^{(i)}| + |\Delta_2 - \Delta_2^{(i)}|) \\ & \quad + \sum_{i=1}^n E|\Delta - \Delta^{(i)}| \min(|\xi_i|, |\Delta - \Delta^{(i)}|)(3 + \lambda(\Delta_2 - \Delta_1 \\ & \quad \left. + 2\delta)) \max(e^{\lambda(W+\Delta)}, e^{\lambda(W^{(i)}+\Delta^{(i)})}) \right\}. \end{aligned}$$

For the general case, let $\Delta_1^* = \min(\Delta_1, \Delta_2)$, $\Delta_2^* = \max(\Delta_1, \Delta_2)$, $\Delta_1^{*(i)} = \min(\Delta_1^{(i)}, \Delta_2^{(i)})$, $\Delta_2^{*(i)} = \max(\Delta_1^{(i)}, \Delta_2^{(i)})$. Then (2.31) follows from (2.33) by noting that $|\Delta_2^* - \Delta_1^*| = |\Delta_2 - \Delta_1|$,

$$Ee^{\lambda(W+\Delta)}I(\Delta_1 \leq W + \Delta \leq \Delta_2) \leq Ee^{\lambda(W+\Delta)}I(\Delta_1^* \leq W + \Delta \leq \Delta_2^*)$$

and

$$\begin{aligned} & |\Delta_1^* - \Delta_1^{*(i)}| + |\Delta_2^* - \Delta_2^{*(i)}| \\ & = |\min(\Delta_1, \Delta_2) - \min(\Delta_1^{(i)}, \Delta_2^{(i)})| + |\max(\Delta_1, \Delta_2) - \max(\Delta_1^{(i)}, \Delta_2^{(i)})| \\ & \leq 2|\Delta_1 - \Delta_1^{(i)}| + 2|\Delta_2 - \Delta_2^{(i)}|. \end{aligned}$$

To prove (2.33) under (2.32), for $a < b$ define

$$f_{a,b}(w) = \begin{cases} 0 & \text{for } w \leq a - \delta, \\ e^{\lambda w}(w - a + \delta) & \text{for } a - \delta \leq w \leq b + \delta, \\ e^{\lambda w}(b - a + 2\delta) & \text{for } w > b + \delta. \end{cases} \tag{2.34}$$

Then

$$\begin{aligned}
 &EWf_{\Delta_1, \Delta_2}(W + \Delta) \\
 &= \sum_{i=1}^n E\xi_i(f_{\Delta_1, \Delta_2}(W + \Delta) - f_{\Delta_1, \Delta_2}(W^{(i)} + \Delta^{(i)})) \\
 &\quad + \sum_{i=1}^n E\xi_i(f_{\Delta_1, \Delta_2}(W^{(i)} + \Delta^{(i)}) - f_{\Delta_1^{(i)}, \Delta_2^{(i)}}(W^{(i)} + \Delta^{(i)})) \\
 &:= I_1 + I_2 \tag{2.35}
 \end{aligned}$$

by using the assumption that $E\xi_i = 0$ and that ξ_i and $(W^{(i)}, \Delta^{(i)}, \Delta_1^{(i)}, \Delta_2^{(i)})$ are independent. Clearly,

$$|f_{a,b}(w) - f_{a_1,b_1}(w)| \leq e^{\lambda w}(|a - a_1| + |b - b_1|)$$

for all $w, a < b$ and $a_1 < b_1$. We have

$$|I_2| \leq \sum_{i=1}^n Ee^{\lambda(W^{(i)} + \Delta^{(i)})} |\xi_i| (|\Delta_1 - \Delta_1^{(i)}| + |\Delta_2 - \Delta_2^{(i)}|). \tag{2.36}$$

To estimate I_1 , write $f = f_{\Delta_1, \Delta_2}$. Noting that

$$0 \leq f'(w) \leq e^{\lambda w} (1 + \lambda(\Delta_2 - \Delta_1 + 2\delta)),$$

we have

$$\begin{aligned}
 &\xi_i(f_{\Delta_1, \Delta_2}(W + \Delta) - f_{\Delta_1, \Delta_2}(W^{(i)} + \Delta^{(i)})) \\
 &= \xi_i \int_{-\xi_i + \Delta^{(i)} - \Delta}^0 f'(W + \Delta + t) dt \\
 &= \xi_i I(|\Delta - \Delta^{(i)}| \leq |\xi_i|/2) \int_{-\xi_i + \Delta^{(i)} - \Delta}^0 f'(W + \Delta + t) dt \\
 &\quad + \xi_i I(|\Delta - \Delta^{(i)}| > |\xi_i|/2) \int_{-\xi_i + \Delta^{(i)} - \Delta}^0 f'(W + \Delta + t) dt \\
 &\geq I(|\Delta - \Delta^{(i)}| \leq |\xi_i|/2) \int_{-\infty}^{\infty} f'(W + \Delta + t) \hat{K}_i(t) dt \\
 &\quad - |\Delta - \Delta^{(i)}| \min(|\xi_i|, |\Delta - \Delta^{(i)}|) (1 + \lambda(\Delta_2 - \Delta_1 + 2\delta)) \\
 &\quad \max(e^{\lambda(W + \Delta)}, e^{\lambda(W^{(i)} + \Delta^{(i)})}),
 \end{aligned}$$

where

$$\hat{K}_i(t) = \xi_i (I\{-\xi_i + \Delta^{(i)} - \Delta \leq t \leq 0\} - I\{0 < t \leq -\xi_i + \Delta^{(i)} - \Delta\}).$$

It is easy to see that when $|\Delta - \Delta^{(i)}| \leq |\xi_i|/2$, ξ_i and $I\{-\xi_i + \Delta^{(i)} - \Delta \leq t \leq 0\} - I\{0 < t \leq -\xi_i + \Delta^{(i)} - \Delta\}$ have the same sign, hence, $\hat{K}_i(t) \geq 0$ and moreover $\tilde{K}_i(t) \geq \hat{K}_i(t)$, where

$$\tilde{K}_i(t) = \xi_i(I\{-\xi_i/2 \leq t \leq 0\} - I\{0 < t \leq -\xi_i/2\}) \geq 0.$$

Therefore

$$I_1 \geq I_3 - \sum_{i=1}^n E|\Delta - \Delta^{(i)}| \min(|\xi_i|, |\Delta - \Delta^{(i)}|)(1 + \lambda(\Delta_2 - \Delta_1 + 2\delta)) \max(e^{\lambda(W+\Delta)}, e^{\lambda(W^{(i)}+\Delta^{(i)})}), \tag{2.37}$$

where

$$I_3 = \sum_{i=1}^n EI(|\Delta - \Delta^{(i)}| \leq |\xi_i|/2) \int_{-\infty}^{\infty} f'(W + \Delta + t)\tilde{K}_i(t)dt.$$

By the fact that $f' \geq 0$ and $f'(w) \geq e^{\lambda w}$ for $\Delta_1 - \delta < w < \Delta_2 + \delta$,

$$\begin{aligned} & I(|\Delta - \Delta^{(i)}| \leq |\xi_i|/2) \int_{-\infty}^{\infty} f'(W + \Delta + t)\tilde{K}_i(t)dt \\ & \geq I\{|\Delta - \Delta^{(i)}| \leq |\xi_i|/2\} \int_{|t| \leq \delta} I(\Delta_1 \leq W + \Delta \leq \Delta_2) f'(W + \Delta + t)\tilde{K}_i(t)dt \\ & \geq I\{|\Delta - \Delta^{(i)}| \leq |\xi_i|/2\} I(\Delta_1 \leq W + \Delta \leq \Delta_2) e^{\lambda(W+\Delta-\delta)} \int_{|t| \leq \delta} \tilde{K}_i(t)dt \\ & = I\{|\Delta - \Delta^{(i)}| \leq |\xi_i|/2\} I(\Delta_1 \leq W + \Delta \leq \Delta_2) e^{\lambda(W+\Delta-\delta)} |\xi_i| \min(\delta, |\xi_i|/2) \\ & \geq I(\Delta_1 \leq W + \Delta \leq \Delta_2) e^{\lambda(W+\Delta-\delta)} \eta_i - I\{|\Delta - \Delta^{(i)}| > |\xi_i|/2\} e^{\lambda(W+\Delta-\delta)} \eta_i \\ & \geq I(\Delta_1 \leq W + \Delta \leq \Delta_2) e^{\lambda(W+\Delta-\delta)} \eta_i \\ & \quad - 2|\Delta - \Delta^{(i)}| \min(|\xi_i|, |\Delta - \Delta^{(i)}|) e^{\lambda(W+\Delta-\delta)}, \end{aligned} \tag{2.38}$$

where $\eta_i = |\xi_i| \min(\delta, |\xi_i|/2)$. Hence

$$I_3 \geq I_4 - 2 \sum_{i=1}^n E|\Delta - \Delta^{(i)}| \min(|\xi_i|, |\Delta - \Delta^{(i)}|) e^{\lambda(W+\Delta-\delta)} \tag{2.39}$$

and

$$\begin{aligned}
 I_4 &= Ee^{\lambda(W+\Delta-\delta)}I(\Delta_1 \leq W + \Delta \leq \Delta_2)\sum_{i=1}^n \eta_i \\
 &\geq (c_2/2)Ee^{\lambda(W+\Delta-\delta)}I(\Delta_1 \leq W + \Delta \leq \Delta_2)I\left(\sum_{i=1}^n \eta_i \geq c_2/2\right) \\
 &\geq (c_2/2)\left\{Ee^{\lambda(W+\Delta-\delta)}I(\Delta_1 \leq W + \Delta \leq \Delta_2) \right. \\
 &\quad \left. -Ee^{\lambda(W+\Delta-\delta)}I\left(\sum_{i=1}^n \eta_i < c_2/2\right)\right\} \\
 &\geq (c_2/2)\left\{Ee^{\lambda(W+\Delta-\delta)}I(\Delta_1 \leq W + \Delta \leq \Delta_2) \right. \\
 &\quad \left. -\left(Ee^{2\lambda(W+\Delta-\delta)}\right)^{1/2}\left(P\left(\sum_{i=1}^n \eta_i < c_2/2\right)\right)^{1/2}\right\}.
 \end{aligned}$$

Note that $\eta_i, 1 \leq i \leq n$ are independent non-negative random variables, by (2.29), and (2.30) and the exponential inequality (cf. Theorem 2.19 in [28])

$$\begin{aligned}
 P\left(\sum_{i=1}^n \eta_i < c_2/2\right) &\leq \exp\left(-\frac{(c_2/2)^2}{2\sum_{i=1}^n E\eta_i^2}\right) \\
 &\leq \exp\left(-\frac{(c_2/2)^2}{2c_1\delta^2}\right) = \exp\left(-\frac{c_2^2}{8c_1\delta^2}\right).
 \end{aligned}$$

Thus

$$\begin{aligned}
 I_4 &\geq (c_2/2)\left\{Ee^{\lambda(W+\Delta-\delta)}I(\Delta_1 \leq W + \Delta \leq \Delta_2) \right. \\
 &\quad \left. -\left(Ee^{2\lambda(W+\Delta-\delta)}\right)^{1/2}\exp\left(-\frac{c_2^2}{16c_1\delta^2}\right)\right\}. \tag{2.40}
 \end{aligned}$$

Clearly, $|f| \leq e^{\lambda w}(\Delta_2 - \Delta_1 + 2\delta)$ and hence

$$EWf_{\Delta_1, \Delta_2}(W + \Delta) \leq Ee^{\lambda(W+\Delta)}|W|(\Delta_2 - \Delta_1 + 2\delta). \tag{2.41}$$

This proves (2.33) by (2.35), (2.36), (2.38), (2.39), (2.40) and (2.41).

3. Self-normalized Limit Theory

Let X, X_1, X_2, \dots be independent and identically distributed (i.i.d.) random variables. Put

$$S_n = \sum_{i=1}^n X_i \text{ and } V_n^2 = \sum_{i=1}^n X_i^2. \tag{3.1}$$

The standardized sum usually means $(S_n - a_n)/b_n$, where a_n and b_n are non-random sequences, while the self-normalized sum refers to S_n/V_n . It is well-known that moment conditions or other related conditions are necessary and sufficient for many classical limit theorems for standardized sums. However, it is now well-understood that the limit theorems for self-normalized sums usually require much less moment assumptions than those for their classical analogues. For example, the classical Chernoff (1952) large deviation

$$\lim_{n \rightarrow \infty} P\left(\frac{S_n}{n} \geq x\right)^{1/n} = \inf_{t \geq 0} e^{-tx} E e^{tX}$$

holds for $x > E(X)$ if $E e^{t_0 X} < \infty$ for some $t_0 > 0$, while the self-normalized large deviation (Shao (1997)) holds without any moment assumptions

$$\lim_{n \rightarrow \infty} P\left(S_n \geq x n^{1/2} V_n\right)^{1/n} = \sup_{b \geq 0} \inf_{t \geq 0} E e^{t(bX - x(X^2 + b^2)/2)} \quad (3.2)$$

for $x > 0$ if $E(X) = 0$ or $EX^2 = \infty$; a Cramér type moderate deviation (Shao (1999))

$$\lim_{n \rightarrow \infty} \frac{P(S_n \geq x V_n)}{1 - \Phi(x)} = 1$$

holds uniformly for $x \in [0, o(n^{1/6})]$ provided that $EX = 0$ and $E|X|^3 < \infty$, while a finite moment generating condition of $\sqrt{|X|}$ is necessary for a similar result in relation to the standard sum $S_n/\sqrt{\text{Var}(S_n)}$ (Linnik (1962), see also Petrov (1975)).

Past two decades have witnessed significant achievements on the self-normalized limit theory. Active development began in the 1990s with the seminal work of Griffin and Kuelbs (1989) on laws of the iterated logarithm for self-normalized sums of i.i.d. variables belonging to the domain of attraction of a normal or stable law. Subsequently, Bentkus and Götze (1996) derived a Berry–Esseen bound for Student’s t -statistic, and Giné, Götze and Mason (1997) proved that the t -statistic has a limiting standard normal distribution if and only if X is in the domain of attraction of a normal law. Moreover, Csörgő, Szyszkowicz and Wang (2003) proved a self-normalized version of the weak invariance principle under the same necessary and sufficient condition. Jing, Shao and Zhou (2004) derived saddlepoint approximations for Student’s t -statistic with no moment assumptions. Bercu, Gassiat and Rio (2002) obtained large and moderate deviation results for self-normalized empirical processes. Self-normalized sums of independent but non-identically distributed X_i have been considered by Bentkus, Bloznelis and Götze (1996), Wang and Jing (1999), Jing, Shao and Wang (2003). We refer to Lai and Shao (2007) for a comprehensive survey on this topic and de la Pena, Lai and Shao (2009) for systematical treatments on the theory and applications of self-normalization.

In this section, we summarize some latest developments on self-normalized limit theorems including the self-normalized saddlepoint approximation without any moment assumptions, a universal self-normalized moderate deviation, Cramér type moderate deviations for the maximum of self-normalized sums and for Studentized U-statistics. Throughout this section let X, X_1, X_2, \dots be a sequence of i.i.d. random variables unless otherwise specified. Put

$$S_n = \sum_{i=1}^n X_i, \quad V_n^2 = \sum_{i=1}^n X_i^2.$$

3.1. Self-normalized saddlepoint approximations. Let $\bar{X} = 1/n \sum_{i=1}^n X_i$ be the sample mean of $\{X_i, 1 \leq i \leq n\}$. Large deviation result provides an exponential rate of convergence for tail probability. However, a more fine-tuned approximation can be offered by saddlepoint approximations. Daniels (1954) showed that the density function of \bar{X} satisfies

$$f_{\bar{X}}(x) = e^{-n(\tau x - \kappa(\tau))} \left(\frac{n}{2\pi\kappa''(\tau)} \right)^{1/2} (1 + O(n^{-1})),$$

where $\kappa(t) = \ln Ee^{tX}$ and τ is the saddlepoint satisfying $\kappa'(\tau) = x$. Lugannani and Rice (1980) obtained the tail probability of \bar{X} :

$$P(\bar{X} \geq x) = 1 - \Phi(\sqrt{n}\hat{w}) + \frac{\phi(\sqrt{n}\hat{w})}{\sqrt{n}} \left(\frac{1}{\hat{u}} - \frac{1}{\hat{w}} + O(n^{-1}) \right),$$

where $\kappa'(\tau) = x$, $\hat{w} = \{2[\tau\kappa'(\tau) - \kappa(\tau)]\}^{1/2}\text{sign}\{\tau\}$, $\hat{u} = \tau[\kappa''(\tau)]^{1/2}$, Φ and ϕ denote the standard normal distribution function and density function, respectively. So, the error incurred by the saddlepoint approximation is $O(n^{-1})$ as against the more usual $O(n^{-1/2})$ associated with the normal approximation. Another desirable feature of saddlepoint approximation is that the approximation is quite satisfactory even when the sample size n is small. The book by Jensen (1995) gives a detailed account of saddlepoint approximations and related techniques. However, a finite moment generating function is an essential requirement for saddlepoint expansions. Daniels and Young (1991) derived saddlepoint approximations for the tail probability of the Student t-statistic under the assumption that the moment generating function of X^2 exists. Note that (3.2) holds without any moment assumption. It is natural to ask whether the saddlepoint approximation is still valid without any moment condition for the t statistic or equivalently, for the self-normalized sum S_n/V_n . Jing, Shao and Zhou (2004) give an affirmative answer to this question. Let $K(s, t) = \ln Ee^{sX+tX^2}$,

$$K_{11}(s, t) = \frac{\partial^2 K(s, t)}{\partial s^2}, \quad K_{12}(s, t) = \frac{\partial^2 K(s, t)}{\partial s \partial t}, \quad K_{22}(s, t) = \frac{\partial^2 K(s, t)}{\partial t^2}.$$

For $0 < x < 1$, let \hat{t}_0 and a_0 be solutions t and a to the equations

$$\frac{EXe^{t(-2aX/x^2+X^2)}}{Ee^{t(-2aX/x^2+X^2)}} = a, \quad \frac{EX^2e^{t(-2aX/x^2+X^2)}}{Ee^{t(-2aX/x^2+X^2)}} = \frac{a^2}{x^2}.$$

It is proved in [42] that $\hat{t}_0 < 0$. Put $\hat{s}_0 = -2a_0\hat{t}_0/x^2$ and define

$$\begin{aligned} \Lambda_0(x) &= \hat{s}_0 a_0 + \hat{t}_0 a_0^2/x^2 - K(\hat{s}_0, \hat{t}_0), \\ \Lambda_1(x) &= 2\hat{t}_0/x^2 + (1, 2a_0/x^2)\Delta^{-1}(1, 2a_0/x^2)', \end{aligned}$$

where

$$\Delta = \begin{pmatrix} K_{11}(\hat{s}_0, \hat{t}_0) & K_{12}(\hat{s}_0, \hat{t}_0) \\ K_{12}(\hat{s}_0, \hat{t}_0) & K_{22}(\hat{s}_0, \hat{t}_0) \end{pmatrix}.$$

Theorem 3.1. *Assume $EX = 0$ or $EX^2 = \infty$ and that*

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} |Ee^{isX+itX^2}|^r ds dt < \infty \tag{3.3}$$

for some $r > 1$. Then for $0 < x < 1$

$$P(S_n/V_n \geq x\sqrt{n}) = 1 - \Phi(\sqrt{nw}) - \frac{\phi(\sqrt{nw})}{\sqrt{n}} \left(\frac{1}{w} - \frac{1}{v} + O(n^{-1}) \right), \tag{3.4}$$

where $w = \sqrt{2\Lambda_0(x)}$, and $v = (-\hat{t}_0/2)^{1/2}x^{3/2}a_0^{-1}(\det \Delta)^{1/2}\Lambda_1(x)^{1/2}$.

Open Question 3. *Without assuming condition (3.3), does (3.4) hold with an error of $O(n^{-1/2})$ instead of $O(n^{-1})$? that is,*

$$P(S_n/V_n \geq x\sqrt{n}) = 1 - \Phi(\sqrt{nw}) - \frac{\phi(\sqrt{nw})}{\sqrt{n}} \left(\frac{1}{w} - \frac{1}{v} + O(n^{-1/2}) \right) \tag{3.5}$$

for fixed $0 < x < 1$? Assume that $a_n \downarrow 0$ and $a_n\sqrt{n} \rightarrow \infty$. Does (3.5) hold uniformly for $a_n \leq x \leq 1/2$?

3.2. A universal self-normalized moderate deviation. In Shao (1997) a self-normalized large deviation result without any moment assumptions (see (3.2)). It is also shown there that the tail probability of S_n/V_n is Gaussian-like when X is in the domain of attraction of the normal law and is sub-Gaussian-like when X is in the domain of attraction of a stable law. In particular, when X is symmetric and in the domain of attraction of a stable law of order α ($0 < \alpha < 2$),

$$\ln P(S_n/V_n \geq x_n) \sim -x_n^2\beta_\alpha \tag{3.6}$$

for any $x_n \rightarrow \infty$ satisfying $x_n = o(\sqrt{n})$, where β_α is the solution of

$$\int_0^\infty \frac{2 - \exp(2x - x^2/\beta) - \exp(-2x - x^2/\beta)}{x^{\alpha+1}} dx = 0.$$

Motivated by (3.2) and (3.6), Jing, Shao and Zhou (2008) establish a universal self-normalized moderate deviation for X in the centered Feller class.

Let C_s denote the support of X , that is,

$$C_s = \{x : P(X \in (x - \epsilon, x + \epsilon)) > 0, \text{ for any } \epsilon > 0\}.$$

We denote the number of elements in C_s by $\text{Card}(C_s)$ and define $\text{Card}(C_s) = \infty$ if C_s does not contain a finite number of elements. The random variable X is said to satisfy condition (H1) if

$$(H1) \quad C_s \cap R^+ \neq \emptyset \text{ and } C_s \cap R^- \neq \emptyset, \text{ where } R^+ = \{x : x > 0\}, R^- = \{x : x < 0\}$$

and satisfy condition (H2) if

$$(H2) \quad EX = 0 \text{ or } EX^2 = \infty.$$

We say $X \in \mathcal{F}_\theta$ ($0 \leq \theta < \infty$) if

$$\limsup_{a \rightarrow \infty} \frac{a^2 \left\{ P(|X| > a) + a^{-1} |EXI(|X| \leq a)| \right\}}{EX^2 I(|X| \leq a)} = \theta. \tag{3.7}$$

X is in the centered Feller class if $X \in \mathcal{F}_\theta$ for some $0 \leq \theta < \infty$.

Theorem 3.2. *Assume that X satisfies conditions (H1) and (H2). Also assume that X is in the centered Feller class. Then*

$$\ln P(S_n/V_n \geq x_n) \sim -n\lambda(x_n^2/n)$$

for any sequence $\{x_n, n \geq 1\}$ with $x_n \rightarrow \infty$ and $x_n = o(\sqrt{n})$ as $n \rightarrow \infty$, where $\lambda(x) = \inf_{b \geq 0} \sup_{t \geq 0} (tx - \ln E \exp \{t(2bX - b^2X^2)\})$. If, in addition, $\text{Card}(C_s) \geq 3$, then

$$\lim_{n \rightarrow \infty} \frac{\ln P(S_n/V_n \geq x_n)}{x_n^2} = -t_0,$$

where $t_0 = \lim_{x \rightarrow 0^+} t_x$, and (t_x, b_x) satisfy the following saddlepoint equations

$$\begin{aligned} Eb(2X - bX^2) \exp \{tb(2X - bX^2)\} &= xE \exp \{tb(2X - bX^2)\}, \\ E(X - bX^2) \exp \{tb(2X - bX^2)\} &= 0. \end{aligned}$$

It is also proved that t_0 is a positive and finite number. Theorem 3.2 together with the subsequence method is ready to give the following law of the iterated logarithm.

Theorem 3.3. *Assume that (H1), (H2) and (3.7) are satisfied. Then*

$$\limsup_{n \rightarrow \infty} \frac{S_n}{V_n \sqrt{\log \log n}} = \frac{1}{\sqrt{t_0}} \text{ a.s.}$$

3.3. Self-normalized Cramér type moderate deviations for the maximum of sums.

Let X, X_1, X_2, \dots be i.i.d. random variables with $E(X) = 0$ and $\sigma^2 = \text{Var}(X) < \infty$. Assume $E|X|^{2+r} < \infty$ for $0 < r \leq 1$. Shao (1999) proves the following self-normalized Cramér moderate deviation:

$$\frac{P(S_n \geq x V_n)}{1 - \Phi(x)} \rightarrow 1 \quad (3.8)$$

holds uniformly for $x \in [0, o(n^{r/(4+2r)})]$; Furthermore, Jing, Shao and Wang (2003) give a rate of convergence:

$$\frac{P(S_n \geq x V_n)}{1 - \Phi(x)} = 1 + O(1) \frac{(1+x)^{2+r} E|X|^{2+r}}{n^{r/2} \sigma^{2+r}} \quad (3.9)$$

for $0 \leq x \leq n^{r/(4+2r)} \sigma / (E|X|^{2+r})^{1/(2+r)}$, where $O(1)$ is bounded by an absolute constant. Similarly to the central limit theorem, it is known that for $x \geq 0$

$$P\left(\max_{1 \leq k \leq n} S_k \geq x \sigma \sqrt{n}\right) \rightarrow 2(1 - \Phi(x))$$

and

$$P\left(\max_{1 \leq k \leq n} S_k \geq x V_n\right) \rightarrow 2(1 - \Phi(x)).$$

In view of (3.8), it is natural to ask whether a similar result holds for the maximum of the self-normalized sums $\max_{1 \leq k \leq n} S_k / V_n$. Hu, Shao and Wang (2009) was the first to prove that if $EX^4 < \infty$, then

$$\lim_{n \rightarrow \infty} \frac{P(\max_{1 \leq k \leq n} S_k \geq x V_n)}{1 - \Phi(x)} = 2$$

uniformly for $x \in [0, o(n^{1/6})]$. Liu, Shao and Wang (2010) recently prove that the moment condition can be reduced to a finite third moment, an optimal assumption. More specifically, they show that

Theorem 3.4. *Let $0 < r \leq 1$. Assume that $EX = 0$ and $E|X|^{2+r} < \infty$. Then*

$$\lim_{n \rightarrow \infty} \frac{P(\max_{1 \leq k \leq n} S_k \geq x V_n)}{1 - \Phi(x)} = 2 \quad (3.10)$$

uniformly in $0 \leq x \leq o(n^{r/(4+2r)})$.

A similar result to (3.10) also holds for independent random variables under some regular conditions. In view of (3.9) and (3.10), we make the following conjectures.

Conjecture 1. Assume $EX = 0$ and $E|X|^{2+r} < \infty$ for $0 < r \leq 1$. Then

$$\frac{P(\max_{1 \leq k \leq n} S_k \geq x V_n)}{1 - \Phi(x)} = 2 + O(1) \frac{(1+x)^{2+r} E|X|^{2+r}}{n^{r/2} \sigma^{2+r}}$$

for $0 \leq x \leq n^{r/(4+2r)} \sigma / (E|X|^{2+r})^{1/(2+r)}$, where $O(1)$ is bounded by an absolute constant.

It is known that

$$\max_{1 \leq k \leq n} |S_k| / V_n \xrightarrow{d} U,$$

where U has the probability density function

$$f(x) = \frac{4}{\pi} \sum_{k=0}^{\infty} \frac{(-1)^k}{2k+1} \exp(-\pi^2(2k+1)^2/(8x^2)). \tag{3.11}$$

It would be interesting to study the moderate deviation for $\max_{1 \leq k \leq n} |S_k| / V_n$.

Conjecture 2. Assume that $EX = 0$ and $E|X|^{2+r} < \infty$ for $0 < r \leq 1$. Then

$$\frac{P(\max_{1 \leq k \leq n} |S_k| \geq x V_n)}{P(U \geq x)} = 1 + o(1)$$

uniformly in $x \in [0, o(n^{r/(4+2r)})]$, where U has the probability density function f given in (3.11).

Note that (cf. Lemma 1.6.1 in [24])

$$P(U \leq x) \sim \frac{4}{\pi} e^{-\pi^2/(8x^2)} \text{ as } x \rightarrow 0.$$

It would be also interesting to see if a self-normalized small deviation holds.

Open Question 4. Assume that $E|X|^3 < \infty$. What is the smallest possible sequence $\{a_n\}$ with $a_n \downarrow 0$ such that

$$\frac{P(\max_{1 \leq k \leq n} |S_k| \leq x V_n)}{P(U \leq x)} = 1 + o(1)$$

uniformly in $x \in (a_n, 1)$?

3.4. Studentized U-statistics. Let X, X_1, X_2, \dots, X_n be i.i.d. random variables, and let $h(x_1, x_2)$ be a real-valued symmetric Borel measurable function. Assume that $\theta = Eh(X_1, X_2)$. An unbiased estimator of θ is the Hoeffding (1948) U -statistic

$$U_n = \binom{n}{2}^{-1} \sum_{1 \leq i < j \leq n} h(X_i, X_j).$$

The U-statistic elegantly and usefully generalizes the notion of a sample mean. Typical examples include (i) sample mean: $h(x_1, x_2) = \frac{1}{2}(x_1 + x_2)$; (ii) sample variance: $h(x_1, x_2) = \frac{1}{2}(x_1 - x_2)^2$; (iii) Gini's mean difference: $h(x_1, x_2) = |x_1 - x_2|$; (iv) one-sample Wilcoxon's statistic: $h(x_1, x_2) = I(x_1 + x_2 \leq 0)$. The non-degenerate U-statistic shares many limiting properties with the sample mean. For example, if $Eh^2(X_1, X_2) < \infty$ and $\sigma_1^2 = \text{Var}(g(X_1)) > 0$, where

$$g(x) = Eh(x, X), \quad (3.12)$$

then the central limit theorem holds, i.e.,

$$\sup_x |P\left(\frac{\sqrt{n}}{2\sigma_1}(U_n - \theta) \leq x\right) - \Phi(x)| \rightarrow 0. \quad (3.13)$$

A systematic presentation of the theory of U-statistics was given in Koroljuk and Borovskikh (1994). We refer the study on uniform Berry-Esseen bound for U-statistics to Alberink and Bentkus (2001, 2002), Wang and Weber (2006) and the references there. One can also refer to Borovskikh and Weber (2003) for moderate deviations. However, since σ_1 is typically unknown, it is necessary to estimate σ_1 first and then substitute it in (3.13). Therefore, what is used in practice is actually the following studentized U-statistic (see, e.g., Arvesen (1969))

$$T_n = \sqrt{n}(U_n - \theta)/R_n, \quad (3.14)$$

where

$$R_n^2 = \frac{4(n-1)}{(n-2)^2} \sum_{i=1}^n (q_i - U_n)^2 \quad \text{with} \quad q_i = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq i}}^n h(X_i, X_j).$$

One can refer to Wang, Jing and Zhao (2000) on uniform Berry-Esseen bound for studentized U-statistics.

When $h(x_1, x_2) = (x_1 + x_2)/2$, T_n is reduced to the Student t-statistic (w.l.o.g., assuming $\theta = 0$):

$$t_n = \frac{S_n}{s_n \sqrt{n}},$$

where $s_n = \left(\frac{1}{n-1} \sum_{i=1}^n (X_i - S_n/n)^2\right)^{1/2}$ is the sample standard deviation. Note that the Student t-statistic is closely related to the self-normalized sum S_n/V_n via the following identities

$$t_n = \frac{S_n}{V_n} \left(\frac{n-1}{n - (S_n/V_n)^2} \right)^{1/2}$$

and

$$\{t_n \geq x\} = \left\{ \frac{S_n}{V_n} \geq x \left(\frac{n}{n + x^2 - 1} \right)^{1/2} \right\}.$$

Therefore, all results for the self-normalized sum S_n/V_n can be converted to the Student t-statistic t_n . In particular, the Cramér type moderate deviations (3.8) and (3.9) remain valid for t_n . Thus, it is natural to ask whether similar results hold for general studentized U-statistics. Lai, Shao and Wang (2009) recently show that the studentized U-statistics share similar properties like the student t-statistic does when the kernel satisfies

$$h^2(x_1, x_2) \leq c_0[\sigma_1^2 + g^2(x_1) + g^2(x_2)] \tag{3.15}$$

for some $c_0 > 0$. This condition is satisfied by the typical examples of U-statistics listed at the beginning of this subsection.

Theorem 3.5. *Assume $0 < \sigma_1^2 = Eg^2(X_1) < \infty$ and that (3.15) holds for some $c_0 > 0$. Then, for any x_n with $x_n \rightarrow \infty$ and $x_n = o(n^{1/2})$,*

$$\ln P(T_n \geq x_n) \sim -x_n^2/2.$$

If in addition $E|g(X_1)|^3 < \infty$, then

$$\frac{P(T_n \geq x)}{1 - \Phi(x)} = 1 + o(1) \tag{3.16}$$

holds uniformly in $x \in [0, o(n^{1/6})]$.

Open Question 5. *Does Theorem 3.5 hold without assuming condition (3.15)?*

Open Question 6. *Lai, Shao and Wang (2009) provide a rate of convergence for (3.16), but the rate seems not optimal. Is a similar result to (3.9) valid?*

Hopefully, the concentration inequality in Theorem 2.7 can be applied to answer the open question above and study Cramér type moderate deviations for Studentized statistics in general. Details will be discussed in a forthcoming paper.

4. Applications

Self-normalization arises naturally in statistics as many statistics may involve some unknown nuisance parameters which need to be estimated from the data. Studentized statistics are typical cases of self-normalization. Since self-normalized limit theorems usually require much less moment assumptions, they provides much wider applicability and also theoretical basis for various procedures commonly used in statistical inference. The self-normalized limit theorems as well as the self-normalization technique have been successfully applied to various statistical problems such as Bahadur slope (He and Shao (1996)), change point analysis (Horvath and Shao (1996)), and the performance of Monte Carlo methods for estimating ratios of normalizing constants (Chen and Shao (1997)).

In this section, we focus on recent applications to the false discovery rate (FDR) in simultaneous tests.

Following Benjamini and Hochberg (1995), consider the problem of testing simultaneously m null hypotheses H_1, H_2, \dots, H_m , of which m_0 are true. R is the number of hypotheses rejected. The table below summarizes the test results

	Declared non-significant	Declared significant	Total
True null hypotheses	U	V	m_0
Non-true null hypotheses	T	S	$m - m_0$
Total	$m - R$	R	m

The FDR is defined by $E(V/R)$. Assume P-values are p_1, p_2, \dots, p_m . Let $p_{(1)} \leq p_{(2)} \leq \dots \leq p_{(m)}$ be the ordered p -values, and denote by $H_{(i)}$ the null hypothesis corresponding to $p_{(i)}$. Let $0 < \theta < 1$ and define

$$k = \max\{i : p_{(i)} \leq i\theta/m\}.$$

The celebrated Benjamini and Hochberg (1995) method shows that the FDR is controlled at level θ if all $H_{(i)}$ are rejected for $i = 1, 2, \dots, k$ and if the tests are independent. However, the P-values are usually unknown in practice and need to be estimated. To control the FDR at level θ based on estimated P-values, the accuracy of the estimators should be of order $o(1/m)$. Fan, Hall and Yao (2007) show that if the normal distribution or the t -distribution is used for estimating the P-values for tests based on Student's t -statistic with sample size n , the level of the simultaneous test is accurate provided $\log(m) = o(n^{1/3})$. The argument in [32] (see also Korosok and Ma (2007)) leads to the following result.

Theorem 4.1. *Let $T_{n,i}$ be the test statistic for H_i . Assume that the true P-value is $p_i = P(T_{n,i} \geq t_{n,i})$ and that there exist $a_{n,i}$ and functions f_i such that*

$$\max_{1 \leq i \leq m} \sup_{x \leq a_{n,i}} \left| \frac{P(T_{n,i} \geq x)}{f_i(x)} - 1 \right| = o(1)$$

as $n \rightarrow \infty$. If $m \leq \theta / (2 \max_{1 \leq i \leq m} f_i(a_{n,i}))$, then the FDR is controlled at level θ based on estimated P-values $\hat{p}_i = f_i(t_{n,i})$.

Now consider the problem of identifying periodically expressed genes in microarray time series data. Let $Y_{t,g}$ denote the observed expression level of gene g at time t , $1 \leq g \leq m$ and $1 \leq t \leq n$, where m is the number of genes. Consider the following model of periodic gene expression

$$Y_{t,g} = \mu_g + \beta_g \cos(\omega t + \phi) + \varepsilon_{t,g},$$

where $\beta_g \geq 0$, $\omega \in (0, \pi)$, $\phi \in (-\pi, \pi]$, μ_g is the mean expression level. For each g , $\varepsilon_{1,g}, \dots, \varepsilon_{n,g}$ are i.i.d. noise sequence with mean zero. We wish to test the null hypothesis $H_{0,g} : \beta_g = 0$ against the alternative hypothesis $H_{1,g} :$

$\beta_g \neq 0$. If $H_{0,g}$ is rejected, then gene g is identified with a periodic pattern in its expression. Let $q = \lfloor (n-1)/2 \rfloor$ and set

$$I_n^{(g)}(\omega_j) = \frac{1}{n} \left| \sum_{k=1}^n Y_{k,g} e^{ik\omega_j} \right|^2,$$

where $\omega_j = 2\pi j/n$, $1 \leq j \leq q$. Define the g -statistic

$$f_{n,g} = \frac{\max_{1 \leq j \leq q} I_n^{(g)}(\omega_j)}{q^{-1} \sum_{j=1}^q I_n^{(g)}(\omega_j)} - \log q.$$

Liu and Shao (2010) show that under $H_{0,g}$,

$$\frac{P(f_{n,g} \geq y)}{1 - \exp(-\exp(-y))} = 1 \tag{4.1}$$

holds uniformly in $y \in [-\log q, o(n^{1/3})]$ provided $E\varepsilon_{1,g}^4 < \infty$. Therefore, the FDR based on the estimated p-values can be controlled at level θ as long as $m = \exp(o(n^{1/3}))$.

References

- [1] I.B. Alberink and V. Bentkus (2002). *Lyapunov type bounds for U-statistics*. Theory Probab. Appl. **46**, 571–588.
- [2] R. Arratia, L. Goldstein and L. Gordon (1990). *Poisson approximation and the Chen-Stein method*. Statist. Sci. **5**, 403–434.
- [3] J.N. Arvesen (1969). *Jackknifing U-statistics*. Ann. Math. Statist. **40**, 2076–2100.
- [4] P. Baldi, Y. Rinott and C. Stein (1989). *A normal approximations for the number of local maxima of a random function on a graph*. In Probability, Statistics and Mathematics, Papers in Honor of Samuel Karlin. T.W. Anderson, K.B. Athreya and D.L. Iglehart eds., Academic Press, 59–81.
- [5] A. Barbour and L.H.Y. Chen (2005). *An Introduction to Stein method*. Lecture Notes Series **4**, Institute for Mathematical Sciences, National University of Singapore, Singapore University Press and World Scientific.
- [6] A.D. Barbour, L. Holst and S. Janson (1992). *Poisson Approximation*. Oxford Studies in Probability 2, Clarendon Press, Oxford.
- [7] Y. Benjamini and Y. Hochberg (1995). *Controlling the false discovery rate: a practical and powerful approach to multiple testing*. J. Roy. Statist. Soc. Ser. B **57**, 289–300.
- [8] V. Bentkus and F. Götze (1996). *The Berry-Esséen bound for Student's statistic*. Ann. Probab. **24**, 491–503.
- [9] B. Bercu, E. Gassiat and E. Rio (2002). *Concentration inequalities, large and moderate deviations for self-normalized empirical processes*. Ann. Probab. **30**, 1576–1604.

-
- [10] Y.V. Borovskikh and N.C. Weber (2003). *Large deviations of U-statistics I*. Lietuvos Matematikos Rinkiny **43**, 13–37.
- [11] S. Chatterjee (2008). *A new method of normal approximation*. Ann. Prob. **36**, 1584–1610.
- [12] S. Chatterjee and Q.M. Shao (2008). *Stein's method of exchangeable pairs with application to the Curie-Weiss model*.
- [13] L.H.Y. Chen (1975). *Poisson approximation for dependent trials*. Ann. Prob. **3**, 534–545.
- [14] L.H.Y. Chen (1993). *Extending the Poisson approximation*. Science **262**, 379–380.
- [15] L.H.Y. Chen, X. Fang and Q.M. Shao (2010). *From Stein identities to moderate deviations*.
- [16] L.H.Y. Chen, L. Goldstein and Q.M. Shao (2010). *Normal Approximation by Stein's Method*.
- [17] L.H.Y. Chen and Q.M. Shao (2001). *A non-uniform Berry–Esseen bound via Stein's method*. Probab. Theory Related Fields **120**, 236–254.
- [18] L.H.Y. Chen and Q.M. Shao (2004). *Normal approximation under local dependence*. Ann. Probab. **32**, 1985–2028.
- [19] L.H.Y. Chen and Q.M. Shao (2005). *Stein's Method for Normal Approximation*. In *An Introduction to Stein's Method* (A.D. Barbour and L. H. Y. Chen, eds), Lecture Notes Series **4**, Institute for Mathematical Sciences, National University of Singapore, Singapore University Press and World Scientific, 1–59.
- [20] L.H.Y. Chen and Q.M. Shao (2007). *Normal approximation for nonlinear statistics using a concentration inequality approach*. Bernoulli **13**, 581–599.
- [21] M.H. Chen and Q.M. Shao (1997). *On Monte Carlo methods for estimating ratios of normalizing constants*. Ann. Statist. **25**, 1563–1594.
- [22] H. Chernoff (1952). *A measure of asymptotic efficiency for tests of a hypothesis based on the sum of observations*. Ann. Math. Statist. **23**, 493–507.
- [23] H. Cramér (1938). *Sur un nouveau theoreme-limite de la theorie des probabilites*. Actualites Scientifiques et Industrielles no. 736. Paris: Hermann et Cie.
- [24] M. Csörgő and P. Révész (1981). *Strong Approximations in Probability and Statistics*. Academic Press, New York.
- [25] M. Csörgő, B. Szyszkowicz and Q. Wang (2003). *Donsker's theorem for self-normalized partial sums processes*. Ann. Probab. **31**, 1228–1240.
- [26] H.E. Daniels (1954). *Saddlepoint approximations in statistics*. Ann. Math. Statist. **25**, 631–650.
- [27] H.E. Daniels and G.A. Young (1991). *Saddlepoint approximation for the studentized mean, with an application to the bootstrap*. Biometrika **78**, 169–179.
- [28] V. de la Peña, T.Z. Lai and Q.M. Shao (2009). *Self-Normalized Processes: Theory and Statistical Applications*. Springer-Verlag, New York.
- [29] A. Dembo and Y. Rinott, Y. (1996). *Some examples of normal approximations by Stein's method*. Random discrete structures. IMA Vol. Math. Appl., **76**, 25–44.

- [30] P. Diaconis (1977). *The distribution of leading digits and uniform distribution mod 1*. Ann. Probab. **5**, 72–81.
- [31] P. Diaconis and S. Holmes (2004). *Stein's method: expository lectures and applications*. IMS Lecture Notes, Vol. **46**, Hayward, CA.
- [32] J. Fan, P. Hall and Q. Yao (2007). *To how many simultaneous hypothesis tests can normal, Student's t or bootstrap calibration be applied?* J. Amer. Stat. Asso. **102**, 1282–1288.
- [33] E. Gine., F. Götze and D.M. Mason (1997). *When is the Student t -statistic asymptotically standard normal?* Ann. Probab. **25**, 1514–1531.
- [34] L. Goldstein and G. Reinert (1997). *Stein's method and the zero bias transformation with application to simple random sampling*. Ann. Appl. Probab. **7**, 935–952.
- [35] P.S. Griffin and J. Kuelbs (1989). *Self-normalized laws of the iterated logarithm*. Ann. Probab. **17**, 1571–1601.
- [36] X. He and Q.M. Shao (1996). *Bahadur efficiency and robustness of studentized score tests*. Ann. Inst. Statist. Math. **48**, 295–314.
- [37] W. Hoeffding (1948). *A class of statistics with asymptotically normal distribution*. Ann. Math. Statist. **19**, 293–325.
- [38] L. Horváth and Q.M. Shao (1996). *Limit theorem for maximum of standardized U -statistics with an application*. Ann. Statist. **24**, 2266–2279.
- [39] Z. Hu, Q.M. Shao and Q. Wang (2009). *Cramér type moderate deviations for the maximum of self-normalized sums*. Electron. J. Probab. **14**, 1181–1197.
- [40] J.L. Jensen (1995). *Saddlepoint Approximations*. Oxford University Press, New York.
- [41] B.Y. Jing, Q.M. Shao and Q.Y. Wang (2003). *Self-normalized Cramér type large deviations for independent random variables*. Ann. Probab. **31**, 2167–2215.
- [42] B.Y. Jing, Q.M. Shao and W. Zhou (2004). *Saddlepoint approximation for Student's t -statistic with no moment conditions*. Ann. Statist. **32**, 2679–2711.
- [43] B.Y. Jing, Q.M. Shao and W. Zhou (2008). *Towards a universal self-normalized moderate deviation*. Trans. Amer. Math. Soc. **360**, 4263–4285.
- [44] V.S. Koroljuk and Yu. V. Borovskich (1994). *Theory of U -statistics*. Kluwer Academic Publishers, Dordrecht.
- [45] M.R. Korosok and S. Ma (2007). *Marginal asymptotics for the “large p , small n ” paradigm: With applications to micorarray data*. Ann. Statist. **35**, 1456–1486.
- [46] T.L. Lai, Q.M. Shao (2007). *Self-normalized limit theorems in probability and statistics*. In: Asymptotic Theory in Probability and Statistics with Applications (Editors: T.L. Lai, L.F. Qian and Q.M. Shao), pp. 3–43, Higher Education Press of China, and International Press.
- [47] T.L. Lai, Q.M. Shao and Q.Y. Wang (2009). *Cramér type large deviations for studentized U -statistics*. ESAIM P&S (to appear).
- [48] Yu. V. Linnik (1962). *Limit theorems for sums of independent random variables, taking account of large deviations*. Theory Probab. Appl. **7**, 121–134.

-
- [49] W. Liu and Q.M. Shao (2010). *Cramér type moderate deviation for the maximum of the periodogram with application to simultaneous tests in gene expression time series*. Ann. Statist. (to appear)
- [50] W. Liu, Q.M. Shao and Q. Wang (2010). *Self-normalized Cramér type Moderate Deviations for the Maximum of Sums*.
- [51] I. Nourdin and G. Peccati (2009). *Stein's method on Weiner Chaos*. Probab. Th. Related Fields **145**, 75C118.
- [52] V.V. Petrov (1975). *Sums of Independent Random Variables*. Springer-Verlag, New York.
- [53] M. Raic (2007). *CLT-related large deviation bounds based on Stein's method*. Adv. Appl. Probab. **39**, 731–752.
- [54] Q.M. Shao (1997). *Self-normalized large deviations*. Ann. Probab. **25** (1997), 285–328.
- [55] Q.M. Shao (1999). *Cramér-type large deviation for Student's t statistic*. J. Theoret. Probab. **12**, 387–398 (1999).
- [56] C. Stein (1972). *A bound for the error in the normal approximation to the distribution of a sum of dependent random variables*. Proc. Sixth Berkeley Symp. Math. Stat. Prob. **2**, 583–602, Univ. California Press. Berkeley, Calif.
- [57] C. Stein (1986). *Approximation Computation of Expectations*. Lecture Notes 7, Inst. Math. Statist., Hayward, Calif.
- [58] Q. Wang and B.Y. Jing (1999). *An exponential non-uniform Berry-Esseen bound for self-normalized sums*. Ann. Probab. **27**, 2068–2088.
- [59] Q. Wang, B.Y. Jing and L. Zhao (2000). *The Berry-Esséen bound for studentized statistics*. Ann. Probab. **28**, 511–535.
- [60] Q. Wang and N.C. Weber (2006). *Exact convergence rate and leading term in the central limit theorem for U -statistics*. Statist. Sinica **16** , 1409–1422.

ℓ_1 -regularization in High-dimensional Statistical Models

Sara van de Geer*

Abstract

Least squares with ℓ_1 -penalty, also known as the Lasso [23], refers to the minimization problem

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \|\beta\|_1 \},$$

where $\mathbf{Y} \in \mathbb{R}^n$ is a given n -vector, and \mathbf{X} is a given $(n \times p)$ -matrix. Moreover, $\lambda > 0$ is a tuning parameter, larger values inducing more regularization. Of special interest is the high-dimensional case, which is the case where $p \gg n$. The Lasso is a very useful tool for obtaining good predictions $\mathbf{X}\hat{\beta}$ of the regression function, i.e., of mean $\mathbf{f}^0 := \mathbf{E}\mathbf{Y}$ of \mathbf{Y} when \mathbf{X} is given. In literature, this is formalized in terms of an oracle inequality, which says that the Lasso predicts almost as well as the ℓ_0 -penalized approximation of \mathbf{f}^0 . We will discuss the conditions for such a result, and extend it to general loss functions. For the selection of variables however, the Lasso needs very strong conditions on the Gram matrix $\mathbf{X}^T\mathbf{X}/n$. These can be avoided by applying a two-stage procedure. We will show this for the adaptive Lasso. Finally, we discuss a modification that takes into account a group structure in the variables, where both the number of groups as well as the group sizes are large.

Mathematics Subject Classification (2010). Primary 62G05; Secondary 62J07.

Keywords. High-dimensional model, ℓ_1 -penalty, oracle inequality, restricted eigenvalue, sparsity, variable selection

1. Introduction

Estimation with ℓ_1 -penalty, also known as the Lasso [23], is a popular tool for prediction, estimation and variable selection in high-dimensional regression

*Seminar for Statistics, ETH Zürich, Rämistrasse 101, 8092 Zürich, Switzerland.
E-mail: geer@stat.math.ethz.ch.

problems. It is frequently used in the linear model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon,$$

where \mathbf{Y} is an n -vector of observations, $\mathbf{X} = (\mathbf{X}_1, \dots, \mathbf{X}_p)$ is the $(n \times p)$ -design matrix and ϵ is a noise vector. For the case of least squares error loss, the Lasso is then

$$\hat{\beta} := \arg \min_{\beta \in \mathbb{R}^p} \{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \|\beta\|_1 \}, \quad (1)$$

where $\lambda > 0$ is a tuning parameter.

A vector β is called *sparse* if it has only a few non-zero entries. *Oracle inequalities* are results of the form: with high probability

$$\|\mathbf{X}(\hat{\beta} - \beta_0)\|_2^2/n \leq \text{constant} \times \lambda^2 s_0, \quad (2)$$

where β_0 is the unknown true regression coefficient, or some sparse approximation thereof, and s_0 is the sparsity index, i.e., the number of non-zero coefficients of β_0 .

The terminology *oracle inequality* is based on the idea of mimicking an oracle that knows beforehand which coefficients β_0 are non-zero. Indeed, suppose that $\mathbf{E}\mathbf{Y} = \mathbf{X}\beta_0$, and that the noise $\epsilon = \mathbf{Y} - \mathbf{X}\beta_0$ has independent components with variance σ^2 . Let $S_0 := \{j : \beta_{j,0} \neq 0\}$, say $S_0 = \{1, \dots, s_0\}$ is the set of indices of the first s_0 variables. Let $\mathbf{X}(S_0) := \{\mathbf{X}_1, \dots, \mathbf{X}_{s_0}\}$ be the design matrix containing these first s_0 variables, and let $\beta_0(S_0)$ be the s_0 non-zero entries of β_0 . Suppose that $\mathbf{X}(S_0)$ has full rank s_0 ($s_0 \leq n$). If S_0 were known, we can apply the least squares estimator based on the variables in S_0

$$\hat{\beta}(S_0) := \left(\mathbf{X}^T(S_0)\mathbf{X}(S_0) \right)^{-1} \mathbf{X}^T(S_0)\mathbf{Y}.$$

From standard least squares theory, we have

$$\mathbb{E} \|\mathbf{X}(S_0)(\hat{\beta}(S_0) - \beta_0(S_0))\|_2^2 = \sigma^2 s_0.$$

Under general conditions, the prediction error of the Lasso behaves as if it knew S_0 , e.g., for i.i.d. centered Gaussian errors with variance σ^2 , the inequality (2) holds with large probability, with λ^2 up to a logarithmic factor $\log p$, of order σ^2/n .

In fact, what we will show in Section 2, is an oracle inequality of the form (2), where β_0 is not necessarily the “true” β , but may be a sparse approximation of the truth. The “optimal” sparse approximation will be called the *oracle*. To make the distinction, we denote the truth (if there is any) as β_{true} , and the oracle by β_{oracle} . As we will see, β_{oracle} will be at least as sparse as β_{truth} , and is possibly much sparser.

Apart from oracle inequalities, one may also consider estimation results, which are bounds on the ℓ_q error $\|\hat{\beta} - \beta_0\|_q$, for some $1 \leq q \leq \infty$. Variable selection refers to estimating the support S_0 of β_0 .

From a numerical point of view, the Lasso is attractive as it is easy to compute and the ℓ_1 -penalty ensures that a number of the estimated coefficients $\hat{\beta}_j$ are exactly zero. Its active set $\hat{S} := \{j : \hat{\beta}_j \neq 0\}$ will generally contain less than n variables, even when originally dealing with $p \gg n$ variables. In theory however, there is in general no guarantee that \hat{S} coincides with S_0 . Indeed, this would be too good to be true, because then we would have a very accurate procedure that in addition can correctly assess its own accuracy. This is somehow in contradiction with statistical uncertainty principles.

What is so special about the ℓ_1 -penalty? The theoretically ideal penalty (at least, in the linear model) for sparse situations is actually the ℓ_0 -penalty $\lambda \|\beta\|_0^0$, where $\|\beta\|_0^0 := \sum_{j=1}^p |\beta_j|^0 = \#\{\beta_j \neq 0\}$. But with this, the minimization problem is computationally intractable. The ℓ_1 -penalty has the advantage of being convex. Minimization with ℓ_1 -penalty can be done using e.g. interior point methods or path-following algorithms. Convexity is important from the computational point of view (as well as from the theoretical point of view as soon as we leave the linear model context). For theoretical analysis, it is important that the ℓ_1 -penalty satisfies the *triangle inequality*

$$\|\beta + \tilde{\beta}\|_1 \leq \|\beta\|_1 + \|\tilde{\beta}\|_1,$$

and is *separable*:

$$\|\beta\|_1 = \|\beta_S\|_1 + \|\beta_{S^c}\|_1,$$

for any $S \subset \{1, \dots, p\}$. Here β_S denotes the vector β with the entries in S^c set to zero, and $\beta_{S^c} = \beta - \beta_S$ has the entries in S set to zero. Note for example that among the ℓ_q -penalties $\lambda \|\beta\|_q^q$ (or $\lambda \|\beta\|_q$, $q \geq 1$), the ℓ_1 -penalty is the only one which unites these three properties.

There has been an explosion of papers on the topic. The theoretical properties - and limitations - of the standard Lasso are by now quite well understood. We mention some of the key papers. Consistency was obtained in [9]. Its prediction error and estimation error is derived in [12], [13] and [1], where also the so-called *restricted eigenvalue conditions* are introduced. The slightly weaker *compatibility condition* is given in [25]. In [8] an alternative to the Lasso is introduced, which is called the Dantzig selector. The papers [3], [4] and [5] also present oracle and estimation bounds, and treat *incoherence assumptions*.

Variable selection with the Lasso is studied in [21] and [32], [16] presents conditions for convergence sup-norm, and [31] for convergence in ℓ_q , $1 \leq q \leq \infty$. Modifications of the Lasso procedure have also been developed, for example, the group Lasso [30], the fused Lasso [24], and the elastic net [34]. Moreover, two-stage procedures have been proposed and studied, such as the adaptive Lasso [33, 10], and the relaxed Lasso [20]. Extension to density estimation is in [6], and to generalized-linear models in [15] (for the case of orthonormal design) and [26].

The present paper puts some of our theoretical results in a single framework. This will reveal the common aspects of various versions of the Lasso (and some links with decoding). We will mainly refer to own work, but stress here that

this work in turn builds upon results and ideas from literature. In Section 2, we present an oracle inequality in the context of the linear model. This is extended to general convex loss in Section 3. Section 4 discusses the restricted eigenvalue condition and the related compatibility condition. We turn to estimation results and variable selection in Section 5. First, we give a bound for the ℓ_2 -error (Subsection 5.1). We then show in Subsection 5.2 that the Lasso needs strong conditions for correctly estimating the support set of the coefficients. We show in Subsection 5.3 that the adaptive Lasso has a limited number of false positive selections but may have less good prediction error than the Lasso. In Section 6, we consider an extension, where the variables are divided into groups, with within each group a certain ordering of the coefficients. We provide an oracle inequality involving sparsity in the number of groups. Section 7 concludes.

2. An Oracle Inequality in the Linear Model

In this section, we present a version of the oracle inequality, which is along the lines of results in [25].

Suppose that the observations \mathbf{Y} are of the form

$$\mathbf{Y} = \mathbf{f}^0 + \epsilon,$$

where \mathbf{f}^0 is some unknown vector in \mathbb{R}^n , and ϵ is a noise vector. Let $\mathbf{X} = \{\mathbf{X}_1, \dots, \mathbf{X}_p\}$ be the design matrix. We assume that \mathbf{X} is normalized, i.e., that

$$\hat{\sigma}_{j,j} = 1, \quad \forall j,$$

where $\{\hat{\sigma}_{j,j}\}$ are the diagonal elements of the Gram matrix

$$\hat{\Sigma} := \mathbf{X}^T \mathbf{X} / n := (\hat{\sigma}_{j,k}).$$

The empirical correlation between the noise ϵ and the j -th variable \mathbf{X}_j is controlled by introducing the set

$$\mathcal{T}(\lambda) := \left\{ \max_{1 \leq j \leq p} 4|\epsilon^T \mathbf{X}_j| / n \leq \lambda \right\}.$$

The tuning parameter λ is to be chosen in such a way that the probability of $\mathcal{T}(\lambda)$ is large.

For any index set $S \subset \{1, \dots, p\}$, and any $\beta \in \mathbf{R}^p$, we let

$$\beta_{j,S} := \beta_j \mathbf{1}\{j \in S\}, \quad j = 1, \dots, p.$$

We sometimes identify $\beta_S \in \mathbb{R}^p$ with the vector in $\mathbb{R}^{|S|}$ containing only the entries in S .

We write the projection of \mathbf{f}^0 on the space spanned by the variables $\{\mathbf{X}_j\}_{j \in S}$ as

$$\mathbf{f}_S := \mathbf{X} \mathbf{b}^S := \arg \min_{f = \mathbf{X} \beta_S} \|f - \mathbf{f}^0\|_2^2.$$

When $p > n$, the Gram matrix $\hat{\Sigma}$ is obviously singular: it has at least $p - n$ eigenvalues equal to zero. We do however need some kind of compatibility of norms, namely the ℓ_1 -norm $\|\beta_S\|_1$ should be compatible with $\|\mathbf{X}\beta\|_2$. Observe that $\|\mathbf{X}\beta\|_2^2/n = \beta^T \hat{\Sigma} \beta$.

Definition compatibility condition Let $L > 0$ be a given constant and S be an index set. We say that the (L, S) -compatibility condition holds if

$$\phi_{\text{comp}}^2(L, S) := \min \left\{ \frac{|S| \beta^T \hat{\Sigma} \beta}{\|\beta_S\|_1^2} : \|\beta_{S^c}\|_1 \leq L \|\beta_S\|_1 \right\} > 0.$$

Section 4 will briefly discuss this condition.

Theorem 2.1. Let $\hat{f} = \mathbf{X}\hat{\beta}$, where $\hat{\beta}$ is the Lasso estimator defined in (1). Then on $\mathcal{T}(\lambda)$, and for all S , it holds that

$$\|\hat{f} - \mathbf{f}^0\|_2^2/n + \lambda \|\hat{\beta} - b_S\|_1 \leq 7 \|\mathbf{f}_S - \mathbf{f}^0\|_2^2/n + \frac{(7\lambda)^2 |S|}{\phi_{\text{comp}}^2(6, S)}. \tag{3}$$

The constants in the above theorem can be refined. We have chosen some explicit values for definiteness. Moreover, the idea is to apply the result to sets S with $\phi_{\text{comp}}(6, S)$ not too small (say bounded from below by a constant not depending on n or p , if possible).

Assuming that $\mathbf{f}^0 := \mathbf{X}\beta_{\text{true}}$ is linear, the above theorem tells us that

$$\|\hat{f} - \mathbf{f}^0\|_2^2/n + \lambda \|\hat{\beta} - \beta_{\text{true}}\|_1 \leq \frac{(7\lambda)^2 |S_{\text{true}}|}{\phi_{\text{comp}}^2(6, S_{\text{true}})}, \tag{4}$$

where $S_{\text{true}} := \{j : \beta_{j,\text{true}} \neq 0\}$. This is an inequality of the form (2), with β_0 taken to be β_{true} . We admit that the constant $\phi_{\text{comp}}^2(6, S_{\text{true}})$ is hiding in the unspecified ‘‘constant’’ of (2). The improvement which replaces β_{true} by a sparse approximation is based on the *oracle* set

$$S_{\text{oracle}} := \arg \min_S \left\{ \|\mathbf{f}_S - \mathbf{f}^0\|_2^2/n + \frac{7\lambda^2 |S|}{\phi_{\text{comp}}^2(6, S)} \right\}, \tag{5}$$

and the oracle predictor

$$f_{\text{oracle}} := \mathbf{f}_{S_{\text{oracle}}} = \mathbf{X}\beta_{\text{oracle}},$$

where

$$\beta_{\text{oracle}} := b^{S_{\text{oracle}}}.$$

By the above theorem

$$\|\hat{f} - \mathbf{f}^0\|_2^2/n + \lambda \|\hat{\beta} - \beta_{\text{oracle}}\|_1 \leq 7 \|f_{\text{oracle}} - \mathbf{f}^0\|_2^2/n + \frac{(7\lambda)^2 |S_{\text{oracle}}|}{\phi_{\text{comp}}^2(6, S_{\text{oracle}})},$$

which is a - possibly substantial - improvement of (4). We think of this oracle as the ℓ_0 -penalized sparse approximation of the truth. Nevertheless, the constant $\phi_{\text{comp}}(6, S_{\text{oracle}})$ can still be quite small and spoil this interpretation.

We end this section with a simple bound for the probability of the set $\mathcal{T}(\lambda)$ for the case of normally distributed errors. It is clear that appropriate probability inequalities can also be derived for other distributions. A good common practice is not to rely on distributional assumptions, and to choose the tuning parameter λ using cross-validation.

Lemma 2.1. *Suppose that ϵ is $\mathcal{N}(0, \sigma^2 I)$ -distributed. Then we have for all $x > 0$, and for*

$$\lambda := 4\sigma \sqrt{\frac{2x + 2 \log p}{n}},$$

$$\mathbb{P}\left(\mathcal{T}(\lambda)\right) \geq 1 - 2 \exp[-x].$$

3. An Oracle Inequality for General Convex Loss

As in [25, 26] one can extend the framework for squared error loss with fixed design to the following scenario. Consider data $\{Z_i\}_{i=1}^n \subset \mathcal{Z}$, where \mathcal{Z} is some measurable space. We denote, for a function $g : \mathcal{Z} \rightarrow \mathbb{R}$, the empirical average by

$$P_n g := \sum_{i=1}^n g(Z_i)/n,$$

and the theoretical mean by

$$P g := \sum_{i=1}^n \mathbb{E} g(Z_i)/n.$$

Thus, P_n is the “empirical” measure, that puts mass $1/n$ at each observation Z_i ($i = 1, \dots, n$), and P is the “theoretical” measure.

Let \mathbf{F} be a (rich) parameter space of real-valued functions on \mathcal{Z} , and, for each $f \in \mathbf{F}$, $\rho_f : \mathcal{Z} \rightarrow \mathbb{R}$ be a loss function. We assume that the map $f \mapsto \rho_f$ is convex. For example, in a density estimation problem, one can consider the loss

$$\rho_f(\cdot) := -f(\cdot) + \log \int e^f d\mu,$$

where μ is a given dominating measure. In a regression setup, one has (for $i = 1, \dots, n$) response variables $Y_i \in \mathcal{Y} \subset \mathbb{R}$ and co-variables $X_i \in \mathcal{X}$ i.e., $Z_i = (X_i, Y_i)$. The parameter f is a regression function. Examples are quadratic loss

$$\rho_f(\cdot, y) = (y - f(\cdot))^2,$$

or logistic loss

$$\rho_f(\cdot, y) = -yf(\cdot) + \log(1 + \exp[f(\cdot)]),$$

etc.

The empirical risk, and theoretical risk, at f , is defined as $P_n\rho_f$, and $P\rho_f$, respectively. We furthermore define the *target* - or *truth* - as the minimizer of the theoretical risk

$$f^0 := \arg \min_{f \in \mathbf{F}} P\rho_f.$$

Consider a linear subspace

$$\mathcal{F} := \left\{ f_\beta(\cdot) = \sum_{j=1}^p \beta_j \psi_j(\cdot) : \beta \in \mathbf{R}^p \right\} \subset \mathbf{F}.$$

Here, $\{\psi_j\}_{j=1}^p$ is a collection of functions on \mathcal{Z} , often referred to as the dictionary. The Lasso is

$$\hat{\beta} = \arg \min_{\beta} \{P_n\rho_{f_\beta} + \lambda\|\beta\|_1\}. \tag{6}$$

We write $\hat{f} = f_{\hat{\beta}}$.

For $f \in \mathbf{F}$, the excess risk is

$$\mathcal{E}(f) := P(\rho_f - \rho_{f^0}).$$

Note that by definition, $\mathcal{E}(f) \geq 0$ for all $f \in \mathbf{F}$.

Before presenting an oracle result of the same spirit as for the linear model, we need three definitions, and in addition some further notation. Let the parameter space $\mathbf{F} := (\mathbf{F}, \|\cdot\|)$ be a normed space. Recall the notation

$$\beta_{j,S} := \beta_j \mathbf{1}\{j \in S\}, \quad j = 1, \dots, p.$$

Our first definition is as in the previous section, but now with a general norm $\|\cdot\|$.

Definition compatibility condition *We say that the (L, S) -compatibility condition is met if*

$$\phi_{\text{comp}}^2(L, S) := \min \left\{ \frac{|S|\|f_\beta\|^2}{\|\beta_S\|_1^2} : \|\beta_{S^c}\|_1 \leq L\|\beta_S\|_1 \right\} > 0.$$

Definition margin condition *Let $\mathbf{F}_{\text{local}} \subset \mathbf{F}$ be some “local neighborhood” of f^0 . We say that the margin condition holds with strictly convex function G , if for all $f \in \mathbf{F}_{\text{local}}$, we have*

$$\mathcal{E}(f) \geq G(\|f - f^0\|).$$

Definition convex conjugate Let G be a strictly convex function on $[0, \infty)$, with $G(0) = 0$. The convex conjugate H of G is defined as

$$H(v) = \sup_u \{uv - G(u)\}, \quad v \geq 0.$$

The best approximation of f^0 using only the variables in S is

$$f_S := f_{b^S} := \arg \min_{f=f_{\beta_S}} \mathcal{E}(f).$$

The function f_S plays here the same role as the projection f_S of the previous section.

For H being the convex conjugate of the function G appearing in the margin condition, set

$$2\varepsilon(\lambda, S) = 3\mathcal{E}(f_S) + 2H\left(\frac{4\lambda\sqrt{|S|}}{\phi_{\text{comp}}(3, S)}\right). \tag{7}$$

For any $M > 0$, we let $\mathbf{Z}_M(S)$ be given by

$$\mathbf{Z}_M(S) := \sup_{\beta: \|\beta - b^S\|_1 \leq M} \left| (P_n - P)(\rho_{f_\beta} - \rho_{f_S}) \right|. \tag{8}$$

Theorem 3.1. Suppose that S is an index set for which $f_\beta \in \mathbf{F}_{\text{local}}$ for all $\|\beta - b^S\|_1 \leq M(\lambda, S)$, where $M(\lambda, S) := \varepsilon(\lambda, S)/(16\lambda)$. Then on the set

$$\mathcal{T}(\lambda, S) := \{ \mathbf{Z}_{M(\lambda, S)}(S) \leq \lambda M(\lambda, S)/8 \},$$

we have

$$\mathcal{E}(\hat{f}) + \lambda \|\hat{\beta} - b^S\|_1 \leq 4\varepsilon(\lambda, S).$$

The typical case is the case of quadratic margin function G , say $G(u) = u^2/2$. Then also the convex conjugate $H(v) = v^2/2$ is quadratic. This shows that Theorem 3.1 is in fact an extension of Theorem 2.1, albeit that the constants do not carry over exactly (the latter due to human inconsistencies). We furthermore remark that - in contrast to the ℓ_0 -penalty - the ℓ_1 -penalty adapts to the margin behavior. In other words, having left the framework of a linear model, the ℓ_1 -penalty exhibits an important theoretical advantage.

One may object that by assuming one is on the set $\mathcal{T}(\lambda, S)$, Theorem 3.1 neglects all difficulties coming from the random nature of our statistical problem. However, contraction and concentration inequalities actually make it possible to derive bounds for the probability of $\mathcal{T}(\lambda, S)$ in a rather elegant way. Indeed, in the case of Lipschitz loss, one may invoke the contraction inequality of [14], which gives the following lemma.

Lemma 3.1. Suppose that $f \mapsto \rho_f$ is Lipschitz:

$$|\rho_f - \rho_{\tilde{f}}| \leq |f - \tilde{f}|.$$

Then one has

$$\mathbf{E}\mathbf{Z}_M(S) \leq 4\lambda_{\text{noise}}M,$$

where

$$\lambda_{\text{noise}} := \mathbf{E} \left(\max_{1 \leq j \leq p} \left| \sum_{i=1}^n \varepsilon_i \psi_j(Z_i) / n \right| \right),$$

and where $\varepsilon_1, \dots, \varepsilon_n$ is a Rademacher sequence independent of Z_1, \dots, Z_n .

Concentration inequalities [17, 18, 2], which say that $\mathbf{Z}_M(S)$ is with large probability concentrated near its expectation, will then allow one to show that for appropriate λ , the set $\mathcal{T}(\lambda, S)$ has large probability.

4. Compatibility and Restricted Eigenvalues

Let Q be a probability measure on \mathcal{Z} , and for $\beta \in \mathbb{R}^p$, let $f_\beta = \sum_{j=1}^p \beta_j \psi_j$, where $\{\psi_j\}_{j=1}^p \subset L_2(Q)$ is a given dictionary. Write the Gram matrix as

$$\Sigma := \int \psi^T \psi dQ, \quad \psi := (\psi_1, \dots, \psi_p).$$

Moreover, let $\|\cdot\|$ be the $L_2(Q)$ -norm induced by the inner product

$$(f, \tilde{f}) := \int f \tilde{f} dQ.$$

Note thus that

$$\|f_\beta\|^2 = \beta^T \Sigma \beta.$$

Definition compatibility and restricted eigenvalue Let $L > 0$ be a given constant and S be an index set. We say that the (Σ, L, S) -compatibility condition holds if

$$\phi_{\text{comp}}^2(\Sigma, L, S) := \min \left\{ \frac{|S| \|f_\beta\|^2}{\|\beta_S\|_1^2} : \|\beta_{S^c}\|_1 \leq L \|\beta_S\|_1 \right\}$$

is strictly positive. We say that the (Σ, L, S) -restricted eigenvalue condition holds if the restricted eigenvalue

$$\phi_{\text{RE}}^2(\Sigma, L, S) := \min \left\{ \frac{\|f_\beta\|^2}{\|\beta_S\|_2^2} : \|\beta_{S^c}\|_1 \leq L \|\beta_S\|_1 \right\}$$

is strictly positive.

The compatibility condition was introduced in [25], and the restricted eigenvalue condition in [1]. It is clear that

$$\phi_{\text{RE}}^2(\Sigma, L, S) \leq \phi_{\text{comp}}^2(\Sigma, L, S).$$

On the other hand, results involving the set S_{true} , for the ℓ_2 -error $\|\hat{\beta} - \beta_{\text{true}}\|_2$ of the Lasso rely on $\phi_{\text{RE}}(\Sigma, L, S_{\text{true}})$ rather than $\phi_{\text{comp}}(\Sigma, L, S_{\text{true}})$ (and improved results, involving the oracle set S_{oracle} , in fact depend on the so-called *adaptive* restricted eigenvalue $\phi_{\text{adap}}(\Sigma, L, S_{\text{oracle}})$, see Subsection 5.1).

It is easy to see that

$$\phi_{\text{RE}}^2(\Sigma, L, S) \leq \Lambda_{\min}^2(S),$$

where $\Lambda_{\min}^2(S)$ is the smallest eigenvalue of the Gram matrix corresponding to the variables in S , i.e.,

$$\Lambda_{\min}^2(S) := \min_{\beta} \frac{\|f_{\beta_S}\|^2}{\|\beta_S\|_2^2}.$$

Conversely, denoting the canonical correlation by

$$\theta(S) := \sup_{\beta} \frac{|(f_{\beta_S}, f_{\beta_S^c})|}{\|f_{\beta_S}\| \|f_{\beta_S^c}\|},$$

one has the following bound.

Lemma 4.1. *Suppose that $\theta(S) < 1$. Then*

$$\phi_{\text{RE}}^2(\Sigma, L, S) \geq (1 - \theta(S))^2 \Lambda_{\min}^2(S).$$

Lemma 4.1 does not exploit the fact that in the definition of the restricted eigenvalue, we restrict the coefficients β to $\|\beta_{S^c}\|_1 \leq L \|\beta_S\|_1$. Using this restriction, the restricted eigenvalue condition can for instance be derived from the restricted isometry property introduced in [7]. The latter paper studies the exact recovery of the true coefficients β_{true} of $f^0 := f_{\beta_{\text{true}}}$, using the linear program

$$\beta_{\text{LP}} := \arg \min \{ \|\beta\|_1 : \|f_{\beta} - f^0\| = 0 \}. \tag{9}$$

The restrictions on the coefficients also allows one to derive bounds for restricted eigenvalues based on those computed with respect to an approximating (potentially non-singular) matrix. For two symmetric $(p \times p)$ -matrices Σ_0 and Σ_1 , we define

$$\|\Sigma_0 - \Sigma_1\|_{\infty} := \max_{1 \leq j \leq k \leq p} |\Sigma_{0,j,k} - \Sigma_{1,j,k}|.$$

The following lemma is proved in [28].

Lemma 4.2. *We have*

$$\phi_{\text{comp}}(\Sigma_1, L, S) \geq \phi_{\text{comp}}(\Sigma_0, L, S) - (L + 1) \sqrt{\|\Sigma_0 - \Sigma_1\|_{\infty} |S|}.$$

Similarly,

$$\phi_{\text{RE}}(\Sigma_1, L, S) \geq \phi_{\text{RE}}(\Sigma_0, L, S) - (L + 1) \sqrt{\|\Sigma_0 - \Sigma_1\|_{\infty} |S|}.$$

5. Estimation and Variable Selection

We present results for the linear model only.

5.1. Estimation.

Consider the model

$$\mathbf{Y} = \mathbf{f}^0 + \epsilon.$$

For estimation in ℓ_2 of the coefficients, we introduce the *adaptive* restricted eigenvalue. For a given S , our adaptive restricted eigenvalue conditions are stronger than in [1], but the result we give is also stronger, as we consider $S_{\text{oracle}} \subset S_{\text{true}}$ instead of S_{true} .

Definition adaptive restricted eigenvalue *We say that the (L, S) -adaptive restricted eigenvalue condition holds if*

$$\phi_{\text{adap}}^2(L, S) := \min \left\{ \frac{\|\mathbf{X}\beta\|_2^2}{n\|\beta_S\|_2^2} : \|\beta_{S^c}\|_1 \leq L\sqrt{|S|}\|\beta_S\|_2 \right\} > 0.$$

Thus,

$$\phi_{\text{adap}}^2(L, S) \leq \phi_{\text{RE}}^2(L, S) \leq \phi_{\text{comp}}^2(L, S).$$

In addition, we consider supersets \mathcal{N} of S , with size $(1 + \text{constant}) \times |S|$. For definiteness, we take the constant to be equal to 1. The minimal adaptive restricted eigenvalue is

$$\phi_{\text{adap}}(L, S, 2|S|) := \min\{\phi_{\text{adap}}(L, \mathcal{N}) : \mathcal{N} \supset S, |\mathcal{N}| = 2|S|\}.$$

Lemma 5.1. *Let $\hat{\beta}$ be the Lasso given in (1). Let*

$$\mathcal{T}(\lambda) := \left\{ \max_{1 \leq j \leq p} 4|\epsilon^T \mathbf{X}_j|/n \leq \lambda \right\}.$$

Then on $\mathcal{T}(\lambda)$, and for $\beta_{\text{oracle}} := b^{S_{\text{oracle}}}$, and $f_{\text{oracle}} := f_{S_{\text{oracle}}}$, with S_{oracle} given in (5), we have

$$\|\hat{\beta} - \beta_{\text{oracle}}\|_2 \leq \frac{10}{\lambda\sqrt{|S_{\text{oracle}}|}} \left\{ \|f_{\text{oracle}} - \mathbf{f}^0\|_2^2/n + \frac{(7\lambda)^2|S_{\text{oracle}}|}{\phi_{\text{adap}}^2(6, S_{\text{oracle}}, 2|S_{\text{oracle}}|)} \right\}.$$

This lemma was obtained in [29].

5.2. Variable selection. We now show that the Lasso is not very good in variable selection, unless rather strong conditions on the Gram matrix are met. To simplify the exposition, we assume in this subsection that there is no noise. We let $\{\psi_j\}_{j=1}^p$ be a given dictionary in $L_2(Q)$, with Gram matrix $\Sigma := \int \psi^T \psi dQ := (\sigma_{j,k})$. Furthermore, for an index set S , we consider the

submatrices

$$\Sigma_{1,1}(S) := (\sigma_{j,k})_{j \in S, k \in S}, \quad \Sigma_{2,2}(S) := (\sigma_{j,k})_{j \notin S, k \notin S},$$

and

$$\Sigma_{2,1}(S) = (\sigma_{j,k})_{j \notin S, k \in S}, \quad \Sigma_{1,2}(S) := (\sigma_{j,k})_{j \in S, k \notin S}.$$

We let, as before, $\Lambda_{\min}^2(S)$ be the smallest eigenvalue of $\Sigma_{1,1}(S)$.

The noiseless Lasso is

$$\beta_{\text{Lasso}} := \arg \min_{\beta} \{ \|f_{\beta} - f^0\|^2 + \lambda \|\beta\|_1 \}.$$

Here,

$$f^0 = f_{\beta_{\text{true}}},$$

is assumed to be linear, with a sparse vector of coefficients β_{true} . Our aim is to estimate $S_{\text{true}} := \{j : \beta_{j,\text{true}} \neq 0\}$ using the Lasso $S_{\text{Lasso}} = \{j : \beta_{j,\text{Lasso}} \neq 0\}$.

The irrerepresentable condition can be found in [32]. We use a slightly modified version.

Definition

Part 1 We say that the irrerepresentable condition is met for the set S , if for all vectors $\tau_S \in \mathbb{R}^{|S|}$ satisfying $\|\tau_S\|_{\infty} \leq 1$, we have

$$\|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S\|_{\infty} < 1. \tag{10}$$

Part 2 Moreover, for a fixed $\tau_S \in \mathbb{R}^{|S|}$ with $\|\tau_S\|_{\infty} \leq 1$, the weak irrerepresentable condition holds for τ_S , if

$$\|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S\|_{\infty} \leq 1.$$

Part 3 Finally, for some $0 < \theta < 1$, the θ -uniform irrerepresentable condition is met for the set S , if

$$\max_{\|\tau_S\|_{\infty} \leq 1} \|\Sigma_{2,1}(S)\Sigma_{1,1}^{-1}(S)\tau_S\|_{\infty} \leq \theta.$$

The next theorem summarizes some results of [28].

Theorem 5.1.

Part 1 Suppose the irrerepresentable condition is met for S_{true} . Then $S_{\text{Lasso}} \subset S_{\text{true}}$.

Part 2 Conversely, suppose that $S_{\text{Lasso}} \subset S_{\text{true}}$, and that

$$|\beta_{j,\text{true}}| > \lambda \sup_{\|\tau_{S_{\text{true}}}\|_{\infty} \leq 1} \|\Sigma_{1,1}^{-1}(S_{\text{true}})\tau_{S_{\text{true}}}\|_{\infty}/2.$$

Then the weak irrerepresentable condition holds for the sign-vector

$$\tau_{\text{true}} := \text{sign}((\beta_{\text{true}})_{S_{\text{true}}}).$$

Part 3 Suppose that for some $\theta < 1/L$, the θ -uniform irrerepresentable condition is met for S . Then the compatibility condition holds with $\phi^2(\Sigma, L, S) \geq (1 - L\theta)^2 \Lambda_{\min}^2(S)$.

One may also verify that the irrerepresentable condition implies exact recovery:

$$\beta_{\text{LP}} = \beta_{\text{true}},$$

where β_{LP} is given in (9).

5.3. The adaptive Lasso. The adaptive Lasso introduced by [33] is

$$\hat{\beta}_{\text{adap}} := \arg \min_{\beta} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda_{\text{init}} \lambda_{\text{adap}} \sum_{j=1}^p \frac{|\beta_j|}{|\hat{\beta}_{j,\text{init}}|} \right\}. \tag{11}$$

Here, $\hat{\beta}_{\text{init}}$ is the one-stage Lasso defined in (1), with initial tuning parameter $\lambda := \lambda_{\text{init}}$, and $\lambda_{\text{adap}} > 0$ is the tuning parameter for the second stage. Note that when $|\hat{\beta}_{j,\text{init}}| = 0$, we exclude variable j in the second stage.

We write $\hat{f}_{\text{init}} := \mathbf{X}\hat{\beta}_{\text{init}}$ and $\hat{f}_{\text{adap}} := \mathbf{X}\hat{\beta}_{\text{adap}}$, with active sets $\hat{S}_{\text{init}} := \{j : \hat{\beta}_{j,\text{init}} \neq 0\}$ and $\hat{S}_{\text{adap}} := \{j : \hat{\beta}_{j,\text{adap}} \neq 0\}$, respectively.

Let

$$\hat{\delta}_{\text{init}}^2 := \|\mathbf{X}\hat{\beta}_{\text{init}} - \mathbf{f}^0\|_2^2/n,$$

be the prediction error of the initial Lasso, and and, for $q > 1$,

$$\hat{\delta}_q := \|\hat{\beta}_{\text{init}} - \beta_{\text{oracle}}\|_q$$

be its ℓ_q -error. Denote the prediction error of the adaptive Lasso as

$$\hat{\delta}_{\text{adap}}^2 := \|\mathbf{X}\hat{\beta}_{\text{adap}} - \mathbf{f}^0\|_2^2/n.$$

The next theorem was obtained in [29]. The first two parts actually repeat the statements of Theorem 2.1 and Lemma 5.1, albeit that we everywhere invoke the smaller minimal adaptive restricted eigenvalue $\phi_{\text{adap}}(6, S_{\text{oracle}}, 2|S_{\text{oracle}}|)$ instead of $\phi_{\text{comp}}(6, S_{\text{oracle}})$, which is not necessary for the bounds on $\hat{\delta}_{\text{init}}^2$ and $\hat{\delta}_1$. This is only to simplify the exposition.

Theorem 5.2. Consider the oracle set $S_0 := S_{\text{oracle}}$ given in (5), with cardinality $s_0 := |S_{\text{oracle}}|$. Let $\phi_0 := \phi_{\text{adap}}(6, S_0, 2s_0)$. Let

$$\mathcal{T}(\lambda_{\text{init}}) := \left\{ \max_{1 \leq j \leq p} 4|\epsilon^T \mathbf{X}_j|/n \leq \lambda_{\text{init}} \right\}.$$

Then on $\mathcal{T}(\lambda_{\text{init}})$, the following statements hold.

- 1) There exists a bound $\delta_{\text{init}}^{\text{upper}} = O(\lambda_{\text{init}} \sqrt{s_0}/\phi_0)$ such that

$$\hat{\delta}_{\text{init}} \leq \delta_{\text{init}}^{\text{upper}}.$$

2) For $q \in \{1, 2, \infty\}$, there exists bounds δ_q^{upper} satisfying

$$\delta_1^{\text{upper}} = O(\lambda_{\text{init}} s_0 / \phi_0^2), \quad \delta_2^{\text{upper}} = O(\lambda_{\text{init}} \sqrt{s_0} / \phi_0^2),$$

$$\delta_\infty^{\text{upper}} = O(\lambda_{\text{init}} \sqrt{s_0} / \phi_0^2),$$

such that

$$\hat{\delta}_q \leq \delta_q^{\text{upper}}, \quad q \in \{1, 2, \infty\}.$$

3) Let δ_2^{upper} and $\delta_\infty^{\text{upper}}$ be such bounds, satisfying $\delta_\infty^{\text{upper}} \geq \delta_2^{\text{upper}} / \sqrt{s_0}$, and $\delta_2^{\text{upper}} = O(\lambda_{\text{init}} \sqrt{s_0} / \phi_0^2)$. Let $|\beta_{\text{oracle}}|_{\text{harm}}^2$ be the trimmed harmonic mean

$$|\beta_{\text{oracle}}|_{\text{harm}}^2 := \left(\sum_{|\beta_{j,\text{oracle}}| > 2\delta_\infty^{\text{upper}}} \frac{1}{|\beta_{j,\text{oracle}}|^2} \right)^{-1}.$$

Suppose that

$$\lambda_{\text{adap}}^2 \asymp \left(\frac{1}{n} \left\| \mathbf{f}_{S_0^{\text{thres}}} - \mathbf{f}^0 \right\|_2^2 + \frac{\lambda_{\text{init}}^2 s_0}{\phi_0^2} \right) \frac{|\beta_{\text{oracle}}|_{\text{harm}}^2}{\lambda_{\text{init}}^2 / \phi_0^2}, \tag{12}$$

where $S_0^{\text{thres}} := \{j : |\beta_{j,\text{oracle}}| > 4\delta_\infty^{\text{upper}}\}$. Then

$$\hat{\delta}_{\text{adap}}^2 = O\left(\frac{1}{n} \left\| \mathbf{f}_{S_0^{\text{thres}}} - \mathbf{f}^0 \right\|_2^2 + \frac{\lambda_{\text{init}}^2 s_0}{\phi_0^2} \right),$$

and

$$|\hat{S}_{\text{adap}} \setminus S_0| = O\left(\frac{\lambda_{\text{init}}^2 s_0}{\phi_0^2} \frac{1}{|\beta_{\text{oracle}}|_{\text{harm}}^2} \right).$$

The value (12) for the tuning parameter seems complicated, but it generally means we take it in such a way that the the adaptive Lasso has its prediction error optimized. The message of the theorem is that when using cross-validation for choosing the tuning parameters, the adaptive Lasso will - when the minimal adaptive restricted eigenvalues are under control - have $O(s_0)$ false positives, and possibly less, e.g., when the trimmed harmonic mean of the oracle coefficients is large. As far as we know, the cross-validated initial Lasso can have $O(s_0)$ false positives only when strong conditions on the Gram matrix $\hat{\Sigma}$ are met, for instance the condition that the maximal eigenvalue of $\hat{\Sigma}$ is $O(1)$ (and in that case the adaptive Lasso wins again by having $O(\sqrt{s_0})$ false positives). On the other hand, the prediction error of the adaptive Lasso is possibly less good than that of the initial Lasso.

6. The Lasso with Within Group Structure

Finally, we study a procedure for regression with group structure. The co-variables are divided into p given groups. The parameters within a group are assumed to either all zero, or all non-zero.

We consider the linear model

$$\mathbf{Y} = \mathbf{X}\beta + \epsilon.$$

As before, ϵ is a vector of noise, which, for definiteness, we assume to be $\mathcal{N}(0, I)$ -distributed. Furthermore, \mathbf{X} is now an $(n \times M)$ -matrix of co-variables. There are p groups of co-variables, each of size T (i.e., $M = pT$), where both p and T can be large. We rewrite the model as

$$\mathbf{Y} = \sum_{j=1}^p \mathbf{X}_j \beta_j + \epsilon,$$

where $\mathbf{X}_j = \{\mathbf{X}_{j,t}\}_{t=1}^T$ is an $(n \times T)$ -matrix and $\beta_j = (\beta_{j,1}, \dots, \beta_{j,T})^T$ is a vector in \mathbb{R}^T . To simplify the exposition, we consider the case where $T \leq n$ and where the Gram matrix within groups is normalized, i.e., $\mathbf{X}_j^T \mathbf{X}_j/n = I$ for all j . The number of groups p can be very large. The group Lasso was introduced by [30]. With large T (say $T = n$), the standard group Lasso will generally not have good prediction properties, even when p is small (say $p = 1$). Therefore, one needs to impose a certain structure within groups. Such an approach has been considered by [19], [22], and [11].

We present results from [27], which are similar to those in [19]. We assume that for all j , there is an ordering in the variables of group j : the larger t , the less important variable $\mathbf{X}_{j,t}$ is likely to be. Given positive weights $\{w_t\}_{t=1}^T$ (which we for simplicity assume to be the same for all groups j), satisfying $0 < w_1 \leq \dots \leq w_T$, we express the structure in group j as the weighted sum

$$\|W\beta_j\|_2^2 := \sum_{t=1}^T w_t^2 \beta_{j,t}^2, \quad \beta_j \in \mathbb{R}^p.$$

The structured group Lasso estimator is defined as

$$\hat{\beta}_{\text{SGL}} := \arg_{\beta \in \mathbb{R}^{pT}} \left\{ \|\mathbf{Y} - \mathbf{X}\beta\|_2^2/n + \lambda \sum_{j=1}^p \|\beta_j\|_2 + \lambda\mu \sum_{j=1}^p \|W\beta_j\|_2 \right\}, \quad (13)$$

where λ and μ are tuning parameters. The idea here is that the variables $X_{j,t}$ with t large are thought of as being less important. For example $X_{j,t}$ could be the t^{th} resolution level in a Fourier expansion, or the t^{th} order interaction term for categorical variables, etc.

Let

$$R^2(t) := \sum_{s>t} \frac{1}{w_s^2}, \quad t = 1, \dots, T.$$

Let $T_0 \in \{1, \dots, T\}$ be the smallest value such that

$$T_0 \geq R(T_0)\sqrt{n}.$$

Take $T_0 = T$ if such a value does not exist. We call T_0 the *hidden truncation level*. The faster the w_j increase, the smaller T_0 will be, and the more structure we have within groups. The choice of T_0 is in a sense inspired by a bias-variance trade-off.

We will throughout take the tuning parameters λ and μ such that $\lambda \geq \sqrt{T_0/n}$ and $\lambda\mu \geq T_0/n$.

Let, for $x > 0$,

$$\nu_0^2 := \nu_0^2(x) = (2x + 2 \log(pT)),$$

and

$$\xi_0^2 := \xi_0^2(x) = 1 + \sqrt{\frac{4x + 4 \log p}{T_0}} + \frac{4x + 4 \log p}{T_0}.$$

Define the set

$$\mathcal{T} := \left\{ \max_{1 \leq j \leq p} \|V_j\|_\infty \leq \nu_0, \max_{1 \leq j \leq p} \xi_j^2/T_0 \leq \xi_0^2 \right\}.$$

Here, $V_j^T := \epsilon^T \mathbf{X}_j / \sqrt{n}$, and $\xi_j^2 = \sum_{t=1}^{T_0} V_{j,t}^2$, $j = 1, \dots, p$.

Define

$$\hat{\Sigma} := \mathbf{X}^T \mathbf{X} / n,$$

and write

$$\|\beta\|_{\hat{\Sigma}}^2 := \beta^T \hat{\Sigma} \beta.$$

When $M = pT$ is larger than n , it is clear that $\hat{\Sigma}$ is singular. To deal with this, we will (as in Lemma 4.2) approximate $\hat{\Sigma}$ by a matrix Σ , which potentially is non-singular. We let Σ_j be the $(T \times T)$ -submatrix of Σ corresponding to the variables in the j^{th} group (as $\hat{\Sigma}_j = I$, we typically take $\Sigma_j = I$ as well), and we write

$$\|\beta\|_{\Sigma}^2 := \beta^T \Sigma \beta, \|\beta_j\|_{\Sigma_j}^2 := \beta_j^T \Sigma_j \beta_j, j = 1, \dots, p.$$

We invoke the notation

$$\text{pen}_1(\beta) := \lambda \sum_j \|\beta_j\|_2, \text{pen}_2(\beta) := \lambda\mu \sum_j \|W\beta_j\|_2,$$

and

$$\text{pen}(\beta) := \text{pen}_1(\beta) + \text{pen}_2(\beta).$$

For an index set $S \subset \{1, \dots, p\}$, we let

$$\beta_{j,S} = \beta_j \mathbf{1}\{j \in S\}, j = 1, \dots, p$$

(recall that β_j is now a T -vector).

Definition *The structured group Lasso (Σ, L, S) -compatibility condition holds if*

$$\phi_{\text{struc}}^2(\Sigma, L, S) := \min \left\{ \frac{|S| \|\beta\|_{\Sigma}^2}{(\sum_{j \in S} \|\beta_j\|_{\Sigma_j})^2} : \text{pen}_1(\beta_{S^c}) + \text{pen}_2(\beta) \leq L \text{pen}_1(\beta_S) \right\}$$

is strictly positive.

Let

$$\mathcal{S}(\Sigma) := \left\{ S : \frac{64n\lambda^2 \|\hat{\Sigma} - \Sigma\|_{\infty} |S|}{\phi_{\text{struc}}^2(\Sigma, \mathfrak{S}, S)} \leq \frac{1}{2} \right\}.$$

By considering only sets $S \in \mathcal{S}(\Sigma)$, we actually put a bound on the sparsity we allow, i.e., we cannot handle very non-sparse problems very well. Mathematically, it is allowed to take $\Sigma = \hat{\Sigma}$, having $\mathcal{S}(\hat{\Sigma})$ being every set S with strictly positive $\phi_{\text{struc}}(\hat{\Sigma}, \mathfrak{S}, S)$. The reason we generalize to approximating matrices Σ is that this helps to check the structured group Lasso (Σ, L, S) -compatibility condition.

Theorem 6.1. *Let*

$$\mathbf{Y} = \mathbf{f}^0 + \epsilon,$$

where ϵ is $\mathcal{N}(0, I)$ -distributed. We have $\mathbf{P}(\mathcal{T}) \geq 1 - 3 \exp[-x]$. Consider the structured group Lasso $\hat{\beta}_{\text{SGL}}$ given in (13), and define $\hat{f}_{\text{SGL}} := \mathbf{X} \hat{\beta}_{\text{SGL}}$. Assume

$$\lambda \geq 8\xi_0 \sqrt{T_0/n}, \quad \lambda\mu \geq 8\nu_0 T_0/n.$$

On \mathcal{T} , we have for all $S \in \mathcal{S}(\Sigma)$ and all β_S ,

$$\|\hat{f}_{\text{SGL}} - \mathbf{f}^0\|_2^2/n + \text{pen}(\hat{\beta}_{\text{struc}} - \beta_S) \leq 4\|f_{\beta_S} - \mathbf{f}^0\|_2^2/n + \frac{(4\lambda)^2 |S|}{\phi_{\text{struc}}^2(\Sigma, \mathfrak{S}, S)} + 8\text{pen}_2(\beta_S).$$

In other words, the structured group Lasso mimics an oracle that selects groups of variables in a sparse way. Note that the tuning parameter λ is now generally of larger order than in the standard Lasso setup (1). This is the price to pay for having large groups. As an extreme case, one may consider the situation with weights $w_t = 1$ for all t . Then $T_0 = T$, and the oracle bound is up to log p -factors the same as the one obtained by the standard Lasso.

7. Conclusion

The Lasso is an effective method for obtaining oracle optimal prediction error or excess risk. For variable selection, the adaptive Lasso or other two-stage procedures can be applied, generally leading to less false positives at the price of reduced predictive power (or a larger number of false negatives). A priori structure in the variables can be dealt with by using a group Lasso, possibly with an additional within group penalty.

Future work concerns modifications that try to cope with large correlations between variables. Moreover, it will be of interest to go beyond generalized linear modeling.

References

- [1] P. Bickel, Y. Ritov, and A. Tsybakov. Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, **37**:1705–1732, 2009.
- [2] O. Bousquet. A Bennet concentration inequality and its application to suprema of empirical processes. *C.R. Acad. Sci. Paris*, **334**:495–550, 2002.
- [3] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation and sparsity via l_1 -penalized least squares. In *Proceedings of 19th Annual Conference on Learning Theory*, COLT 2006. Lecture Notes in Artificial Intelligence 4005, 379–391, Heidelberg, 2006. Springer Verlag.
- [4] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Aggregation for Gaussian regression. *Annals of Statistics*, **35**:1674–1697, 2007.
- [5] F. Bunea, A. Tsybakov, and M.H. Wegkamp. Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, **1**:169–194, 2007.
- [6] F. Bunea, A.B. Tsybakov, and M.H. Wegkamp. Sparse Density Estimation with l_1 Penalties. In *Learning Theory 20th Annual Conference on Learning Theory*, COLT 2007, San Diego, CA, USA, June 13–15, 2007: Proceedings, 530–543. Springer, 2007.
- [7] E. Candès and T. Tao. Decoding by linear programming. *IEEE Transactions on Information Theory*, **51**:4203–4215, 2005.
- [8] E. Candès and T. Tao. The Dantzig selector: statistical estimation when p is much larger than n . *Annals of Statistics*, **35**:2313–2351, 2007.
- [9] E. Greenshtein and Y. Ritov. Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli*, **10**:971–988, 2004.
- [10] J. Huang, S. Ma, and C.-H. Zhang. Adaptive Lasso for sparse high-dimensional regression models. *Statistica Sinica*, **18**:1603–1618, 2008.
- [11] Koltchinskii, V. and Yuan, M. Sparse recovery in large ensembles of kernel machines. In *Conference on Learning Theory, COLT*, 229–238, 2008.
- [12] V. Koltchinskii. Sparsity in penalized empirical risk minimization. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, **45**:7–57, 2009.
- [13] V. Koltchinskii. The Dantzig selector and sparsity oracle inequalities. *Bernoulli*, **15**:79–828, 2009.
- [14] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and processes*. Springer, 1991.
- [15] Loubes, J.M. and van de Geer, S. Adaptive estimation with soft thresholding penalties. *Statistica Neerlandica*, **56**, 454–479, 2002.
- [16] K. Lounici. Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators. *Electronic Journal of Statistics*, **2**:90–102, 2008.
- [17] P. Massart. About the constants in Talagrand's concentration inequalities for empirical processes. *Annals of Probability*, **28**:863–884, 2000.
- [18] P. Massart. Some applications of concentration inequalities to statistics. *Ann. Fac. Sci. Toulouse*, **9**:245–303, 2000.

-
- [19] L. Meier, S. van de Geer and P. Bühlmann. High-dimensional additive modeling. *Annals of Statistics*, **37**:3779–3821, 2009.
- [20] N. Meinshausen. Relaxed Lasso. *Computational Statistics and Data Analysis*, **52**:374–393, 2007.
- [21] N. Meinshausen and P. Bühlmann High dimensional graphs and variable selection with the Lasso. *The Annals of Statistics*, **34**:1436–1462, 2006.
- [22] P. Ravikumar, H. Liu, J. Lafferty, and L. Wasserman. SpAM: sparse additive models. *Advances in neural information processing systems*, **20**:1201–1208, 2008.
- [23] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, **58**:267–288, 1996.
- [24] R. Tibshirani, M. Saunders, S. Rosset, J. Zhu and K. Knight. Sparsity and smoothness via the fused Lasso. *Journal of the Royal Statistical Society Series B*, **67**:91–108, 2005.
- [25] S. van de Geer. The deterministic Lasso. *The JSM Proceedings*, 2007.
- [26] S. van de Geer. High-dimensional generalized linear models and the Lasso. *Annals of Statistics*, **36**:614–645, 2008.
- [27] S. van de Geer. The Lasso with within group structure. *IMS Lecture Notes*, submitted, 2010.
- [28] S. van de Geer and P. Bühlmann. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics*, **3**:1360–1392, 2009.
- [29] S. van de Geer, P. Bühlmann, and S. Zhou. Prediction and variable selection with the adaptive Lasso. Seminar for Statistics, ETH Zürich, preprint available at ArXiv: 1001.5176, 2010.
- [30] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, **68**:49–67, 2006.
- [31] T. Zhang. Some sharp performance bounds for least squares regression with L1 regularization. *Annals of Statistics*, **37**:2109–2144, 2009.
- [32] P. Zhao and B. Yu. On model selection consistency of Lasso. *Journal of Machine Learning Research*, **7**:2541–2567, 2006.
- [33] H. Zou. The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, **101**:1418–1429, 2006.
- [34] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, **67**:301–320, 2005.

Bayesian Regularization

Aad van der Vaart*

Abstract

We consider the recovery of a curve or surface from noisy data by a nonparametric Bayesian method. This entails modelling the surface as a realization of a “prior” stochastic process, and viewing the data as arising by measuring this realization with error. The conditional distribution of the process given the data, given by Bayes’ rule and called “posterior”, next serves as the basis of all further inference. As a particular example of priors we consider Gaussian processes. A nonparametric Bayesian method can be called successful if the posterior distribution concentrates most of its mass near the surface that produced the data. Unlike in classical “parametric” Bayesian inference the quality of the Bayesian reconstruction turns out to depend on the choice of the prior. For instance, it depends on the fine properties of the sample paths of a Gaussian process prior, with good results obtained only if these match the properties of the true surface. The Bayesian solution to overcome the problem that these fine properties are typically unknown is to put additional priors on hyperparameters. For instance, sample paths of a Gaussian process prior are rescaled by a random amount. This leads to mixture priors, to which Bayes’ rule can be applied as before. We show that this leads to minimax precision in several examples: adapting to unknown smoothness or sparsity. We also present abstract results on hierarchical priors.

Mathematics Subject Classification (2010). Primary 62H30, 62-07; Secondary 65U05, 68T05.

Keywords. Posterior distribution, nonparametric Bayes, Gaussian process prior, regression, classification, density estimation, rate of contraction, adaptation, sparsity.

1. Introduction

The last decades have seen a growing interest in Bayesian methods for recovering curves, surfaces or other high-dimensional objects from noisy measurements.

*Dept. Mathematics, VU University Amsterdam, De Boelelaan 1081, Amsterdam, The Netherlands. E-mail: aad@few.vu.nl.

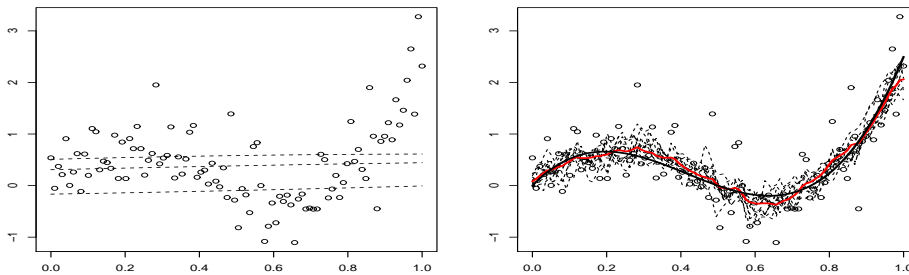


Figure 1. Observations in a regression problem. Three realizations from a Gaussian process prior (left panel) and 10 from the posterior distribution (right panel). The true regression curve and the posterior mean are indicated in the right panel. The Bayesian updating is successful: the realizations from the posterior are much closer to the truth than those from the prior.

The object θ is modelled as a realization from some *prior* probability distribution Π , and the observed data X is viewed as drawn from a probability density $x \mapsto p_\theta(x)$ that depends on the realization of θ . The *posterior* distribution of the “parameter” θ is then given by Bayes’ rule as

$$d\Pi(\theta|X) \propto p_\theta(X) d\Pi(\theta).$$

Any question that is expressible in the parameter can in principle be answered by querying this distribution. Practical implementation of this Bayesian paradigm is nontrivial if the parameter space is infinite-dimensional, but has been made possible by modern computing. For instance, MCMC methods allow to generate a Markov chain $\theta_1, \theta_2, \dots, \theta_B$ with the posterior as its stationary distribution, and questions can be answered by simple averaging procedures. In particular, if the θ are functions, then the average $B^{-1} \sum_{i=1}^B \theta_i$ gives an approximation to the mean of the posterior distribution, of a precision that is controlled by the number B of simulated values. Estimates of the spread of the posterior distribution can similarly be obtained.

In this paper we are concerned not with computational issues, but with the quality of the posterior distribution itself. To investigate this we put ourselves in a non-Bayesian framework, where it is assumed that the data X are generated according to the density p_{θ_0} determined by a fixed parameter θ_0 , and view the posterior distribution as just a random measure on the parameter space. The Bayesian procedure is considered accurate if this random measure concentrates its mass near the parameter θ_0 . We wish this to be true for many θ_0 simultaneously, preferably uniformly in θ_0 belonging to a class of test models. For instance, a set of surfaces known to have a certain number of bounded derivatives.

Except in very special cases this question can be investigated only in an asymptotic setting. We consider data X^n depending on an index n (for instance sample size) and study the resulting sequence of posterior distributions

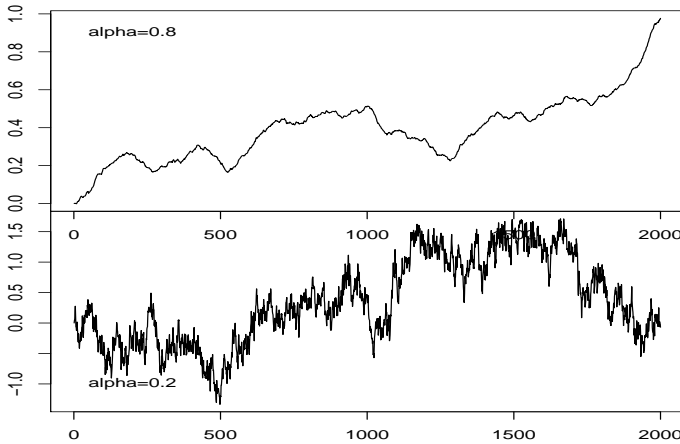


Figure 2. Realizations of a fractional Brownian Motion with Hurst index 0.8 (top panel) and 0.2 (bottom panel). The fine properties of the sample paths determine the accuracy of a Bayesian reconstruction using these priors.

$d\Pi_n(\theta | X^n)$ as $n \rightarrow \infty$. In a setting where the informativeness of the data increases indefinitely with n , we desire that this sequence contracts to the Dirac measure at θ_0 , meaning complete recovery “in the limit”. Given a metric structure d on the parameter space, we can more precisely measure the *rate of contraction*. We say that this is at least ε_n if, for any sequence of constants $M_n \rightarrow \infty$,

$$\Pi_n(\theta: d(\theta, \theta_0) < M_n \varepsilon_n | X^n) \rightarrow 1.$$

The convergence can be in mean, or in the almost sure sense. Thus the posterior distribution puts almost all its mass on balls of radius of the order ε_n around θ_0 .

In classical finite-dimensional problems, with θ a vector in Euclidean space and n the sample size, the rate of contraction ε_n is typically $n^{-1/2}$, relative to for instance the Hellinger distance, for any prior with full support. For smoothly parameterized models the *Bernstein-von Mises theorem* also shows that the posterior asymptotically resembles a normal distribution centered at the maximum likelihood estimator $\hat{\theta}_n$ and with scale equal to $1/n$ times the Fisher information:

$$\left\| \Pi_n(\theta \in \cdot | X^n) - N(\hat{\theta}_n(X^n), n^{-1} I_{\theta}) \right\| \rightarrow 0.$$

The prior distribution does not appear in this approximation and is said to “wash out” as $n \rightarrow \infty$. In nonparametric problems this is very different. First there are many priors which do not lead to contraction of the posterior at all. Second many natural priors yield a rate of contraction that depends on the combination of the prior and the true parameter. The positive news is that a good match between prior and θ_0 may lead to an *optimal rate of contraction*, equal to the minimax rate for a problem.

In practice such “good matches” may not be easy to achieve. It is never trivial to have a proper intuitive understanding of a prior probability distribution on an infinite-dimensional set. Furthermore, and more importantly, one does not know the fine properties of the true parameter θ_0 . Figures 2 and 3 illustrate these points. The Hurst index of fractional Brownian has a strong influence on the appearance of the sample paths of this process, but it is not clear what influence this difference has on estimating a particular true function θ_0 . Whereas the visible appearance of the two priors in Figure 2 is different, it is almost impossible to distinguish between the sample paths in the three bottom panels of Figure 3, which are realizations of one, two and three times integrated Brownian motion. All these priors lead to different rates of contraction for a given θ_0 .

The elegant solution to the dilemma of prior choice (a classical point of criticism to Bayesian methods) is to work with many priors at the same time. We start with a collection of priors Π_α , indexed by some parameter α in an index set A , which is assumed to contain at least one appropriate prior for each possible truth θ_0 . Next we combine these priors by putting a prior distribution, a *hyper prior*, on the index α . If A is countable and the hyper prior is denoted by $(\lambda_\alpha: \alpha \in A)$, then this just leads to the overall prior

$$\Pi = \sum_{\alpha} \lambda_{\alpha} \Pi_{\alpha}.$$

Inference, using Bayes’ rule, proceeds as before. The hope is that the data will automatically “use” the priors Π_α that are appropriate for θ_0 , and produce a posterior that contracts at an optimal rate, given that at least one of the priors Π_α would produce this rate if used on its own.

This automatic *adaptation* of the posterior distribution sounds too good to be true. Obviously, it depends strongly on the weights $(\lambda_\alpha: \alpha \in A)$ and the prior distributions Π_α . Because the latter often possess very different “dimensionalities”, finding appropriate weights can be delicate. However, quite natural schemes turn out to do the job, although sometimes a logarithmic factor is lost. In this paper we first consider adaptation to the regularity of a surface θ_0 using Gaussian process priors, next consider adaptation to sparsity, and finally present an abstract result. We present theorems without proofs; these can be found in the papers [10], [2] and [4]. These papers also give references to the considerable literature on adaptation by non-Bayesian methods.

2. Gaussian Process Priors

Imagine estimating a curve or surface $w: T \rightarrow \mathbb{R}$ on some domain T , for instance a regression or density function, by modelling this apriori as the sample path of a Gaussian process $W = (W_t: t \in T)$, and next letting Bayes’ rule do the work and come up with the resulting posterior distribution. There is a great

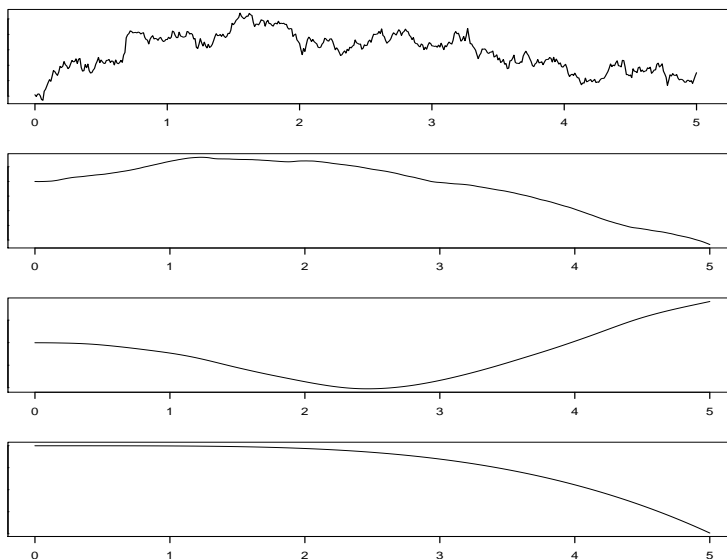


Figure 3. Realizations of 0, 1, 2 or 3 times integrated Brownian Motion. The fine properties of the sample paths determine the accuracy of a Bayesian reconstruction using these priors.

variety of Gaussian processes that can be employed. If we restrict to centered processes, then each is characterized by its covariance function

$$(s, t) \mapsto \text{cov}(W_s, W_t).$$

Because there are so many different covariance functions, Gaussian process priors have gained some popularity (see e.g. [1]).

Often more insight in the prior modelling can be gained from visualizing the sample paths of the process. There are very rough processes, like Brownian motion, but also infinitely smooth ones. Not surprisingly the regularity of the prior influences the posterior. Perhaps it is surprising that this influence does not disappear if the informativeness of the observations increases indefinitely: the prior does not wash out. This influence concerns the regularity of the sample paths of the posterior, but more importantly the concentration of the posterior near the true curve.

We measure this by the rate of contraction of the posterior distribution. For Gaussian priors this depends on two quantities (see [7]). First a Gaussian distribution has a certain concentration near its mean, measured in a *small ball probability*. For instance, for uniform balls in the one-dimensional case this is the probability that the process remains within bands at heights $\pm\varepsilon$, for small ε of course (see Figure 4). More generally, if the Gaussian variable W takes its values in a Banach space, the *exponent of the small ball probability* can be

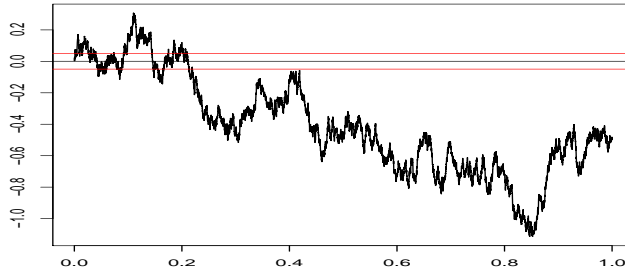


Figure 4. The small ball probability of Brownian Motion relative to the uniform norm is the probability that a Brownian sample path remains between bands at heights $-\varepsilon$ and ε . (The depicted realization does not and hence does not contribute to the small ball probability at this ε .)

defined as

$$\phi_0(\varepsilon) = -\log P(\|W\| < \varepsilon).$$

If the small ball probability is small (the exponent ϕ_0 increases rapidly as $\varepsilon \downarrow 0$), then the rate of contraction of the posterior will be small as well. In fact, if the norm $\|\cdot\|$ matches up with the statistical setting (see later for examples), then the rate of contraction ε_n is not faster than the solution to

$$\phi_0(\varepsilon_n) \sim n\varepsilon_n^2.$$

This is not necessarily bad, as the true curve may be intrinsically hard to estimate. However, the small ball probability is a property of the prior only, not of the true curve, and hence is one property through which the prior may express itself in the posterior. For instance, Brownian motion has small ball exponent $(1/\varepsilon)^2$ and hence will never give contraction rates faster than $n^{-1/4}$. Its small ball probabilities are small, because its sample paths rarely stay close to zero.

The second quantity that determines the rate of contraction is the position of the true curve relative to the prior. Clearly if it is outside the support of the prior, then the posterior will not even be consistent. A position inside the support can be quantified by its position relative to the *reproducing kernel Hilbert space* (RKHS) of the prior. (See [9] for an introduction to RKHS appropriate to prior distributions.) Being inside the RKHS gives the fastest rate, but other positions give some rate, which can be computed from the RKHS-norm. If w_0 is the true surface and $\|\cdot\|_{\mathbb{H}}$ denotes the RKHS-norm, then the crucial quantity is

$$\phi_{w_0}(\varepsilon) := \phi_0(\varepsilon) + \inf_{\|h-w_0\| < \varepsilon} \|h\|_{\mathbb{H}}^2.$$

Under general conditions it can be shown ([7, 8]) that the rate of contraction is ε_n if

$$\phi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2.$$

Small ball probabilities and RKHSs are somewhat complicated objects, but there is big literature that permits obtaining contraction rates for many examples.

Unless one is a true Bayesian, and believes strongly in the fine properties of the prior, the dependence of the contraction rate on the prior is not good news. It is possible to alleviate this dependence by combining priors. We shall study this here for adapting to unknown smoothness of a surface w_0 . In this case an elegant way of combining Gaussian priors is to rescale the sample paths of a given process. Running a process for a longer time and mapping its time domain to a shorter interval creates more variability (see Figure 5), whereas rescaling time in the other direction smoothes the sample paths. The scaling variable can be viewed as a bandwidth, and the obvious Bayesian approach is to choose this from a prior.

We shall illustrate with one such a construction. It employs a fixed prior distribution, constructed by rescaling a smooth Gaussian random field. For definiteness we use the squared exponential process combined with an inverse Gamma bandwidth. The *squared exponential process* is the centered Gaussian process $W = (W_t: t \in \mathbb{R}^d)$ with covariance function, for $\|\cdot\|$ the Euclidean norm on \mathbb{R}^d ,

$$\text{cov}(W_s, W_t) = \exp(-\|t - s\|^2). \quad (1)$$

The Gaussian field W is well known to have a version with analytic sample paths $t \mapsto W_t$. To make it suitable as a prior for surfaces that are less smooth, we rescale the sample paths by an independent random variable A distributed as the d th root of a Gamma variable. As a prior distribution for a function on the domain $[0, 1]^d$ we consider the law of the process

$$(W_{At}: t \in [0, 1]^d).$$

The inverse $1/A$ of the variable A can be viewed as a bandwidth parameter. For large A the prior sample path $t \mapsto W_{At}$ is obtained by shrinking the long sample path $t \mapsto W_t$ indexed by $t \in [0, A]^d$ to the unit cube $[0, 1]^d$. Thus it employs “more randomness” and becomes suitable as a prior model for less regular functions if A is large (cf. [8]).

We measure the quality of the recovery of a surface w_0 using this prior by studying the rate of contraction of the corresponding posterior for surfaces w_0 belonging to two scales of test models. First the scale of *Hölder spaces* $C^\alpha[0, 1]^d$ consists of functions $w: [0, 1]^d \rightarrow \mathbb{R}$ that have partial derivatives up to order $\alpha > 0$. For noninteger α it is understood, as usual, that the partial derivatives of order $\lfloor \alpha \rfloor$ are Lipschitz of order $\alpha - \lfloor \alpha \rfloor$. Second we consider a scale of infinitely smooth functions. Let $\mathcal{A}^{\gamma, r}(\mathbb{R}^d)$ be the space of functions $f: \mathbb{R}^d \rightarrow \mathbb{R}$ with Fourier transform \hat{f} satisfying $\int e^{\gamma\|\lambda\|^r} |\hat{f}|^2(\lambda) d\lambda < \infty$. These functions are infinitely often differentiable and “increasingly smooth” as γ or r increase; they extend to functions that are analytic on a strip in \mathbb{C}^d containing \mathbb{R}^d if $r = 1$ and to entire functions if $r > 1$.

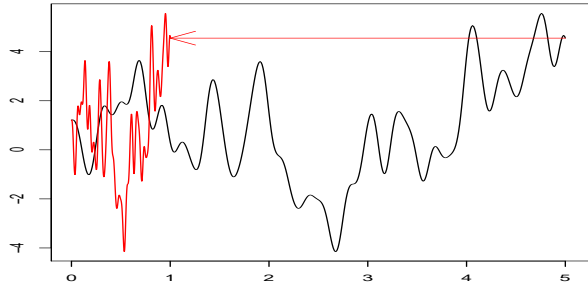


Figure 5. A realization of the squared exponential processes and its rescaling to the unit interval.

Typical minimax rates, relative to metrics that depend on the statistical setting, for these scales are $n^{-\alpha/(2\alpha+d)}$ if w_0 belongs to the Hölder space of order α , and $n^{-1/2}(\log n)^\kappa$ if it belongs to $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$, for the logarithmic exponent κ depending on r and increasing to $(d + 1)/2$ as $r \uparrow \infty$. Thus higher precision is possible for smoother surfaces w_0 , with the precision approaching the parametric precision $n^{-1/2}$ as the regularity increases.

Of course, it is necessary to describe how the data X^n relates to the surface w_0 . Rather than describing this in abstract terms, we give three examples: density estimation, regression, and classification.

Example 2.1 (*Density estimation*). A sample path of a Gaussian process is not a suitable model for a probability density. We transform it by exponentiation and renormalization, and as a prior distribution Π for a probability density $f_0: [0, 1]^d \rightarrow \mathbb{R}$ on the unit cube we use the distribution of

$$t \mapsto \frac{e^{W_{At}}}{\int_{[0,1]^d} e^{W_{As}} ds}.$$

We assume that the data X^n consists of a random sample X_1, \dots, X_n from a continuous, positive density f_0 on the unit cube $[0, 1]^d \subset \mathbb{R}^d$, and measure the rate of contraction of the posterior distribution by the *Hellinger distance*, the L_2 -distance between root-densities. To link to the preceding discussion on estimating a function w_0 , we represent the true density as $f_0 = e^{w_0}$.

Example 2.2 (*Fixed design regression*). A sample path of a Gaussian process can be used without transformation as a prior for a regression function. We consider data X^n consisting of independent variables X_1, \dots, X_n satisfying the regression relation $X_i = w_0(t_i) + \varepsilon_i$, for independent $N(0, \sigma_0^2)$ -distributed error variables ε_i and known elements t_1, \dots, t_n of the unit cube $[0, 1]^d$. The law of the random field $(W_{At}: t \in [0, 1]^d)$ is used as a prior for w_0 . If the standard deviation σ_0 of the errors is unknown, we endow it with a prior distribution as well, which we assume to be supported on a given interval $[a, b] \subset (0, \infty)$ that contains σ_0 , with a Lebesgue density that is bounded away from zero. The rate

of contraction is measured by the *empirical L_2 -norm*, with square

$$\|w\|_n^2 = n^{-1} \sum_{i=1}^n w^2(t_i),$$

the L_2 -norm corresponding to the empirical distribution of the design points.

Example 2.3 (*Classification*). In the classification problem the data X^n consists of a random sample $(Y_1, \Delta_1), \dots, (Y_n, \Delta_n)$, where Y_i takes values in the unit cube $[0, 1]^d$ and Δ_i takes values in the set $\{0, 1\}$. The statistical problem is to estimate the binary regression function r_0 given by $r_0(y) = P(\Delta_1 = 1 | Y_1 = y)$. Because this function has range $(0, 1)$, we transform a Gaussian process prior through a link function $\Psi: \mathbb{R} \rightarrow (0, 1)$, which we take to be the logistic or the normal distribution function. Thus, as a prior Π on r_0 we use the law of the process $(\Psi(W_{At}): t \in [0, 1]^d)$. The rate of contraction is measured by the $L_2(G)$ -norm relative to the marginal distribution G of Y_1 applied to the binary regression functions. We link up to the preceding discussion by setting $w_0 = \Psi^{-1}(r_0)$.

The rates of contraction for the three examples are the same, where we define $w_0 = \log f_0$, $w_0 = w_0$ and $w_0 = \Psi^{-1}(r_0)$ in the three examples, and use the metrics as indicated.

Theorem 2.1. *For every of the three examples:*

- If $w_0 \in C^\alpha[0, 1]^d$ for some $\alpha > 0$, then the posterior contracts at rate $n^{-\alpha/(2\alpha+d)} (\log n)^{(4\alpha+d)/(4\alpha+2d)}$.
- If w_0 is the restriction of a function in $\mathcal{A}^{\gamma,r}(\mathbb{R}^d)$, then the posterior contracts at rate $n^{-1/2} (\log n)^{d+1}$ if $r \geq 2$ and $n^{-1/2} (\log n)^{d+1+d/(2r)}$ if $r < 2$.

The theorem shows that the posterior distribution contracts at the minimax rate times a logarithmic factor, both in the Hölder and infinitely smooth scales. The remarkable fact is that near minimaxity is obtained in the Hölder scale for any $\alpha > 0$, even though the prior is fixed and does not refer to any of the individual spaces. Moreover, if the true surface happens to be infinitely smooth, then the posterior automatically produces a near parametric rate, without further work.

This adaptation is caused by the random rescaling of the squared exponential process. The latter process has analytic sample paths, and (apparently) can be roughened sufficiently by shrinkage (Figure 5). The effect of the rescaling can be seen by considering the rate of contraction without it. If the variable A would be replaced by a constant, then typical elements w_0 of each Hölder class are recovered by the posterior at no faster rate than $(\log n)^{-\nu}$, where the power ν increases with α , but the rate becomes never polynomial in n ([11]). Thus modelling a surface that is α -smooth, but not infinitely smooth, by a prior that is infinitely smooth leads to disastrous results.

The logarithmic factors in the rate are a bit disappointing. The theorem presents only an upper bound on the rates of contraction. However, we conjecture that a logarithmic factor is necessary for the present prior, even for the case of Hölder spaces, although the power $(4\alpha + d)/(4\alpha + 2d)$ may not be optimal. Such a loss of logarithmic factor is not characteristic of the Bayesian approach, as other priors can avoid it. As the loss is modest, in practice a simple prior of the type considered here may be preferable.

3. Sparsity

Suppose that we observe a random vector $X^n = (X_1, \dots, X_n)$ in \mathbb{R}^n such that

$$X_i = \theta_i + \varepsilon_i, \quad i = 1, \dots, n, \tag{2}$$

for independent standard normal random variables ε_i and an unknown vector of means $\theta = (\theta_1, \dots, \theta_n)$. We are interested in the estimation of θ when the vector is thought to be *sparse*. This problem is of interest on its own, for instance as a model for high-throughput experiments when many variables are measured simultaneously with noise. Applied with the θ_i equal to the coefficients of an expansion of a curve or surface in a basis, the model is also related to curve estimation.

Sparsity can be made precise in various ways. We take it here to mean that only few coordinates of θ are nonzero. For $\theta_0 = (\theta_{0,1}, \dots, \theta_{0,n})$ the parameter vector that actually produces the data, set

$$p_{0,n} = \#\{1 \leq i \leq n: \theta_{0,i} \neq 0\}.$$

We assume that $p_{0,n} \ll n$, and wish to estimate the parameter accurately relative to the Euclidean norm on \mathbb{R}^n . If the set of nonzero parameters were known a-priori, then the model would effectively be $p_{0,n}$ -dimensional, and the parameter vector could be estimated with mean square error of the order $p_{0,n}$. It turns out that not knowing the sparse set of coordinates needs to incur only to a logarithmic loss, resulting in minimax rates of the order $p_{0,n} \log(n/p_{0,n})$ (see [5]). The Gaussianity of the perturbations ε_i is essential for this, as it causes coordinates X_i with zero mean to be close to zero.

Besides the set of nonzero parameters, we assume the number $p_{0,n}$ of nonzero parameters to be unknown. A Bayesian approach starts by putting a prior π_n on this number, which is next extended to a full prior on the set of all possible sequences $\theta = (\theta_1, \dots, \theta_n)$ in \mathbb{R}^n . Precisely, a prior Π_n on \mathbb{R}^n is constructed in three steps

- (P1) A dimension p is chosen according to a prior probability measure π_n on $\{0, 1, 2, \dots, n\}$.
- (P2) Given p a subset $S \subset \{0, 1, \dots, n\}$ of size $|S| = p$ is chosen at random.

(P3) Given (p, S) the coordinates of the vector $\theta_S = (\theta_i: i \in S)$ are chosen independently from a given Lebesgue density g on \mathbb{R} (if $p \geq 1$) and the remaining coordinates θ_{S^c} are set equal to 0.

Next Bayes' rule yields the *posterior distribution* $B \mapsto \Pi_n(B|X^n)$, the conditional distribution of θ given X^n if the conditional distribution of X given θ is taken equal to the normal distribution $N_n(\theta, I)$. The probability of a Borel set $B \subset \mathbb{R}^n$ under the posterior distribution can be decomposed as

$$\frac{\sum_{p=1}^n \pi_n(p) \binom{n}{p}^{-1} \sum_{|S|=p} \int_{(\theta_S, 0) \in B} \prod_{i \in S} \phi(X_i - \theta_i) \prod_{i \notin S} \phi(X_i) g_S(\theta_S) \prod_{i \in S} d\theta_i}{\sum_{p=1}^n \pi_n(p) \binom{n}{p}^{-1} \sum_{|S|=p} \int \prod_{i \in S} \phi(X_i - \theta_i) \prod_{i \notin S} \phi(X_i) g_S(\theta_S) \prod_{i \in S} d\theta_i}.$$

Here $(\theta_S, 0)$ is the vector in \mathbb{R}^n formed by adding coordinates $\theta_i = 0$ to $\theta_S = (\theta_i: i \in S)$, at the positions left open by $S \subset \{1, \dots, n\}$.

This formidable object is a random probability distribution on \mathbb{R}^n , which we study under the assumption that the vector $X^n = (X_1, \dots, X_n)$ is distributed according to a multivariate normal distribution with mean vector θ_0 and covariance matrix the identity.

The statistical problem of recovering θ_0 from X^n is equivariant in θ_0 , and hence the location of θ_0 in \mathbb{R}^n should not play a role in its recovery rate. However, a Bayesian procedure (with proper priors) necessarily favours certain regions of the parameter space. Depending on the choice of priors in (P3) this may lead to a shrinkage effect, yielding suboptimal behaviour for true parameters θ_0 that are far from the origin. This can be prevented by choosing priors with sufficiently heavy tails, for instance a product of Laplace densities. More generally, we assume that g has the form e^h for $h: \mathbb{R} \rightarrow \mathbb{R}$ satisfying

$$|h(x) - h(y)| \lesssim 1 + |x - y|, \quad \forall x, y \in \mathbb{R}. \tag{3}$$

This covers densities e^h with a uniformly Lipschitz function $h: \mathbb{R} \rightarrow \mathbb{R}$, such as Laplace and t -densities. It also covers densities of the form $e^{-|x|^\alpha}$ for some $\alpha \in (0, 1]$. However, the Gaussian density is excluded.

If the true parameter θ_0 is sparse, then one would hope that the posterior distribution concentrates on sparse vectors as well. The following theorem shows that this is true as soon as the priors π_n decrease exponentially.

Theorem 3.1. *If there exist constants $c < 1$ and $\bar{p}_n \rightarrow \infty$ such that $\pi_n(p) \leq c \pi_n(p - 1)$ for every $p \geq \bar{p}_n$, and the density in (P3) has finite second moment and is of the form $g = e^h$ for h satisfying (3), then, for a sufficiently large constant C ,*

$$\Pi_n(\theta: \#\{1 \leq i \leq n: \theta_i \neq 0\} \geq p_{0,n} + C\bar{p}_n | X^n) \xrightarrow{P} 0.$$

The theorem applies for instance to the geometric and Poisson distributions π_n (truncated to $\{0, 1, \dots, n\}$), with $\bar{p}_n \sim 1$. It shows that in these cases the posterior distribution will concentrate on the union of subspaces of dimension of the order the true number $p_{0,n}$ of nonzero coordinates. The following theorem shows that these priors also yield good concentration of the posterior near the true parameter.

Theorem 3.2. *Let π_n and the density in (P3) satisfy the conditions of the preceding theorem. If $r_n \rightarrow \infty$ is a sequence of numbers with $r_n^2 \geq Cp_{0,n} \log(n/p_{0,n}) \vee \log(1/\pi_n(Cp_{0,n}))$ for a sufficiently large constant C , then*

$$\Pi_n(\theta: \|\theta - \theta_0\| > r_n | X^n) \xrightarrow{P} 0.$$

For the truncated geometric and Poisson distributions π_n , the square rate of contraction r_n can be seen to be of the order $p_{0,n} \log(n/p_{0,n})$. Thus these priors are minimax optimal.

4. An Abstract Theorem

In this final section we present a general result on Bayesian regularization, formulated in the spirit of the general result on rates of contraction in [3]. We suppose that the data consists of a random sample X_1, \dots, X_n from a density p_0 relative to some given dominating measure μ on a given measurable space $(\mathcal{X}, \mathcal{A})$. Given a countable collection of sets of densities $\mathcal{P}_{n,\alpha}$, indexed by a parameter $\alpha \in A_n$, each provided with a prior distribution $\Pi_{n,\alpha}$, and a prior distribution $\lambda_n = (\lambda_{n,\alpha}: \alpha \in A_n)$ on A_n , we consider the posterior distribution relative to the prior that first chooses α according to λ_n and next p according to $\Pi_{n,\alpha}$ for the chosen α . Thus the overall prior is the mixture $\Pi_n = \sum_{\alpha \in A_n} \lambda_{n,\alpha} \Pi_{n,\alpha}$, and the corresponding posterior distribution is

$$\begin{aligned} \Pi_n(B|X_1, \dots, X_n) &= \frac{\int_B \prod_{i=1}^n p(X_i) d\Pi_n(p)}{\int \prod_{i=1}^n p(X_i) d\Pi_n(p)} \\ &= \frac{\sum_{\alpha \in A_n} \lambda_{n,\alpha} \int_{p \in \mathcal{P}_{n,\alpha}: p \in B} \prod_{i=1}^n p(X_i) d\Pi_{n,\alpha}(p)}{\sum_{\alpha \in A_n} \lambda_{n,\alpha} \int_{p \in \mathcal{P}_{n,\alpha}} \prod_{i=1}^n p(X_i) d\Pi_{n,\alpha}(p)}. \end{aligned} \tag{4}$$

To ensure that this expression is well defined, we assume that each collection $\mathcal{P}_{n,\alpha}$ of densities is equipped with a σ -field such that the maps $(x, p) \mapsto p(x)$ are jointly measurable.

We aim at a result of the following type. For a given p_0 there exists a “best” model \mathcal{P}_{n,β_n} that gives a posterior rate of contraction ε_{n,β_n} if it would be combined with the prior Π_{n,β_n} . The hierachical Bayesian procedure would *adapt* to the set of models if the posterior distributions (4) contract at the rate ε_{n,β_n} for this p_0 . We want this to be true for any p_0 in some model $\mathcal{P}_{n,\alpha}$. Obviously the weights λ_n play a crucial role for this.

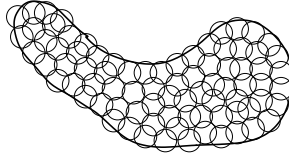


Figure 6. The covering number is the minimal number of balls of a certain radius needed to cover the model.

The result is formulated in terms of neighbourhoods of p_0 within $\mathcal{P}_{n,\alpha}$ of Kullback-Leibler and Hellinger types, given by

$$B_{n,\alpha}(\varepsilon) = \left\{ p \in \mathcal{P}_{n,\alpha} : -P_0 \log \frac{p}{p_0} \leq \varepsilon^2, P_0 \left(\log \frac{p}{p_0} \right)^2 \leq \varepsilon^2 \right\},$$

$$C_{n,\alpha}(\varepsilon) = \{ p \in \mathcal{P}_{n,\alpha} : h(p, p_0) \leq \varepsilon \}. \quad (5)$$

Here $h(p, q) = \|\sqrt{p} - \sqrt{q}\|_2$ is the Hellinger distance between the μ -densities p and q , the $L_2(\mu)$ -norm between their roots.

For β_n a given element of A_n , thought to be the index of a best model for a given fixed true density p_0 , we split the index set in the indices that give a faster or slower rate: for a fixed constant $H \geq 1$,

$$A_{n, \gtrsim \beta_n} := \left\{ \alpha \in A_n : \varepsilon_{n,\alpha}^2 \leq H \varepsilon_{n,\beta_n}^2 \right\},$$

$$A_{n, < \beta_n} := \left\{ \alpha \in A_n : \varepsilon_{n,\alpha}^2 > H \varepsilon_{n,\beta_n}^2 \right\}.$$

Even though we do not assume that A_n is ordered, we shall write $\alpha \gtrsim \beta_n$ and $\alpha < \beta_n$ if α belongs to the sets $A_{n, \gtrsim \beta_n}$ or $A_{n, < \beta_n}$, respectively. The set $A_{n, \gtrsim \beta_n}$ contains β_n and hence is never empty, but the set $A_{n, < \beta_n}$ can be empty (if β_n is the “smallest” possible index). In the latter case conditions involving $\alpha < \beta_n$ are understood to be automatically satisfied.

As in [3] we make assumptions on the complexity of the models and on the concentration of the priors. The complexity is measured by covering numbers. The ε -covering numbers of a metric space (\mathcal{P}, d) are denoted by $N(\varepsilon, \mathcal{P}, d)$, and are defined as the minimal numbers of balls of radius ε needed to cover \mathcal{P} (see Figure 6).

The complexity bound takes the form: for some constants E_α ,

$$\sup_{\varepsilon \geq \varepsilon_{n,\alpha}} \log N\left(\frac{\varepsilon}{3}, C_{n,\alpha}(2\varepsilon), h\right) \leq E_\alpha n \varepsilon_{n,\alpha}^2, \quad \alpha \in A_n. \quad (6)$$

This inequality may actually be read as a definition of a rate $\varepsilon_{n,\alpha}$, which could be set equal to the smallest value that satisfies the inequality. This definition has its roots in the work of Lucien le Cam and has nothing to do with Bayesian methods. Essentially $\varepsilon_{n,\alpha}$ corresponds to the maximal precision of estimation

that can be obtained for the model $\mathcal{P}_{n,\alpha}$ by any statistical method, Bayesian or non-Bayesian. More complex models have larger covering numbers and therefore larger rates $\varepsilon_{n,\alpha}$.

The conditions on the priors involve comparisons of the prior masses of balls of various sizes in various models. These conditions are split in conditions on the models that are smaller or bigger than the best model: for given constants $\mu_{n,\alpha}, L, H, I$,

$$\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}} \frac{\Pi_{n,\alpha}(C_{n,\alpha}(i\varepsilon_{n,\alpha}))}{\Pi_{n,\beta_n}(B_{n,\beta_n}(\varepsilon_{n,\beta_n}))} \leq \mu_{n,\alpha} e^{Li^2 n \varepsilon_{n,\alpha}^2}, \quad \alpha < \beta_n, \quad i \geq I, \quad (7)$$

$$\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}} \frac{\Pi_{n,\alpha}(C_{n,\alpha}(i\varepsilon_{n,\beta_n}))}{\Pi_{n,\beta_n}(B_{n,\beta_n}(\varepsilon_{n,\beta_n}))} \leq \mu_{n,\alpha} e^{Li^2 n \varepsilon_{n,\beta_n}^2}, \quad \alpha \gtrsim \beta_n, \quad i \geq I. \quad (8)$$

A final condition requires that the prior mass in a ball of radius $\varepsilon_{n,\alpha}$ in a big model (i.e. small α) is significantly smaller than in a small model: for some constants I, B ,

$$\sum_{\alpha \in A_n: \alpha < \beta_n} \frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}} \frac{\Pi_{n,\alpha}(C_{n,\alpha}(IB\varepsilon_{n,\alpha}))}{\Pi_{n,\beta_n}(B_{n,\beta_n}(\varepsilon_{n,\beta_n}))} = o(e^{-2n\varepsilon_{n,\beta_n}^2}). \quad (9)$$

Theorem 4.1. *Assume there exist positive constants $B, E_\alpha, L, H \geq 1, I > 2$ such that (6), (7), (8) and (9) hold, and, constants E and \underline{E} such that $E \geq \sup_{\alpha \in A_n: \alpha \gtrsim \beta_n} E_\alpha \varepsilon_{n,\alpha}^2 / \varepsilon_{n,\beta_n}^2$ and $\underline{E} \geq \sup_{\alpha \in A_n: \alpha < \beta_n} E_\alpha$ (with $\underline{E} = 0$ if $A_{n,<\beta_n} = \emptyset$),*

$$B > \sqrt{H}, \quad B^2/9 > (H\underline{E}) \vee E + 1, \quad B^2 I^2 (1/9 - 2L) > 3.$$

Furthermore, assume that $\sum_{\alpha \in A_n} \sqrt{\mu_{n,\alpha}} \leq \exp(n\varepsilon_{n,\beta_n}^2)$. If $\beta_n \in A_n$ for every n and satisfies $n\varepsilon_{n,\beta_n}^2 \rightarrow \infty$, then the posterior distribution (4) satisfies

$$\Pi_n \left(p: h(p, p_0) \geq IB\varepsilon_{n,\beta_n} | X_1, \dots, X_n \right) \xrightarrow{P} 0.$$

In many situations relatively crude bounds on the prior mass bounds (7), (8) and (9) are sufficient. In particular, the following lower bound is often useful: for a positive constant F ,

$$\Pi_{n,\beta_n}(B_{n,\beta_n}(\varepsilon_{n,\beta_n})) \geq \exp[-Fn\varepsilon_{n,\beta_n}^2]. \quad (10)$$

This correspond to the ‘‘crude’’ prior mass condition of [3]. Combined with the trivial bound 1 on the probabilities $\Pi_{n,\alpha}(C)$ in (7) and (8), we see that these conditions hold (for sufficiently large I) if, for all $\alpha \in A_n$,

$$\frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}} \leq \mu_{n,\alpha} e^{n(\varepsilon_{n,\alpha}^2 \vee \varepsilon_{n,\beta_n}^2)}.$$

This appears to be a mild requirement. On the other hand, the similarly adapted version of condition (9) still requires that

$$\sum_{\alpha \in A_n: \alpha < \beta_n} \frac{\lambda_{n,\alpha}}{\lambda_{n,\beta_n}} \Pi_{n,\alpha}(C_{n,\alpha}(IB\varepsilon_{n,\alpha})) = o(e^{-(F+2)n\varepsilon_{n,\beta_n}^2}).$$

Such a condition may be satisfied because the prior probabilities $\Pi_{n,\alpha}(C_{n,\alpha}(IB\varepsilon_{n,\alpha}))$ are very small. For instance, a reverse bound of the type (10) for α instead of β_n would yield this type of bound for fairly general model weights $\lambda_{n,\alpha}$, since $\varepsilon_{n,\alpha} \geq H\varepsilon_{n,\beta_n}$ for $\alpha < \beta_n$. Alternatively, the condition could be forced by choice of the model weights $\lambda_{n,\alpha}$, for general priors $\Pi_{n,\alpha}$. For instance, weights of the type

$$\lambda_{n,\alpha} \propto \mu_\alpha \exp[-Cn\varepsilon_{n,\alpha}^2]$$

satisfy all conditions. In [6] they are applied to several examples of interest. Note that they strongly downweight big models (with large $\varepsilon_{n,\alpha}$) relative to small models.

The posterior distribution (4) can be viewed as a mixture of the posterior distributions on the various models, with the weights given by the posterior distribution of the model index. Our second result shows that models that are “bigger” than the optimal model asymptotically have vanishing zero posterior mass and hence zero weight in this mixture.

Theorem 4.2. *Under the conditions of Theorem 4.1,*

$$\begin{aligned} \Pi_n(\alpha \in A_{n,<\beta_n} | X_1, \dots, X_n) &\xrightarrow{P} 0, \\ \Pi_n(\alpha \in A_{n,\gtrsim\beta_n} : h(p_0, \mathcal{P}_{n,\alpha}) > IB\varepsilon_{n,\beta_n} | X_1, \dots, X_n) &\xrightarrow{P} 0. \end{aligned}$$

The first assertion of the theorem is pleasing. It can be interpreted in the sense that the models that are bigger than the model \mathcal{P}_{n,β_n} that contains the true distribution eventually receive negligible posterior weight. The second assertion makes a similar claim about the smaller models, but it is restricted to the smaller models that keep a certain distance to the true distribution. Such a restriction appears not unnatural, as the posterior looks at the data through the likelihood and hence will judge a model by its approximation properties rather than its parametrization. That big models with similarly good approximation properties are not favoured is caused by the fact that (under our conditions) the prior mass on the big models is more spread out, yielding relatively little prior mass near good approximants within the big models.

References

- [1] Carl Edward Rasmussen and Christopher K.I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, Cambridge, Massachusetts, 2006.

-
- [2] I. Castillo and A.W. van der Vaart. Needles and straws in a haystack: posterior concentration for possibly sparse sequences. *Preprint*, 2010.
 - [3] Subhashis Ghosal, Jayanta K. Ghosh, and Aad W. van der Vaart. Convergence rates of posterior distributions. *Ann. Statist.*, 28(2):500–531, 2000.
 - [4] Subhashis Ghosal, Jüri Lember, and Aad van der Vaart. Nonparametric Bayesian model selection and averaging. *Electron. J. Stat.*, 2:63–89, 2008.
 - [5] GOLUBEV, G. K. Reconstruction of sparse vectors in white Gaussian noise. *Problemy Peredachi Informatsii* 38, 1 (2002), 75–91.
 - [6] Jüri Lember and Aad van der Vaart. On universal Bayesian adaptation. *Statist. Decisions*, 25(2):127–152, 2007.
 - [7] A. W. van der Vaart and J. H. van Zanten. Rates of contraction of posterior distributions based on Gaussian process priors. *Ann. Statist.*, 36(3):1435–1463, 2008.
 - [8] Aad van der Vaart and Harry van Zanten. Bayesian inference with rescaled Gaussian process priors. *Electron. J. Stat.*, 1:433–448 (electronic), 2007.
 - [9] A. W. van der Vaart and J. H. van Zanten. Reproducing kernel Hilbert spaces of Gaussian priors. *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, ser. Inst. Math. Stat. Collect. Beachwood, OH: Inst. Math. Statist., vol. 3, 200–222, 2008.
 - [10] A. W. van der Vaart and J. H. van Zanten. Adaptive Bayesian estimation using a Gaussian random field with inverse gamma bandwidth. *Ann. Statist.*, 37(5B):2655–2675, 2009.
 - [11] A. W. van der Vaart and J. H. van Zanten. Information Rates of Nonparametric Gaussian Process Methods. *Preprint.*, 2010.

This page is intentionally left blank

Section 14

Combinatorics

This page is intentionally left blank

Flag Enumeration in Polytopes, Eulerian Partially Ordered Sets and Coxeter Groups

Louis J. Billera*

Abstract

We discuss the enumeration theory for flags in Eulerian partially ordered sets, emphasizing the two main geometric and algebraic examples, face posets of convex polytopes and regular CW -spheres, and Bruhat intervals in Coxeter groups. We review the two algebraic approaches to flag enumeration – one essentially as a quotient of the algebra of noncommutative symmetric functions and the other as a subalgebra of the algebra of quasisymmetric functions – and their relation via duality of Hopf algebras. One result is a direct expression for the Kazhdan-Lusztig polynomial of a Bruhat interval in terms of a new invariant, the complete \mathbf{cd} -index. Finally, we summarize the theory of combinatorial Hopf algebras, which gives a unifying framework for the quasisymmetric generating functions developed here.

Mathematics Subject Classification (2010). Primary 06A11; Secondary 05E05, 16T30, 20F55, 52B11.

Keywords. Convex polytope, Eulerian poset, Coxeter group, Kazhdan-Lusztig polynomial, \mathbf{cd} -index, quasisymmetric function, Hopf algebra

1. Introduction: Face Enumeration in Convex Polytopes

We begin with an introduction to the enumeration of faces in convex polytopes. For a d -dimensional convex polytope Q , let $f_i = f_i(Q)$ be the number of i -dimensional faces of Q . Thus f_0 is the number of vertices, f_1 the number of

*Partially supported by the US National Science Foundation grant DMS-0555268.
Department of Mathematics, Malott Hall, Cornell University, Ithaca, NY 14850-4201 USA.
E-mail: billera@math.cornell.edu.

edges, \dots , f_{d-1} the number of *facets* (or defining inequalities) of Q . The f -vector of Q is the vector

$$f(Q) = (f_0, f_1, \dots, f_{d-1}).$$

The central problem of this area is to determine when a vector of nonnegative integers $f = (f_0, f_1, \dots, f_{d-1})$ is $f(Q)$ for some d -polytope Q . The case $d = 2$ is clear ($f_0 = f_1 \geq 3$); $d = 3$ was settled by Steinitz in 1906 [54]. The cases $d = 4$ and higher remain open except for special classes of polytopes.

1.1. Simplicial polytopes. A polytope is *simplicial* if all faces are simplices, for example, if its vertices are in general position. Their duals are the *simple* polytopes, which include polytopes with facets in general position. If Q and Q^* are dual d -dimensional polytopes, then their f -vectors are related by $f_i(Q) = f_{d-1-i}(Q^*)$. The f -vectors of simplicial (and, consequently, simple) polytopes have been completely determined.

The h -vector (h_0, \dots, h_d) of a simplicial d -polytope is defined by the polynomial relation

$$\sum_{i=0}^d h_i x^{d-i} = \sum_{i=0}^d f_{i-1} (x-1)^{d-i}. \quad (1)$$

The h -vector and the f -vector of a polytope mutually determine each other via the formulas

$$h_i = \sum_{j=0}^i (-1)^{i-j} \binom{d-j}{i-j} f_{j-1} \quad \text{and} \quad f_{i-1} = \sum_{j=0}^i \binom{d-j}{i-j} h_j,$$

for $0 \leq i \leq d$, so characterizing f -vectors of simplicial polytopes is equivalent to characterizing their h -vectors. This is done in the so-called *g-theorem*, conjectured by McMullen [42] and proved by Billera and Lee [17, 18] (for the sufficiency of the conditions) and Stanley [47] (for their necessity). Given (h_0, \dots, h_d) , define $g_0 := h_0$ and $g_i := h_i - h_{i-1}$ for $1 \leq i \leq \lfloor \frac{d}{2} \rfloor$.

Theorem 1.1 (*g-theorem*). $(h_0, h_1, \dots, h_d) \in \mathbb{Z}^{d+1}$ is the h -vector of a simplicial convex d -polytope if and only if

1.1.1. $h_i = h_{d-i}$, for all i ,

1.1.2. $g_0 = 1$, $g_i \geq 0$, for $1 \leq i \leq \lfloor \frac{d}{2} \rfloor$, and

1.1.3. $g_{i+1} \leq g_i^{\langle i \rangle}$ for $i \geq 1$.

The relations in (1.1.1) are known as the *Dehn-Sommerville equations* and date to the early 20th century. The nonnegativity relations (1.1.2) are known as the *generalized lower bound conditions*. These plus the inequalities (1.1.3) are

known as the *Macaulay conditions*. The quantity $g_i^{(i)}$ is computed by expressing g_i canonically as the sum of a sequence of binomial coefficients and altering them by adding 1 to the top and bottom of each. See [18] for details.

Conditions (1.1.2) and (1.1.3) characterize sequences of natural numbers that count monomials in an order ideal of monomials (a set of monomials closed under the division order). They are similar to, but are not quite the same as, the conditions of Kruskal and Katona for f -vectors of general simplicial complexes, but with g_i in place of f_i . Equivalently, (1.1.2) and (1.1.3) say the g_i 's form the Hilbert function of some graded algebra. The necessity proof of Stanley [47] proceeds by producing a commutative ring with this Hilbert function. See, for example, [11] for complete definitions and references.

1.2. Counting flags in polytopes. For general convex polytopes, the situation for f -vectors is much less satisfactory. In particular, the only equation they all satisfy is the Euler relation

$$f_0 - f_1 + f_2 - \cdots \pm f_{d-1} = 1 - (-1)^d.$$

Already in $d = 4$, we do not know all linear inequalities on f -vectors, and at this point, there is little hope of giving an analog to the Macaulay conditions.

A possible solution is to try to solve a harder problem: count not *faces*, but *chains* of faces. For a d -dimensional polytope Q and a set S of possible dimensions, define $f_S(Q)$ to be the number of chains of faces of Q having dimensions prescribed by the set S . The function

$$S \mapsto f_S(Q)$$

is called the flag f -vector of Q . It was first studied by Stanley in the context of balanced simplicial complexes, a natural extension of order complexes of graded posets [46].

The flag f -vector of a polytope includes the ordinary f -vector, by counting chains of one element: ($f_S : |S| = 1$). It also has a straightforwardly defined flag h -vector that turns out to be a finely graded Hilbert function. Most important for an algebraic approach to flag f -vectors, they satisfy an analog of the Dehn-Sommerville equations, which cut their dimension down to the Fibonacci numbers, compared to $\lfloor \frac{n}{2} \rfloor$ for f -vectors of simplicial polytopes.

In what follows, we discuss the development of the theory of flag vectors of polytopes, and where it has led. We thank Margaret Bayer, Saúl Blanco and Stephanie van Willigenburg for reading and offering corrections on earlier drafts of this paper.

2. Eulerian Posets and the cd-index

The best setting in which to study the flag f -vector of a d -polytope Q is that of its lattice of faces $P = \mathcal{F}(Q)$, an Eulerian graded poset of rank $d + 1$. We define

the **cd**-index and g -polynomials for Eulerian posets and discuss inequalities on these for polytopes and certain spherical subdivisions.

2.1. Flag enumeration in graded posets. A *graded* poset is a poset P with elements $\widehat{0}$ and $\widehat{1}$ such that $\widehat{0} \leq x \leq \widehat{1}$ for all $x \in P$ and with rank function $\rho : P \rightarrow \mathbb{N}$ so that for each $x \in P$, $\rho(x)$ is the length k of any maximal chain $\widehat{0} = x_0 < x_1 < \dots < x_k = x$. The *rank* of P is $\rho(P) := \rho(\widehat{1})$.

The *flag f -vector* of a graded poset P of rank $n + 1$ is the function $S \mapsto f_S = f_S(P)$, where for $S = \{i_1, \dots, i_k\} \subseteq [n] := \{1, \dots, n\}$,

$$f_S = \left| \{ y_1 < y_2 < \dots < y_k \mid y_j \in P, \rho(y_j) = i_j \} \right|.$$

Included is the case $S = \emptyset$, where usually $f_\emptyset = 1$, although later we will let f_\emptyset be unspecified.

To begin to understand flag f -vectors of convex polytopes, it might be helpful first to be able to determine all flag f -vectors of graded posets, or at least determine all linear inequalities satisfied by flag f -vectors of graded posets. The former is an analog of the Kruskal-Katona conditions on f -vectors of simplicial complexes and remains open. The latter are analogs of the Dehn-Sommerville and generalized lower bound relations for graded posets. They are completely determined.

First, it is easy to determine that there are no linear equations that hold for the flag f -vectors of all graded posets [19, Proposition 1.1]. For inequalities, the situation is more interesting. For example, for graded posets of rank 4, it can be shown that the inequality

$$f_{\{1,3\}} - f_{\{1\}} + f_{\{2\}} - f_{\{3\}} \geq 0$$

always holds [15, Example 3].

More generally, a subset of the form $\{i, i + 1, \dots, i + k\} \in [n]$ is called an *interval*. For an antichain of intervals $\mathcal{I} \subset 2^{[n]}$, define the *blocking family*

$$b[\mathcal{I}] = \{S \subseteq [n] \mid S \cap I \neq \emptyset, \forall I \in \mathcal{I}\}.$$

Theorem 2.1 ([15]). *A linear form $\sum_{S \subseteq [n]} a_S f_S$ satisfies $\sum_S a_S f_S(P) \geq 0$ for all graded posets P of rank $n + 1$ if and only if for all antichains of intervals $\mathcal{I} \subset 2^{[n]}$,*

$$\sum_{S \in b[\mathcal{I}]} a_S \geq 0.$$

Corollary. *The closed convex cone generated by all flag f -vectors of graded posets is polyhedral and has the (Catalan many) extreme rays $e_{\mathcal{I}} = \sum_{S \in b[\mathcal{I}]} e_S$, where $\{e_S \mid S \subseteq [n]\}$ are the unit vectors in \mathbb{R}^{2^n} .*

Example 1. We consider the case of graded posets of rank 3. The flag f -vector in this case is the vector $f = (f_\emptyset, f_{\{1\}}, f_{\{2\}}, f_{\{1,2\}})$, and there are 5 extreme rays corresponding to 5 antichains of intervals.

\mathcal{I}	\emptyset	$\{\{1, 2\}\}$	$\{\{1\}, \{2\}\}$	$\{\{1\}\}$	$\{\{2\}\}$
$e_{\mathcal{I}}$	$(1, 1, 1, 1)$	$(0, 1, 1, 1)$	$(0, 0, 0, 1)$	$(0, 1, 0, 1)$	$(0, 0, 1, 1)$

2.2. Eulerian posets and the cd-index. A graded poset P is said to be *Eulerian* if for all $x \leq y \in P$,

$$\mu(x, y) = (-1)^{\rho(y) - \rho(x)}$$

where μ is the Möbius function of P . Equivalently, P is Eulerian if for each subinterval $[x, y] \subseteq P$, the number of elements of even rank is equal to number of elements of odd rank. Face posets of polytopes and spheres are Eulerian.

Again, two natural problems arise, to determine all flag f -vectors of Eulerian posets or, at least, to determine all linear inequalities satisfied by flag f -vectors of Eulerian posets. Here, all the linear equations are known. There are 2^n flag numbers $f_S, S \subseteq [n]$, for graded posets of rank $n + 1$. For Eulerian posets, these are not independent evaluations. In fact, for Eulerian posets, only Fibonacci many f_S are linearly independent.

We consider the first few cases. Note that we consider f_\emptyset to be variable, which will be important later for several reasons.

$n = 0$: f_\emptyset is the only flag number.

$n = 1$: $f_\emptyset, f_{\{1\}}$ are the relevant flag numbers, but $f_{\{1\}} = 2f_\emptyset$ (Euler relation).

$n = 2$: $f_\emptyset, f_{\{1\}}, f_{\{2\}}, f_{\{1,2\}}$ are all the flag numbers, but $f_{\{1\}} = f_{\{2\}}$ (Euler relation) and $f_{\{1,2\}} = 2f_{\{2\}}$.

$n = 3$: $f_\emptyset, f_{\{1\}}, f_{\{2\}}, f_{\{3\}}, f_{\{1,2\}}, f_{\{1,3\}}, f_{\{2,3\}}, f_{\{1,2,3\}}$ are the flag numbers, but $f_{\{1\}} - f_{\{2\}} + f_{\{3\}} = 2f_\emptyset$ (Euler relation), $f_{\{1,2\}} = 2f_{\{2\}}, f_{\{2,3\}} = 2f_{\{2\}}, f_{\{1,3\}} = f_{\{2,3\}}$ and $f_{\{1,2,3\}} = 2f_{\{1,3\}}$. Only $f_\emptyset, f_{\{1\}}$ and $f_{\{2\}}$ are independent.

$n = 4$: Only $f_\emptyset, f_{\{1\}}, f_{\{2\}}, f_{\{3\}}, f_{\{1,3\}}$ are independent.

Thus the first few dimensions of the linear space spanned by all flag numbers of Eulerian posets of rank $n + 1$ are 1, 1, 2, 3 and 5. The relevant relations for P are all derived from Euler relations in P and in intervals $[x, y]$ of P . Details of these equations will appear later.

There is much less known about inequalities for flag numbers of Eulerian posets. The cones of all flag vectors are known for Eulerian posets through rank 6. The best references for this are [8, 9].

For $S \subseteq [n]$ let the *flag h -vector* be defined by

$$h_S = \sum_{T \subseteq S} (-1)^{|S|-|T|} f_T.$$

For noncommuting indeterminates \mathbf{a} and \mathbf{b} let $u_S = u_1 u_2 \cdots u_n$ be defined by

$$u_i = \begin{cases} \mathbf{b} & \text{if } i \in S \\ \mathbf{a} & \text{if } i \notin S. \end{cases}$$

Let $\mathbf{c} = \mathbf{a} + \mathbf{b}$ and $\mathbf{d} = \mathbf{ab} + \mathbf{ba}$. Then for Eulerian posets, the generating function

$$\Psi_P = \sum_S h_S(P) u_S \tag{2}$$

is always a polynomial in \mathbf{c} and \mathbf{d} ; this polynomial $\Phi_P(\mathbf{c}, \mathbf{d})$ is called the *\mathbf{cd} -index* of P . This invariant was first explicitly defined by Bayer and Klapper in [6], following an unpublished suggestion of J. Fine.

Example 2. Let $P = \mathcal{B}_3$, the Boolean algebra of rank 3, *i.e.*, the poset of all subsets of a 3-element set ordered by inclusion. We have $f_\emptyset = 1$, $f_{\{1\}} = 3$, $f_{\{2\}} = 3$, and $f_{\{1,2\}} = 6$ so $h_\emptyset = 1$, $h_{\{1\}} = 2$, $h_{\{2\}} = 2$, $h_{\{1,2\}} = 1$, and

$$\begin{aligned} \Psi_P &= \mathbf{aa} + 2\mathbf{ba} + 2\mathbf{ab} + \mathbf{bb} \\ &= (\mathbf{a} + \mathbf{b})^2 + (\mathbf{ab} + \mathbf{ba}) \\ &= \mathbf{c}^2 + \mathbf{d} = \Phi_P \end{aligned}$$

Another invariant for Eulerian posets that implicitly enumerates flags is the following extension of the h -vector and associated g -vector defined in §1.1. This definition, originally due to MacPherson in the context of convex polytopes and their associated toric varieties, was given in the context of Eulerian posets by Stanley in [48]. For an Eulerian poset P of rank $n + 1 \geq 0$, we define two polynomials $f(P, x), g(P, x) \in \mathbb{Z}[x]$ recursively as follows. If $n + 1 = 0$, then $f(P, x) = g(P, x) = 1$. If $n + 1 > 0$, then

$$f(P, x) = \sum_{y \in P \setminus \{\hat{1}\}} g([\hat{0}, y], x) (x - 1)^{n - \rho(y)}. \tag{3}$$

If $f(P, x) = \sum_{i=0}^n \kappa_i x^i$ has been defined, then we define

$$g(P, x) = \kappa_0 + (\kappa_1 - \kappa_0) x + \cdots + \left(\kappa_{\lfloor \frac{n}{2} \rfloor} - \kappa_{\lfloor \frac{n}{2} \rfloor - 1} \right) x^{\lfloor \frac{n}{2} \rfloor}. \tag{4}$$

For an Eulerian poset P , the vector $(h_0, \dots, h_n) = (\kappa_n, \dots, \kappa_1, \kappa_0)$ is what is sometimes called the *toric h -vector* of P . When P is the face poset of a simplicial polytope (or any simplicial complex), this toric h -vector coincides with the usual simplicial h -vector defined in (1). Since for Eulerian P , $h_i = h_{n-i}$ (see [48] or

[51, Theorem 3.14.9]), our definition of $g(P, x)$ agrees with the usual notion of the simplicial g -vector.

That the toric h and g -vectors are functions of the flag f -vector was first noted by Bayer [3]. Formulas expressing these in terms of the flag f -vector (for general graded posets) and the **cd**-index (for Eulerian posets) are given in [7]. We note that in [7], this distinction between κ_i and h_i is not made, so their formulas for h_i are, in reality, for h_{n-i} (which equals h_i in the Eulerian case).

2.3. Inequalities for flags in polytopes and spheres. There are by now many inequalities known to hold for the g -polynomial and the **cd**-index of convex polytopes and more general spheres. These all give inequalities on the flag f -vectors of these objects. We summarize most of these here.

- Among all n -dimensional polytopes, the g -polynomial is termwise minimized on the n -simplex Δ_n . Since always $g_0 = 1$, this is equivalent to saying that $g_i \geq 0$ for $i \geq 1$ (the generalized lower bound theorem). This was proved by Stanley in [47] and [48] for simplicial and then all rational polytopes using the cohomology of toric varieties, and extended to all polytopes by Karu [37], by means of the theory of combinatorial intersection cohomology. See [21] or [53] for a discussion of this combinatorial cohomology theory.

- For polytopes and, in fact, for all Cohen-Macaulay graded posets (so for face posets of balls and spheres), $h_S \geq 0$ (Stanley, [46]).

- If we write $\Phi_P = \sum_w [w]_P w$ over **cd**-words w , then $[w]_P \geq 0$ for polytopes (more generally for S -shellable CW -spheres; Stanley [49]).

- Among all n -dimensional zonotopes, the **cd**-index is termwise minimized on the n -cube C_n . Equivalently, among all decompositions of the $(n - 1)$ -sphere induced by central hyperplane arrangements in \mathbb{R}^n , the **cd**-index is termwise minimized by the n -dimensional crosspolytope (Billera, Ehrenborg and Readdy [14]).

- Among all n -dimensional polytopes, the **cd**-index is termwise minimized on the n -simplex Δ_n (Billera and Ehrenborg [13]).

- If Q is a polytope, then termwise as polynomials

$$g(Q) \geq g(F) \cdot g(Q/F)$$

for any any face $F \subseteq Q$, where Q/F is any polytope whose face lattice is the interval $[F, Q]$. This was shown by Braden and MacPherson [22] for rational polytopes using cohomology of toric varieties. Again, it follows for all polytopes by combinatorial intersection cohomology; see [21] for a discussion of this.

- For *any* polytope Q and face $F \subseteq Q$, we have termwise as **cd**-polynomials,

$$\Phi_Q \geq \begin{cases} \mathbf{c} \cdot \Phi_F \cdot \Phi_{Q/F} \\ \Phi_F \cdot \mathbf{c} \cdot \Phi_{Q/F} \\ \Phi_F \cdot \Phi_{Q/F} \cdot \mathbf{c}, \end{cases}$$

where $\Phi_Q, \Phi_F, \Phi_{Q/F}$ are the **cd**-indices of (the face lattices of) $Q, F, P/F$, respectively (Billera and Ehrenborg [13]).

• For a polytope Q , let $[w]_Q$ denote the coefficient of the **cd**-word w in the **cd**-index of Q . Then for all **cd**-words u and v

$$[u\mathbf{d}v]_Q \geq [u\mathbf{c}^2v]_Q$$

(Ehrenborg [31]).

• If Q is an n -dimensional polytope with v vertices, then for any S ,

- (a) $f_S(Q) \leq f_S(C(v, n))$,
- (b) $h_S(Q) \leq h_S(C(v, n))$ and
- (c) $\Phi_Q \leq \Phi_{C(v, n)}$,

where $C(v, n)$ is the cyclic n -polytope with v vertices, *i.e.*, the convex hull of v points on the moment curve (t, t^2, \dots, t^n) . This is known as the Upper Bound Theorem. The first inequality for the case $|S| = 1$ was proved by McMullen [41] by proving the first two inequalities for all simplicial polytopes in this case. The latter result was extended to all triangulated spheres by Stanley [45]. The first inequality for general S was observed by Bayer and Billera [4]. In full generality, this result is due to Billera and Ehrenborg [13].

• For P a Gorenstein* poset (*i.e.*, one that is both Eulerian and Cohen-Macaulay), $\Phi_P \geq 0$. Gorenstein* posets include all face-posets of regular CW -spheres. This result was conjectured by Stanley in [49] and proved by Karu [38, 39].

• For P a Gorenstein* lattice of rank $n + 1$, Φ_P is bounded below termwise by the **cd**-index of the n -dimensional simplex. This generalizes the result of Billera-Ehrenborg for **cd**-indices of n -dimensional polytopes. This result was also conjectured by Stanley [50] and was proved by Ehrenborg and Karu [32].

There is one outstanding conjecture of Stanley in this area that remains open. What follows is Conjecture 4.2(d) in [48]. The second part is Conjecture 4.3 in [50]. It covers, in particular, g -polynomials of all triangulated spheres. (That the h -polynomial of a triangulated sphere is nonnegative is a consequence of the Cohen-Macaulayness of its face ring [45].)

Conjecture 1 ([48]). For P a Gorenstein* lattice, the g -polynomial, and so the h -polynomial, is nonnegative.

We should note here that there is no guarantee in any of these cases that there are only finitely many irredundant linear inequalities, although in none of these cases have more than finitely many been found. In a related instance, however, Nyman [43] has found that for rank 3 geometric lattices, countably many linear inequalities are necessary to describe their flag f -vectors.

3. Algebraic Approaches to Counting Flags

In this section, we will consider two different algebras that arise in the study of flag f -vectors of graded posets. In the end, we will see that these algebras are, in fact, directly related to each other via duality of Hopf algebras. Especially interesting is how each one handles the case of Eulerian posets.

3.1. The convolution product and derived inequalities. We will write $f_S^{(n)}, S \subseteq [n - 1]$, when counting chains in a poset of rank n , and we consider $f_S^{(n)}(\cdot)$ to be an operator on posets of rank n . Alternatively, we can define $f_S^{(n)}(P) \equiv 0$ when the rank of P is not n .

Given $f_S^{(n)}$ and $f_T^{(m)}, S \subseteq [n - 1], T \subseteq [m - 1]$ and P a poset of rank $n + m$, define

$$f_S^{(n)} * f_T^{(m)}(P) = \sum_{x \in P : \rho(x)=n} f_S^{(n)}([\widehat{0}, x]) \cdot f_T^{(m)}([x, \widehat{1}]).$$

It is easy to see that $f_S^{(n)} * f_T^{(m)} = f_{S \cup \{n\} \cup (T+n)}^{(n+m)}$, where $T+n := \{x+n \mid x \in T\}$.

For example, $f_{\{1\}}^{(2)} * f_{\{2\}}^{(3)} = f_{\{1,2,4\}}^{(5)}$ and $f_{\emptyset}^{(2)} * f_{\emptyset}^{(3)} = f_{\{2\}}^{(5)}$.

This convolution product was first considered by Kalai [36], who used it to produce new flag vector inequalities for polytopes from known ones. It is immediate, that this works as well for graded posets or for Eulerian posets (in fact, for any class of posets closed under taking intervals).

Proposition 3.1 ([36]). *If the linear forms $G_1 = \sum \alpha_S f_S^{(n)}$ and $G_2 = \sum \beta_S f_S^{(m)}$ satisfy $G_1(P_1) \geq 0$ and $G_2(P_2) \geq 0$ for all polytopes (respectively, graded posets, Eulerian posets) P_1 and P_2 of ranks n and m , then $G_1 * G_2(P) \geq 0$ for all polytopes (graded posets, Eulerian posets) P of rank $n + m$.*

Example 3. Polygons have at least 3 vertices, so $f_{\{1\}}^{(3)} - 3f_{\emptyset}^{(3)} \geq 0$ for all polygons. (Note that rank is one more than dimension, so $f_1^{(3)}$ counts *vertices*.) Thus

$$\left(f_{\{1\}}^{(3)} - 3f_{\emptyset}^{(3)} \right) * f_{\emptyset}^{(1)} = f_{\{1,3\}}^{(4)} - 3f_{\{3\}}^{(4)} \geq 0$$

for all 3-polytopes (*i.e.*, the number of vertices in 2-faces is at least three times the number of 2-faces).

Most of the inequalities described earlier are of the form

$$G(P) = \sum \alpha_S f_S^{(n)}(P) \geq 0$$

and so can be convolved to give further inequalities. As an example we consider the coefficients of the **cd**-index. Let $w = \mathbf{c}^{n_1} \mathbf{d} \mathbf{c}^{n_2} \mathbf{d} \mathbf{c}^{n_3} \dots \mathbf{c}^{n_p} \mathbf{d} \mathbf{c}^{n_{p+1}}$ be a **cd**-word, and define m_0, \dots, m_p by $m_0 = 1$ and $m_i = m_{i-1} + n_i + 2$. Then the coefficient of w in the **cd**-index is given by

$$[w] = \sum_{i_1, \dots, i_p} (-1)^{(m_1-i_1)+(m_2-i_2)+\dots+(m_p-i_p)} k_{\{i_1 i_2 \dots i_p\}}, \tag{5}$$

where the sum is over all p -tuples (i_1, i_2, \dots, i_p) such that $m_{j-1} \leq i_j \leq m_j - 2$ and

$$k_S = \sum_{T \subseteq S} (-2)^{|S|-|T|} f_T.$$

Using (5), we can see the **cd**-indices for Eulerian posets of ranks 1–5 are

$$\begin{aligned} &f_\emptyset^{(1)} \\ &f_\emptyset^{(2)} \mathbf{c} \\ &f_\emptyset^{(3)} \mathbf{c}^2 + (f_{\{1\}}^{(3)} - 2f_\emptyset^{(3)}) \mathbf{d} \\ &f_\emptyset^{(4)} \mathbf{c}^3 + (f_{\{1\}}^{(4)} - 2f_\emptyset^{(4)}) \mathbf{dc} + (f_{\{2\}}^{(4)} - f_{\{1\}}^{(4)}) \mathbf{cd} \\ &f_\emptyset^{(5)} \mathbf{c}^4 + (f_{\{1\}}^{(5)} - 2f_\emptyset^{(5)}) \mathbf{dc}^2 + (f_{\{2\}}^{(5)} - f_{\{1\}}^{(5)}) \mathbf{cdc} + (f_{\{3\}}^{(5)} - f_{\{2\}}^{(5)} + f_{\{1\}}^{(5)} - 2f_\emptyset^{(5)}) \mathbf{c}^2 \mathbf{d} \\ &\quad + (f_{\{1,3\}}^{(5)} - 2f_{\{3\}}^{(5)} - 2f_{\{1\}}^{(5)} + 4f_\emptyset^{(5)}) \mathbf{d}^2, \end{aligned}$$

so, for example, we know from the nonnegativity of the **cd**-index that

$$f_{\{1,3\}}^{(5)} - 2f_{\{3\}}^{(5)} - 2f_{\{1\}}^{(5)} + 4f_\emptyset^{(5)} \geq 0$$

for all 4-dimensional convex polytopes.

We remark here that Stenson [56] has shown that the set of inequalities on polytopes derived by convolution from the nonnegativity of the g_i and the set derived from the fact that $[w]$ is bounded below by its value on the simplex do not imply each other.

3.2. Relations on flag numbers and the enumeration algebra. Eulerian posets of rank d , as well as polytopes of dimension $d - 1$, satisfy the Euler relations

$$f_\emptyset^{(d)} - f_{\{1\}}^{(d)} + f_{\{2\}}^{(d)} - \dots + (-1)^{d-1} f_{\{d-1\}}^{(d)} + (-1)^d f_\emptyset^{(d)} = 0.$$

Since by Proposition 3.1, the convolution product preserves equalities, we can convolve the trivially nonnegative forms $f_S^{(k)}$ with Euler relations to get relations for posets of higher ranks of the form

$$f_S^{(k)} * (f_\emptyset^{(d)} - f_{\{1\}}^{(d)} + f_{\{2\}}^{(d)} - \dots + (-1)^{d-1} f_{\{d-1\}}^{(d)} + (-1)^d f_\emptyset^{(d)}) * f_T^{(l)} = 0. \tag{6}$$

These are enough to generate all linear relations on flag f -vectors on polytopes.

Theorem 3.2 ([5]). *All linear relations on the $f_S^{(d)}$ for polytopes, and so for Eulerian posets, are derived from those coming from the Euler relations as in (6).*

The equations in [5] are identical to those in equation (6), although they originally were expressed without the use of the convolution. The proof there that these are all the equations consists of producing Fibonacci many polytopes whose flag f -vectors span. These can be made for each dimension by considering repeated operations of forming *pyramids* P and *prisms* B starting with an edge, never taking two B 's in a row. The number of words of length $d - 1$ in P and B , with no repeated B , is a Fibonacci number. A simpler algebraic proof that flag f -vectors of polytopes span that does not give a specific basis is given in [14], where it is shown also that zonotopes will suffice. See also [36] for another basis.

There is a simple algebraic way of capturing the notion of convolution product and relations on flag numbers in Eulerian posets. Let

$$A = \mathbb{Q}\langle y_1, y_2, \dots \rangle = A_0 \oplus A_1 \oplus A_2 \dots$$

be the free associative \mathbb{Q} -algebra on noncommuting y_i , graded by $\deg(y_i) = i$. Here

$$A_n = \text{span}_{\mathbb{Q}}\{ y_{i_1} y_{i_2} \dots y_{i_k} \mid i_1 + i_2 + \dots + i_k = n \}.$$

We say $\beta = (\beta_1, \dots, \beta_k)$ is a *composition* of integer $n > 0$ (written $\beta \models n$) if each $\beta_i > 0$ and $|\beta| := \beta_1 + \dots + \beta_k = n$. There is a simple bijection between compositions of $n + 1$ and subsets of $[n] := \{1, \dots, n\}$ given by

$$\beta = (\beta_1, \dots, \beta_k) \models n + 1 \mapsto S(\beta) := \{ \beta_1, \beta_1 + \beta_2, \dots, \beta_1 + \dots + \beta_{k-1} \} \subseteq [n]$$

and

$$S = \{ i_1, \dots, i_{k-1} \} \subseteq [n] \mapsto \beta(S) := (i_1, i_2 - i_1, i_3 - i_2, \dots, n + 1 - i_{k-1}) \models n + 1.$$

We will freely move between indexing by compositions and the associated subsets in the rest of this paper.

Via the association of y_k and $f_{\emptyset}^{(k)}$ and so of

$$y_{\beta} := y_{\beta_1} \dots y_{\beta_k}, \quad \beta = (\beta_1, \dots, \beta_k) \models n + 1 \quad \text{and} \quad f_{S(\beta)}^{(n+1)} = f_S^{(n+1)}, \quad S \subseteq [n],$$

multiplication in A can be seen to be the analog of Kalai's convolution of flag f -vectors, in which

$$f_S^{(n)} * f_T^{(m)} = f_{S \cup \{n\} \cup (T+n)}^{(n+m)}.$$

Example 4. With this association $f_{\{1\}}^{(3)} = y_1 y_2$ and so

$$f_{\{1\}}^{(3)} * f_{\{1\}}^{(3)} = y_1 y_2 y_1 y_2 = f_{\{1,3,4\}}^{(6)}.$$

In general, we get an association between elements $G \in A_n$ and expressions of the form $\sum_{S \subseteq [n-1]} \alpha_S f_S^{(n)}$. Multiplying a form G in this algebra by y_i on the

right (left) corresponds to summing G evaluated on all faces (or links of faces) of corank (rank) i .

For $k \geq 1$ define in A_k the form

$$\chi_k := \sum_{i+j=k} (-1)^i y_i y_j = \sum_{i=0}^k (-1)^i f_i^{(k)},$$

the k^{th} Euler relation, where we take $y_0 = 1$ and $f_0^{(k)} = f_k^{(k)} = f_\emptyset^{(k)}$. Thus in A_4 ,

$$\chi_4 = y_0 y_4 - y_1 y_3 + y_2 y_2 - y_3 y_1 + y_4 y_0 = 2f_\emptyset^{(4)} - f_{\{1\}}^{(4)} + f_{\{2\}}^{(4)} - f_{\{3\}}^{(4)},$$

the Euler relation for posets of rank 4. By Theorem 3.2, the 2-sided ideal

$$I_{\mathcal{E}} = \langle \chi_k : k \geq 1 \rangle \subset A$$

is the space of all relations on Eulerian posets. We define

$$A_{\mathcal{E}} = A/I_{\mathcal{E}},$$

and think of $A_{\mathcal{E}}$ as the algebra of functionals on Eulerian posets. It turns out that it too is a free associative algebra, the algebra of *odd jumps*.

Theorem 3.3 ([19]). *There is an isomorphism of graded algebras,*

$$A_{\mathcal{E}} \cong \mathbb{Q}\langle y_1, y_3, y_5, \dots \rangle,$$

and so $\dim_{\mathbb{Q}}(A_{\mathcal{E}})_n$ is the n^{th} Fibonacci number.

3.3. Quasisymmetric function of a graded poset. Note that the algebras A and $A_{\mathcal{E}}$ discussed in the last section were noncommutative. We can also associate a pair of commutative algebras to the flag vectors of graded and Eulerian posets.

Let $QSym \subset \mathbb{Q}[[x_1, x_2, \dots]]$ be the algebra of all *quasisymmetric functions*

$$QSym := QSym_0 \oplus QSym_1 \oplus \dots$$

where

$$QSym_n := \text{span}_{\mathbb{Q}}\{M_\beta \mid \beta = (\beta_1, \dots, \beta_k) \models n\}$$

and

$$M_\beta := \sum_{i_1 < i_2 < \dots < i_k} x_{i_1}^{\beta_1} x_{i_2}^{\beta_2} \dots x_{i_k}^{\beta_k}.$$

Here $M_0 = 1$ so $QSym_0 = \mathbb{Q}$; otherwise $k > 0$, each $\beta_i > 0$, and $\beta_1 + \dots + \beta_k = n$. Alternatively, $QSym$ consists of all bounded degree power series in $\mathbb{Q}[[x_1, x_2, \dots]]$ such that for all $\beta \models n$, the coefficient of $x_{i_1}^{\beta_1} x_{i_2}^{\beta_2} \dots x_{i_k}^{\beta_k}$ is the same as that of $x_1^{\beta_1} x_2^{\beta_2} \dots x_k^{\beta_k}$ whenever $i_1 < \dots < i_k$.

For example, $(1, 2, 1) \models 4$ and $M_{(1,2,1)} = \sum_{i_1 < i_2 < i_3} x_{i_1}^1 x_{i_2}^2 x_{i_3}^1$. We can index also by subsets. For $S \subseteq [n]$, define

$$M_S = M_S^{(n+1)} := M_{\beta(S)},$$

so, for example, if $S = \{1, 3\} \subseteq [3]$ then $\beta(S) = (1, 2, 1) \models 4$ and

$$M_{\{1,3\}} = M_{\{1,3\}}^{(4)} = M_{(1,2,1)}.$$

This basis $\{M_\beta \mid \beta \models n, n \geq 0\}$ is known as the *monomial* basis for $QSym$.

We note that quasisymmetric functions arise naturally as weight enumerators of P -partitions of labeled posets [34]. In this context, a more natural basis arises as weight enumerators of labeled chains,

$$L_S = \sum_{T \supseteq S} M_T.$$

Here $S \subseteq T \subseteq [n]$ and S is the *descent set* of the labeling. This is known as the *fundamental* basis for $QSym$. See [52, §7.19] for further discussion.

We summarize here the basics of the use of quasisymmetric functions in the theory of flag f -vectors of graded posets and, in particular, Eulerian posets. For a finite graded poset P , with rank function $\rho(\cdot)$, we define the formal power series

$$F(P) := \sum_{\hat{0}=u_0 \leq \dots \leq u_{k-1} < u_k = \hat{1}} x_1^{\rho(u_0, u_1)} x_2^{\rho(u_1, u_2)} \dots x_k^{\rho(u_{k-1}, u_k)}, \tag{7}$$

where the sum is over all finite *multichains* in P whose last two elements are distinct and $\rho(x, y) = \rho(y) - \rho(x)$. See [30] for general properties of $F(P)$. In particular, we have the following.

Proposition 3.4. *For a graded poset P ,*

- 3.4.1.** $F(P) \in QSym$, in fact $F(P) \in QSym_n$ where $n = \rho(P)$,
- 3.4.2.** $F(P_1 \times P_2) = F(P_1) F(P_2)$,
- 3.4.3.** $F(P) = \sum_\alpha f_\alpha M_\alpha = \sum_\alpha h_\alpha L_\alpha$, where f_α and h_α are the flag f and flag h -vectors, respectively, of P .

We define next an interesting subalgebra of $QSym$ that turns out to be related to Eulerian posets. For a **cd**-word w of degree n ,

$$w = \mathbf{c}^{n_1} \mathbf{d} \mathbf{c}^{n_2} \mathbf{d} \dots \mathbf{c}^{n_k} \mathbf{d} \mathbf{c}^m,$$

where $\deg \mathbf{c} = 1$ and $\deg \mathbf{d} = 2$, let

$$\mathcal{I}_w = \{\{i_1 - 1, i_1\}, \{i_2 - 1, i_2\}, \dots, \{i_k - 1, i_k\}\},$$

where $i_j = \deg(\mathbf{c}^{n_1} \mathbf{d} \mathbf{c}^{n_2} \mathbf{d} \cdots \mathbf{c}^{n_j} \mathbf{d})$. Note that \mathcal{I}_w consists of disjoint intervals in $[n]$, all of size 2. These and more general *even* antichains of intervals have been related to extremes of the cone of Eulerian flag vectors in [8, 9].

The *peak algebra* Π is defined to be the subalgebra of $QSym$ generated by the *peak quasisymmetric functions*

$$\Theta_w = \sum_{T \in b[\mathcal{I}_w]} 2^{|T|+1} M_T^{(n+1)}, \tag{8}$$

where w is any **cd**-word (including empty **cd**-word $\mathbf{1}$, for which $\mathcal{I}_1 = \emptyset$). Note that if $\deg w = n$, then $\deg \Theta_w = n + 1$; there are Fibonacci many Θ_w of each degree.

The peak algebra was first defined by Stembridge [55], where peak quasisymmetric functions arise naturally as weight enumerators of *enriched P-partitions* of labeled posets.

A subset $S \subseteq [n]$ is *sparse* if $1 \notin S$ and $i \in S \Rightarrow i - 1 \notin S$. We can associate a sparse subset $S_w \subseteq [n]$ to a **cd**-word of degree n by associating

$$w = \mathbf{c}^{n_1} \mathbf{d} \mathbf{c}^{n_2} \mathbf{d} \cdots \mathbf{c}^{n_k} \mathbf{d} \mathbf{c}^m \quad \text{and} \quad S_w = \{i_1, i_2, \dots, i_k\} \subset [n],$$

where $i_j = \deg(\mathbf{c}^{n_1} \mathbf{d} \mathbf{c}^{n_2} \mathbf{d} \cdots \mathbf{c}^{n_j} \mathbf{d})$. Stembridge considers the basis for Π to be labeled by sets S of the form S_w . In this context, his basis Θ_S arises as weight enumerators of labeled chains, where S is the *peak set* of the labeling. (A *peak* is a descent preceded by an ascent.)

3.4. Peak functions and Eulerian posets. The main result for our purposes with respect to the subalgebra Π is due to Bergeron, Mykytiuk, Sottile and van Willigenburg [10].

Theorem 3.5. *If P is an Eulerian poset, then $F(P) \in \Pi$.*

The proof of Theorem 3.5 depends on connections between the enumeration algebra $\mathbb{Q}\langle y_1, y_2, \dots \rangle$ and the algebra of quasisymmetric functions $QSym$ as well as between the quotient $A_{\mathcal{E}}$ and the subalgebra Π of peak functions. Now the algebras Π and $A_{\mathcal{E}}$ both have Hilbert series given by the Fibonacci sequence, although they are surely not isomorphic: Π is commutative, $A_{\mathcal{E}}$ is not. The connection comes via duality of Hopf algebras. We summarize this briefly here.

Let B be a graded algebra. The product on the algebra B can be viewed as a homogeneous linear map

$$B \otimes B \longrightarrow B, \quad a \otimes b \mapsto a \cdot b$$

A coalgebra C has instead a *coproduct* $C \longrightarrow C \otimes C$, as well as a *counit*, an analog of the unit in an algebra. A *Hopf algebra* H has both product and coproduct (plus unit and counit), as well as a map $S : H \rightarrow H$ known as the *antipode*. (In the case of graded Hopf algebras, the antipode is uniquely specified

by the product and coproduct; see, e.g. [30, Lemma 2.1] or the Appendix in [10].) In the dual vector space H^* to a Hopf algebra H , the adjoint of the product on H

$$H^* \otimes H^* \longleftarrow H^*$$

gives a coproduct on H^* , and the adjoint of the coproduct on H

$$H^* \longleftarrow H^* \otimes H^*$$

gives a product on H^* , making H^* a Hopf algebra as well. H^* is the *dual Hopf algebra* to H .¹

The four algebras we have discussed are all graded Hopf algebras, with the coproducts defined below. In [33], the *integral* Hopf algebra $NC = \mathbb{Z}\langle y_1, y_2, \dots \rangle$ (called there the *noncommutative symmetric functions*) was shown to be dual to the Hopf algebra of quasisymmetric functions with integral coefficients, with coproducts

$$\Delta(M_\beta) = \sum_{\beta = \beta_1 \cdot \beta_2} M_{\beta_1} \otimes M_{\beta_2}$$

for $QSym$ and

$$\Delta(y_k) = \sum_{i+j=k} y_i \otimes y_j.$$

for NC . So, for example,

$$\Delta(M_{(2,1,1)}) = 1 \otimes M_{(2,1,1)} + M_{(2)} \otimes M_{(1,1)} + M_{(2,1)} \otimes M_{(1)} + M_{(2,1,1)} \otimes 1$$

and $\Delta(y_2) = 1 \otimes y_2 + y_1 \otimes y_1 + y_2 \otimes 1$, where, as before, we take $y_0 = 1$.

In [10], these coproducts on $QSym$ and A , respectively, are shown to extend to coproducts on Π and $A_{\mathcal{E}}$, and they proved [10, Theorem 5.4]:

Theorem 3.6 ([10]). *These coproducts make Π and $A_{\mathcal{E}}$ into a dual pair of Hopf algebras.*

Theorem 3.5 follows directly from this: For any graded poset P , the quasisymmetric function $F(P) = \sum_S f_S(P)M_S$ defines a functional $A \rightarrow \mathbb{Q}$, defined by $\sum_S \alpha_S f_S \mapsto \sum_S \alpha_S f_S(P)$, in $A^* = QSym$. Theorem 3.6 implies that Π is the kernel of the restriction of this functional to functionals on the ideal $I_{\mathcal{E}}$. By the definition of $I_{\mathcal{E}}$, any Eulerian P has an $F(P)$ in this kernel, so $F(P) \in \Pi$.

This leads immediately to the following question: For an Eulerian poset P , what is the representation of $F(P)$ in terms of the basis of peak functions $\{\Theta_w\}$ for Π ? Equivalently, what is the dual basis in $A_{\mathcal{E}}$ to the basis $\{\Theta_w\}$? This was answered in [16].

¹In reality, we are considering the *graded dual* $H^* = \oplus H_i^*$ of the graded Hopf algebra $H = \oplus H_i$ [2]. All products and coproducts we describe will be homogeneous maps.

Theorem 3.7 ([16]). *If P is any Eulerian poset, then*

$$F(P) = \sum_w \frac{1}{2^{|w|_{\mathbf{d}}+1}} [w]_P \Theta_w,$$

where the $[w]_P$ are the coefficients of the **cd**-index of P and $|w|_{\mathbf{d}}$ is the number of \mathbf{d} 's in w .

Corollary. *The elements*

$$\frac{1}{2^{|w|_{\mathbf{d}}+1}} [w] \in A_{\mathcal{E}}$$

form a dual basis to the basis Θ_w in Π .

Since, in terms of the theory of P -partitions, the subalgebra Π and the basis $\{\Theta_w\}$ arise naturally when considering the algebra $QSym$, one sees that the **cd**-index is a natural, in fact, inescapable, invariant in the context of flag enumeration in Eulerian posets. We see in the next section how these ideas lead to an interesting new invariant in the theory of Bruhat intervals on Coxeter groups.

4. Bruhat Intervals in Coxeter Groups

A Coxeter group is a group W generated by a finite set S with the relations $s^2 = e$ for all $s \in S$ (e is the identity of W) and otherwise only relations of the form

$$(ss')^{m(s,s')} = e,$$

for $s \neq s' \in S$ with $m(s,s') = m(s',s) \geq 2$. There are many examples of such groups, including the symmetry groups of regular polytopes (and so the symmetric groups) and the finite reflection groups. See [35] and [20] for general background, especially the latter for the combinatorial theory of Coxeter groups discussed here.

Given a Coxeter system (W, S) (the set of generators is a critical component), each $v \in W$ can be written $v = s_1 s_2 \cdots s_k$ with $s_i \in S$. If k is minimal among all such expressions for v , then $s_1 s_2 \cdots s_k$ is called a *reduced expression* for v and $k = l(v)$ is called the *length* of v .

The Bruhat order on (W, S) is a partial order on the set W , defined as follows. If $v = s_1 s_2 \cdots s_k$ is a reduced expression for v , then $u \leq v$ for $u \in W$ if some (reduced) expression for u is a subword $u = s_{i_1} s_{i_2} \cdots s_{i_\ell}$, $i_1 < i_2 < \cdots < i_\ell$, of v .

It was shown by Verma [57] that for each $u \leq v \in W$ the *Bruhat interval* $[u, v]$ is an Eulerian poset of rank $l(u, v) := l(v) - l(u)$. Thus, as an Eulerian poset, the interval $[u, v]$ has a **cd**-index. This was first studied in any detail by Reading [44], who showed that there were no equations other than those described in Theorem 3.2 that held for the flag vectors of all Bruhat intervals.

Here we extend the **cd**-index of a Bruhat interval to the *complete cd*-index, a nonhomogeneous **cd**-polynomial of degree $l(u, v) - 1$ that includes enough information to compute important invariants for the interval, including its R -polynomial and its Kazhdan-Lusztig polynomial. The remainder of this section represents mostly joint work with Francesco Brenti [12].

4.1. R -polynomial and Kazhdan-Lusztig polynomial. Let $\mathcal{H}(W)$ be the *Hecke algebra* associated to W , i.e. the free $\mathbb{Z}[q, q^{-1}]$ -module having the set $\{T_v \mid v \in W\}$ as a basis and multiplication such that for all $v \in W$ and $s \in S$

$$T_v T_s = \begin{cases} T_{vs}, & \text{if } l(vs) > l(v) \\ qT_{vs} + (q - 1)T_v, & \text{if } l(vs) < l(v). \end{cases}$$

Note that were we to set $q = 1$, then this would give precisely the integral group ring of W . $\mathcal{H}(W)$ is an associative algebra having T_e as unity, in which each T_v is invertible. For $v \in W$,

$$(T_{v^{-1}})^{-1} = q^{-l(v)} \sum_{u \leq v} (-1)^{l(u,v)} R_{u,v}(q) T_u,$$

where $R_{u,v}(q) \in \mathbb{Z}[q]$.

The polynomials $R_{u,v}$ are called the *R-polynomials* of W . For $u, v \in W$, $u \leq v$, $\deg(R_{u,v}) = l(u, v)$ and $R_{u,u}(q) = 1$. It is customary to set $R_{u,v}(q) \equiv 0$ if $u \not\leq v$.

The *Kazhdan-Lusztig polynomial* $P_{u,v}$ of a Bruhat interval $[u, v]$ is defined by the following theorem. A proof can be found in [35, §9-11] of an equivalent statement. The version here is [20, Theorem 5.1.4].

Theorem 4.1. *There is a unique family of polynomials $\{P_{u,v}(q)\}_{u,v \in W} \subset \mathbb{Z}[q]$, such that, for all $u, v \in W$,*

4.1.1. $P_{u,v}(q) = 0$ if $u \not\leq v$;

4.1.2. $P_{u,u}(q) = 1$;

4.1.3. $\deg(P_{u,v}(q)) \leq \lfloor \frac{l(u,v)-1}{2} \rfloor$, if $u < v$, and

4.1.4.

$$q^{l(u,v)} P_{u,v} \left(\frac{1}{q} \right) = \sum_{u \leq z \leq v} R_{u,z}(q) P_{z,v}(q),$$

if $u \leq v$.

The main conjectures in this area are that for all Coxeter systems (W, S) and all Bruhat intervals $[u, v]$ in W , the Kazhdan-Lusztig polynomial $P_{u,v}$ is nonnegative, and depends only on the *poset* $[u, v]$, and not on the underlying

group. The first conjecture is known to hold, for example, for all finite Coxeter groups and the second for all *lower* intervals, that is, intervals where $u = e$, the identity element of W [24]. Both conjectures are known to hold when the interval $[u, v]$ is a lattice². See the discussion pp. 161–162 and 171–172 of [20] for references.

4.2. The complete quasisymmetric function of a Bruhat interval and the complete cd-index. While the R -polynomial of a Bruhat interval may have negative terms, there is an associated polynomial that has nonnegative coefficients with a direct combinatorial interpretation. The following is [20, Proposition 5.3.1].

Proposition 4.2. *For $u \leq v \in W$, there exists a (necessarily unique) polynomial $\tilde{R}_{u,v}(q) \in \mathbb{N}[q]$ such that*

$$R_{u,v}(q) = q^{\frac{l(u,v)}{2}} \tilde{R}_{u,v} \left(q^{\frac{1}{2}} - q^{-\frac{1}{2}} \right).$$

For a Bruhat interval $[u, v]$, we use the \tilde{R} -polynomials to define a nonhomogeneous analog of the quasisymmetric function $F(P)$ of a graded poset. For Bruhat interval $[u, v]$, the *complete quasisymmetric function* is defined by

$$\tilde{F}(u, v) := \sum_{u=u_0 \leq \dots \leq u_{k-1} < u_k=v} \tilde{R}_{u_0, u_1}(x_1) \tilde{R}_{u_1, u_2}(x_2) \cdots \tilde{R}_{u_{k-1}, u_k}(x_k). \tag{9}$$

Again, the sum is over all finite multichains in $[u, v]$ whose last two elements are distinct. It is straightforward to show that \tilde{F} is multiplicative [12, Proposition 2.6], that is, for Bruhat intervals $[u_i, v_i]$, $\tilde{F}([u_1, v_1] \times [u_2, v_2]) = \tilde{F}(u_1, v_1) \tilde{F}(u_2, v_2)$.³

To give an analog of Proposition 3.4 for $\tilde{F}(u, v)$, we need to define the Bruhat graph of the interval $[u, v]$. Let $T = \{wsw^{-1} \mid w \in W, s \in S\}$ be the set of all conjugates of the generators in W . Elements of T are called *reflections*, while elements of S are called *simple reflections*.

We define the *Bruhat graph* of a Coxeter system (W, S) to be the directed graph $B(W, S)$ obtained by taking W as vertex set and putting a directed edge from x to y if and only if $x^{-1}y \in T$ and $l(x) < l(y)$. We can consider the edge (x, y) of $B(W, S)$ to be labeled by the reflection $t = x^{-1}y$.

The Bruhat graph of an interval $[u, v]$ is the subgraph of $B(W, S)$ induced by the elements in $[u, v]$; it contains the Hasse diagram of the poset (directed

²This follows since in this case $P_{u,v}(q) = g([u, v]^*, q)$, which depends only on the poset $[u, v]$ (see Remark 1 in §4.2 and Remark 2 in §4.3). By an unpublished result of Dyer, lattice Bruhat intervals are face posets of polytopes, so nonnegativity follows from the generalized lower bound theorem for polytopes.

³In fact both F and \tilde{F} are maps of Hopf algebras (see [30, Proposition 4.4] and [12, Remark 2.8]). This will also be a consequence of the results discussed in §5.

in increasing Bruhat order) as a spanning subgraph. The Bruhat graph was first defined by Dyer [28], who showed the graph (not including the labeling) to depend only on the isomorphism class of the poset $[u, v]$ and not on the underlying group.

A reflection subgroup of W is any subgroup W' of W generated by a subset of T . For $w \in W$, define $N(w) := \{t \in T : l(tw) < l(w)\}$. Reflection subgroups W' are Coxeter groups, with simple reflections $S' = \{t' \in T : N(t') \cap W' = \{t'\}\}$ [26, 27]. See also [35, §8.2]. A reflection subgroup (W', S') is said to be *dihedral* if $|S'| = 2$.

A total ordering $<_T$ on the set of all reflections T in (W, S) is called a *reflection ordering* if it satisfies the following: For any dihedral reflection subgroup (W', S') , where $S' = \{a, b\}$, either $a <_T aba <_T ababa <_T \dots <_T babab <_T bab <_T b$ or $b <_T bab <_T babab <_T \dots <_T ababa <_T aba <_T a$. The existence of reflection orderings for any Coxeter system was shown by Dyer in [29].

Example 5. The symmetric group $W = \mathcal{S}_n$ is a Coxeter group (often denoted A_{n-1}) with Coxeter generators given by the *adjacent* transpositions $s_i = (i \ i+1)$, $i = 1, \dots, n-1$. Here, reflections are *all* transpositions $(i \ j)$, and lexicographic order is a reflection order. Thus in \mathcal{S}_4 , $(12) <_T (13) <_T (14) <_T (23) <_T (24) <_T (34)$.

Given a reflection ordering on the interval $[u, v]$, directed u - v paths in its Bruhat graph are labeled by reflections, and so they have a well-defined descent set in this ordering. For $\alpha \models k$, $k \leq n + 1 = l(u, v)$, we denote by $b_\alpha = b_\alpha(u, v)$ the number of paths of length k having descent set $S = S(\alpha)$. Further, define

$$c_\alpha(u, v) = \sum_{\{\beta \models n \mid \alpha \preceq \beta\}} b_\beta(u, v)$$

where \preceq denotes refinement of compositions (parts of β are sums of successive parts of α). Using the quantities b_α and c_α , we can express the complete quasisymmetric function $\tilde{F}(u, v)$ in terms of the fundamental and monomial bases for $QSym$.

Proposition 4.3 ([12]). $\tilde{F}(u, v) = \sum_\alpha c_\alpha(u, v) M_\alpha = \sum_\alpha b_\alpha(u, v) L_\alpha$

Thus we see that $c_\alpha(u, v)$ and $b_\alpha(u, v)$ are analogs of the flag f - and flag h -numbers. Note that it is possible that the quantities $c_\alpha(u, v)$ can be *greater than 1* for $\alpha \models k$, $k < l(u, v)$, that is, there can be more than one rising Bruhat path of less than maximum length.

Since the Bruhat order on $[u, v]$ is always Eulerian, we know $F([u, v]) \in \Pi$, but usually $\tilde{F}(u, v) \neq F([u, v])$. In [23, Theorem 8.4], Brenti showed that the coefficients $c_\alpha(u, v)$ satisfy the equations (6), and so by [16, Proposition 1.3], we can conclude

Theorem 4.4. For any Bruhat interval $[u, v]$, $\tilde{F}(u, v) \in \Pi$, in fact

$$\tilde{F}(u, v) \in \Pi_{l(u,v)} \oplus \Pi_{l(u,v)-2} \oplus \Pi_{l(u,v)-4} \oplus \cdots .$$

The last assertion follows since the $b_\alpha(u, v)$ count directed paths from u to v of length $|\alpha|$ in the Bruhat graph $B(W, S)$, and all of these must have length $k \equiv l(u, v) \pmod{2}$. This is true since for any reflection t , $l(ut) - l(u)$ is odd, and so the length of every Bruhat path has the same parity.

Since $\tilde{F}(u, v) \in \Pi$, we can express it in terms of the peak basis Θ_w . We define the *complete cd-index* of the Bruhat interval $[u, v]$

$$\tilde{\Phi}_{u,v} := \sum_w [w]_{u,v} w$$

by the unique expression

$$\tilde{F}(u, v) = \sum_w [w]_{u,v} \left[\frac{1}{2^{|w|_d+1}} \Theta_w \right],$$

where the sum is over all **cd**-words w with $\deg(w) = l(u, v) - 1, l(u, v) - 3, \dots$

In [29], Dyer shows that the polynomial $\tilde{R}_{u,v}(q)$ enumerates *rising* paths in the Bruhat graph of $[u, v]$, *i.e.*, the coefficient of q^k is the number of paths of length k with empty descent set (see [29, Corollary 3.4] or [20, Theorem 5.3.4]). In [29, §4], he also shows that the reflection labeling of the Bruhat graph gives an *EL*-labeling on the maximal length Bruhat paths in $[u, v]$. Together, they imply that the leading term of $\tilde{R}_{u,v}(q)$ is 1, since, in particular, an *EL*-labeling will always have a unique rising path.

Remark 1. One consequence of this is that $c_\alpha(u, v) = f_\alpha([u, v])$ when $\alpha \models l(u, v)$ and so $\tilde{F}(u, v) = F([u, v]) +$ lower terms. Thus the top-degree terms of $\tilde{\Phi}_{u,v}$ (*i.e.*, those of degree $l(u, v) - 1$) constitute the ordinary **cd**-index of the underlying poset $[u, v]$, *i.e.*, $\tilde{\Phi}_{u,v} = \Phi_{[u,v]} +$ lower terms. If $[u, v]$ is a lattice, then $\tilde{\Phi}_{u,v} = \Phi_{[u,v]}$.⁴

By Dyer’s *EL*-labeling (or by the earlier *CL*-labeling of Björner and Wachs; see [20, Corollary 2.7.6]), the poset $[u, v]$ is Gorenstein*, so by the result of Karu, $\Phi_{[u,v]} \geq 0$. The following is Conjecture 6.1 in [12].

Conjecture 2 ([12]). For all Bruhat intervals $[u, v]$, $\tilde{\Phi}_{u,v} \geq 0$.

We can easily see that all the pure **c** coefficients $[c^{k-1}]_{u,v} = b_{(k)}(u, v) \geq 0$, where (k) is the composition with one part. Since $b_{(k)}$ counts the rising paths of length k , we get the following [12, Corollary 2.10].

⁴It is a consequence of an unpublished result of Dyer that $\tilde{\Phi}_{u,v} = \Phi_{[u,v]}$ if and only if $[u, v]$ is a lattice.

Proposition 4.5. For $u < v$, $\tilde{R}_{u,v}(q) = q \tilde{\Phi}_{u,v}(q, 0)$.

There is some evidence for Conjecture 2; see [12, §6], where one consequence is proved. Further, if d_{min} is the least degree of a term in $\tilde{\Phi}_{u,v}$, it is known that if $d_{min} \leq 2$ or if $[\mathbf{c}^{d_{min}}] = 1$, then $[w]_{u,v} \geq 0$ if $\deg w = d_{min}$.⁵

In [44], Reading also showed that for lower intervals $[e, v]$, $\Phi_{[e,v]}$ is termwise less than or equal to the **cd** index of the Boolean algebra $\mathcal{B}_{l(v)}$ of rank $l(v)$. We conjecture that this also bounds the complete **cd**-index of lower intervals, in the following sense.

Conjecture 3. For all lower Bruhat intervals $[e, v]$, $\tilde{\Phi}_{e,v}(1, 1) \leq \Phi_{\mathcal{B}_{l(v)}}(1, 1)$.

4.3. Kazhdan-Lusztig polynomial and the complete **cd-index.** We note here that if we were only interested in the complete **cd**-index of the interval $[u, v]$, it could have been defined directly by means of a nonhomogeneous **ab** polynomial $\tilde{\Psi}_{u,v}$ defined analogously to (2), using the quantities b_α in place of h_S (see [12, Proposition 2.9]). However, the form of the quasisymmetric function $\tilde{\Phi}_{u,v}$ given in Proposition 4.3 leads directly to a way of expressing the Kazhdan-Lusztig polynomial $P_{u,v}$ in terms of the coefficients of the complete **cd**-index.

We first consider a family of polynomials $B_k(q)$. We call these *ballot polynomials*, since the coefficient of q^i in $B_k(q)$ is the number of ways k ballots can be cast so that the losing candidate receives i votes, while the winning candidate is never behind. Define

$$B_k(q) := \sum_{i=0}^{\lfloor k/2 \rfloor} \frac{k+1-2i}{k+1} \binom{k+1}{i} q^i. \tag{10}$$

The constant term of $B_k(q)$ is always 1 and, when k is even, the lead term is a Catalan number.

For $n \geq 0$ define the *dihedral poset* D_n of rank $n + 1$ to be a graded poset with two elements at each rank $1 \leq i \leq n$ where $x \leq y$ if $\rho(x) \leq \rho(y)$. Since each interval in a dihedral poset is dihedral, it is easy to see that D_n is Eulerian for each $n \geq 0$, and it is an easy calculation to see that $\Phi_{D_n} = \mathbf{c}^n$. D_n is the underlying Bruhat poset of a dihedral group of order $2n + 2$, and it follows from discussion following [20, Proposition 5.1.8] that $P_{D_n} = 1$. It is straightforward to verify that $\tilde{\Phi}_{D_n} = \mathbf{c} \cdot \tilde{\Phi}_{D_{n-1}} + \tilde{\Phi}_{D_{n-2}}$, with $\tilde{\Phi}_{D_0} = 1$ and $\tilde{\Phi}_{D_1} = \mathbf{c}$, and so $\tilde{\Phi}_{D_n} = \sum_{j=0}^{\lfloor n/2 \rfloor} \binom{n-j}{n-2j} \mathbf{c}^{n-2j}$. In fact $\tilde{\Phi}_{u,v}(\mathbf{c}, \mathbf{d}) = \tilde{\Phi}_{u,v}(\mathbf{c}, 0)$ if and only if $[u, v]$ is dihedral.⁵ As for the g -polynomial of D_n , the following is [48, Proposition 2.5].

Proposition 4.6. The g -polynomial of the dihedral poset D_n is the alternating ballot polynomial $B_n(-q)$.

⁵These are results that will appear in the forthcoming Cornell Ph.D. Thesis of S.A. Blanco.

In [20, Theorem 5.5.7], an expression is given for $P_{u,v}$, $u < v$, in terms of the $b_\alpha(u, v)$ and universal polynomials Υ_α that enumerate an implicitly defined set of lattice paths. By expressing this in terms of the complete **cd**-index of $[u, v]$, the resulting paths are now explicit, and we can get an expression for $P_{u,v}$ in terms of only the coefficients $[w]_{u,v}$ of the complete **cd** index $\tilde{\Phi}_{u,v}$ and shifts of the alternating ballot polynomials $B_k(-q)$.

A **cd**-word w is said to be *even* if it is a word in \mathbf{c}^2 and \mathbf{d} . For an even **cd**-word $w = \mathbf{c}^{n_1} \mathbf{d} \mathbf{c}^{n_2} \mathbf{d} \cdots \mathbf{d} \mathbf{c}^{n_k}$, let $C_w = C_{n_1/2} \cdots C_{n_k/2}$, where $C_i = \frac{1}{2i+1} \binom{2i+1}{i}$, the i^{th} Catalan number. Finally, let $|w| := \deg w$ and $|w|_{\mathbf{d}}$ be the number of \mathbf{d} 's in w . The following is [12, Theorem 4.1].

Theorem 4.7. *For any Bruhat interval $[u, v]$ of rank $l(u, v) = n + 1$,*

$$P_{u,v}(q) = \sum_{i=0}^{\lfloor n/2 \rfloor} a_i q^i B_{n-2i}(-q)$$

where

$$a_i = a_i(u, v) = [\mathbf{c}^{n-2i}]_{u,v} + \sum_{\mathbf{d}w \text{ even}} (-1)^{\frac{|w|}{2} + |w|_{\mathbf{d}}} C_{\mathbf{d}w} [\mathbf{c}^{n-2i} \mathbf{d}w]_{u,v}. \tag{11}$$

Note that the coefficient $a_i(u, v)$ of $q^i B_{n-2i}(-q)$ in this expression for $P_{u,v}$ depends only on **cd**-words beginning with \mathbf{c}^{n-2i} that are otherwise even. The expression for $P_{u,v}(q) = p_0 + p_1q + \cdots$ in terms of the $a_i(u, v)$ can be inverted to give

$$a_j = \sum_{i=0}^j \binom{n-j-i}{n-2j} p_i, \tag{12}$$

for $j = 0, \dots, \lfloor n/2 \rfloor$. Thus if $P_{u,v}(q) \geq 0$ then $a_i(u, v) \geq 0$ for $i = 0, \dots, \lfloor n/2 \rfloor$. The conjectured nonnegativity of $P_{u,v}$ leads to the following, which is [25, Conjecture 6.6] as well as [12, Conjecture 4.11].

Conjecture 4. For each Bruhat interval $[u, v]$ of rank $l(u, v) = n + 1$, $a_i(u, v) \geq 0$ for $i = 0, 1, \dots, \lfloor n/2 \rfloor$.

Remark 2. We note that if we restrict the $[w]_{u,v}$ in (11) to those of degree n only, then we get the formula of Bayer and Ehrenborg [7, Theorem 4.2] for the g -polynomial of the dual poset $[u, v]^*$. Thus the difference $P_{u,v}(q) - g([u, v]^*, q)$ is a function of the lower-degree **cd**-coefficients only (and their only function in this expression). Example 4.6 of [12] gives a pair of rank 6 Bruhat intervals in $W = \mathcal{S}_5$ having the same **cd**-index but unequal Kazhdan-Lusztig polynomials, and thus unequal complete **cd**-indices.

Finally, we point out that as far as combinatorial invariance is concerned, $P_{u,v}$, $a_i(u, v)$ and $[w]_{u,v}$ are all equivalent. We say that an invariant of Bruhat

intervals is *combinatorially invariant* if its value on a Bruhat interval $[u, v]$ depends only on the isomorphism type of the poset $[u, v]$.

Proposition 4.8. *The following are equivalent for all Coxeter systems (W, S) .*

4.8.1 *For all $u \leq v \in W$, $P_{u,v}$ is combinatorially invariant.*

4.8.2 *For all $u \leq v \in W$ and $i = 0, \dots, \lfloor \frac{l(u,v)-1}{2} \rfloor$, $a_i(u, v)$ is combinatorially invariant.*

4.8.3 *For all $u \leq v \in W$, and all **cd**-words of degree $n, n - 2, \dots$, where $n = l(u, v) - 1$, $[w]_{u,v}$ is combinatorially invariant.*

The equivalence of 4.8.1 and 4.8.3 is discussed in [12, Remark 4.13].

5. Epilog: Combinatorial Hopf Algebras

There is a general enumeration theory that explains the existence of the quasisymmetric functions such as $F(P)$ and $\tilde{F}(u, v)$ as well as many other quasisymmetric generating functions that arise in combinatorial theory. Originally formulated by Aguiar in [1] in the context of infinitesimal Hopf algebras, it was later expanded by Aguiar, Bergeron and Sottile and reformulated for Hopf algebras [2]. We summarize this theory and a more recent extension below.

Let $H = H_0 \oplus H_1 \oplus H_2 \oplus \dots$ be a graded connected Hopf algebra (say, over \mathbb{Q}). This means $H_0 \cong \mathbb{Q}$ and the product and coproduct are homogeneous maps. A *character* of H is an algebra morphism $\zeta : H \rightarrow \mathbb{Q}$, and the pair (H, ζ) is called a *combinatorial Hopf algebra*. A morphism $f : (H', \zeta') \rightarrow (H, \zeta)$ of combinatorial Hopf algebras is a morphism of graded Hopf algebras $f : H' \rightarrow H$ such that $\zeta' = \zeta \circ f$.

Example 6. Let \mathcal{P} be the \mathbb{Q} -vector space with basis consisting of all isomorphism classes of graded posets. We define a product on \mathcal{P} by $P_1 \cdot P_2 := P_1 \times P_2$, the Cartesian product of posets, and coproduct by $\Delta(P) = \sum_{x \in P} [\hat{0}, x] \otimes [x, \hat{1}]$. The unit element of \mathcal{P} is the poset $\mathbf{1}$ with one element $\hat{0} = \hat{1}$, and the counit is $\epsilon(P) = \delta_{P, \mathbf{1}}$. See, for example, [30]. If we take ζ to be the usual zeta function for posets, defined by $\zeta(P) = 1$ for all posets P , the pair (\mathcal{P}, ζ) is called the *combinatorial Hopf algebra of posets* [2].

The Hopf algebra $QSym$ becomes a combinatorial Hopf algebra with the canonical character $\zeta_{\mathbb{Q}}$ defined by $\zeta_{\mathbb{Q}}(M_{\alpha}) = 1$ if $\alpha = (n)$, $n \geq 0$, $\zeta_{\mathbb{Q}}(M_{\alpha}) = 0$ otherwise. The main result [2, Theorem 4.1] is that the combinatorial Hopf algebra $(QSym, \zeta_{\mathbb{Q}})$ is a terminal object in the category of combinatorial Hopf algebras, that is, for any combinatorial Hopf algebra (H, ζ_H) , there is a unique

morphism $F : (H, \zeta_H) \rightarrow (QSym, \zeta_Q)$. For $(H, \zeta_H) = (\mathcal{P}, \zeta)$ from Example 6, the morphism F is the one given in (7).

Further, each combinatorial Hopf algebra (H, ζ_H) has a special subalgebra Π_H , called the *odd subalgebra*, and the morphism F satisfies $F(\Pi_H) \subseteq \Pi_{QSym}$ [2, Proposition 6.1]. Now $\Pi_{QSym} = \Pi$, the peak algebra with basis given in (8), and $\Pi_{\mathcal{P}}$ contains the subalgebra of all Eulerian posets. Together, this gives another proof of Theorem 3.5.

The author and Aguiar are currently working to extend the theory of combinatorial Hopf algebras to the case of nonhomogeneous polynomial characters. One outcome is an alternate definition of the complete quasisymmetric function \tilde{F} defined in (9) and a new proof of Theorem 4.4.

References

- [1] M. Aguiar, Infinitesimal Hopf algebras and the **cd**-index of polytopes, *Discrete Comput. Geometry* **27** (2002), 3–28.
- [2] M. Aguiar, N. Bergeron, and F. Sottile, Combinatorial Hopf algebras and generalized Dehn-Sommerville relations, *Compos. Math.* **142** (2006), 1–30.
- [3] M.M. Bayer, The extended f -vectors of 4-polytopes, *J. Combin. Theory Ser. A* **44** (1987), 141–151.
- [4] M.M. Bayer and L.J. Billera, Counting faces and chains in polytopes and posets, in *Combinatorics and Algebra*, C. Greene, ed., Contemporary Mathematics, vol. 34, Amer. Math. Soc., Providence, 1984.
- [5] M.M. Bayer and L.J. Billera, Generalized Dehn-Sommerville relations for polytopes, spheres and Eulerian partially ordered sets, *Inventiones Math.* **79** (1985), 143–157.
- [6] M.M. Bayer and A. Klapper, A new index for polytopes, *Discrete Comput. Geometry* **6** (1991), 33–47.
- [7] M.M. Bayer and R. Ehrenborg, The toric h -vector of partially ordered sets, *Trans. Amer. Math. Soc.* **352** (2000), 4515–4531.
- [8] M.M. Bayer and G. Hetyei, Flag vectors of Eulerian partially ordered sets, *Europ. J. Combinatorics* **22** (2001), 5–26.
- [9] M.M. Bayer and G. Hetyei, Generalizations of Eulerian partially ordered sets, flag numbers, and the Möbius function, *Discrete Math.* **256** (2002), 577–593.
- [10] N. Bergeron, S. Mykytiuk, F. Sottile and S. van Willigenburg, Non-commutative Pieri operators on posets, *J. Comb. Theory Ser. A* **91** (2000), 84–110.
- [11] L.J. Billera and A. Björner, Face numbers of polytopes and complexes, *Handbook of Discrete and Computational Geometry*, J.E. Goodman and J. O'Rourke, eds., CRC Press, Boca Raton and New York, 1997.
- [12] L.J. Billera and F. Brenti, Quasisymmetric functions and Kazhdan-Lusztig polynomials, *Israel Journal of Mathematics* (to appear), arXiv:0710.3965.

-
- [13] L.J. Billera and R. Ehrenborg, Monotonicity of the \mathbf{cd} -index for polytopes, *Math. Z.* **233** (2000), 421–441.
- [14] L.J. Billera, R. Ehrenborg and M. Readdy, The \mathbf{cd} -index of zonotopes and arrangements, *Mathematical Essays in Honor of Gian-Carlo Rota*, B.E. Sagan and R.P. Stanley, eds., Birkhäuser, Boston, 1998, pp. 23–40.
- [15] L.J. Billera and G. Hetyei, Linear inequalities for flags in graded posets, *J. Comb. Theory Ser. A* **89** (2000), 77–104.
- [16] L.J. Billera, S.K. Hsiao and S. van Willigenburg, Peak quasisymmetric functions and Eulerian enumeration, *Adv. in Math.* **176** (2003), 248–276.
- [17] L.J. Billera and C.W. Lee, Sufficiency of McMullen’s Conditions for f -vectors of Simplicial Polytopes, *Bull. (New Series) Amer. Math. Soc.* **2** (1980), 181–185.
- [18] L.J. Billera and C.W. Lee, A Proof of the Sufficiency of McMullen’s Conditions for f -vectors of Simplicial Convex Polytopes, *Jour. Combinatorial Theory (A)* **31** (1981), 237–255.
- [19] L.J. Billera and N. Liu, Noncommutative enumeration in graded posets, *J. Alg. Combinatorics* **12** (2000), 7–24.
- [20] A. Björner and F. Brenti, *Combinatorics of Coxeter Groups*, Springer, New York, 2005.
- [21] T.C. Braden, Remarks on the combinatorial intersection cohomology of fans, *Pure Appl. Math. Q.* **2** (2006), no. 4, part 2, 1149–1186.
- [22] T.C. Braden and R. MacPherson, Intersection homology of toric varieties and a conjecture of Kalai, *Comment. Math. Helv.* **74** (1999), 442–455.
- [23] F. Brenti, Lattice paths and Kazhdan-Lusztig polynomials, *Jour. Amer. Math. Soc.* **11** (1998), 229–259.
- [24] F. Brenti, F. Caselli, and M. Marietti, Special matchings and Kazhdan-Lusztig polynomials, *Advances in Math.* **202** (2006), 555–601.
- [25] F. Caselli, Non-negativity properties of R -polynomials, *European J. Combin.* **27** (2006), 1005–1021.
- [26] V.V. Deodhar, A note on subgroups generated by reflections in Coxeter groups, *Arch. Math.* **53** (1989), 543–546.
- [27] M. Dyer, Reflection subgroups of Coxeter systems, *J. Algebra* **135** (1990), 57–73.
- [28] M. Dyer, On the “Bruhat graph” of a Coxeter system, *Compos. Math.*, **78** (1991), 185–191.
- [29] M. Dyer, Hecke algebras and shellings of Bruhat intervals, *Compos. Math.*, **89** (1993), 91–115.
- [30] R. Ehrenborg, On posets and Hopf algebras, *Adv. in Math.* **119** (1996), 1–25.
- [31] R. Ehrenborg, Lifting inequalities for polytopes, *Adv. in Math.* **193** (2005), 205–222.
- [32] R. Ehrenborg and K. Karu, Decomposition theorem for the \mathbf{cd} -index of Gorenstein posets. *J. Algebraic Combin.* **26** (2007), 225–251.
- [33] I. M. Gel’fand, D. Krob, A. Lascoux, B. Leclerc, V. Retakh and J.-Y. Thibon, Noncommutative symmetric functions, *Adv. in Math.* **112** (1995), 218–348.

-
- [34] I.M. Gessel, Multipartite P -partitions and inner products of Schur functions, in *Combinatorics and Algebra*, C. Greene, ed., Contemporary Mathematics, vol. 34, Amer. Math. Soc., Providence, 1984.
- [35] J. E. Humphreys, *Reflection Groups and Coxeter Groups*, Cambridge University Press, Cambridge, 1990.
- [36] G. Kalai, A new basis of polytopes, *J. Comb. Theory Ser. A* **49** (1988), 191–208.
- [37] K. Karu, Hard Lefschetz theorem for nonrational polytopes, *Inventiones Math.* **157** (2004), 419–447.
- [38] K. Karu, The \mathbf{cd} -index of fans and posets, *Compos. Math.*, **142** (2006), 701–718.
- [39] K. Karu, Lefschetz decomposition and the \mathbf{cd} -index of fans. *Proceedings of Gökova Geometry-Topology Conference 2005*, 59–74, Gökova Geometry/Topology Conference (GGT), Gökova, 2006.
- [40] D. Kazhdan and G. Lusztig, Representations of Coxeter groups and Hecke algebras, *Inventiones Math.* **53** (1979), 165–184.
- [41] P. McMullen, The maximum numbers of faces of a convex polytope, *Mathematika* **17** (1970), 179–184.
- [42] P. McMullen, The numbers of faces of simplicial polytopes, *Israel J. Math.* **9** (1971), 559–570.
- [43] K. Nyman, Linear inequalities for rank 3 geometric lattices, *Discrete Comput. Geom.* **31** (2004), 229–242.
- [44] N. Reading, The \mathbf{cd} -index of Bruhat intervals, *Electron. J. Combin.* **11** (2004), Research Paper 74.
- [45] R. Stanley, The upper bound conjecture and Cohen-Macaulay rings, *Studies in Appl. Math.* **54** (1975), 135–142.
- [46] R. Stanley, Balanced Cohen-Macaulay complexes, *Trans. Amer. Math. Soc.* **249** (1979), 139–157.
- [47] R. Stanley, The number of faces of a simplicial convex polytope, *Adv. in Math.* **35** (1980), 236–238.
- [48] R. Stanley, Generalized H -vectors, intersection cohomology of toric varieties, and related results, *Adv. Stud. Pure Math.* **11** (1987), 187–213.
- [49] R. Stanley, Flag f -vectors and the \mathbf{cd} -index, *Math. Z.* **216** (1994), 483–499.
- [50] R. Stanley, A survey of Eulerian posets, in *Polytopes: abstract, convex and computational*, T. Bisztriczky, P. McMullen, R. Schneider, and A. Ivic Weiss, eds., NATO Advanced Science Institute Series, vol. C 440, Kluwer Acad. Publ., Dordrecht, 1994.
- [51] R. Stanley, *Enumerative Combinatorics, Vol. 1*, Cambridge Studies in Advanced Mathematics, Vol. 49, Cambridge University Press, Cambridge, UK, 1997.
- [52] R. Stanley, *Enumerative Combinatorics, Vol. 2*, Cambridge Studies in Advanced Mathematics, Vol. 62, Cambridge University Press, Cambridge, UK, 1999.
- [53] R. Stanley, Recent developments in algebraic combinatorics, *Israel J. Math.* **143** (2004), 317–339.

-
- [54] E. Steinitz, Über the Eulerschen polyederrelationen, *Archiv für Mathematik und Physik* **11** (1906), 86–88.
- [55] J. Stembridge, Enriched P -partitions, *Trans. Amer. Math. Soc.* **349** (1997), 763–788.
- [56] C. Stenson, Families of tight inequalities for polytopes, *Discrete Comput. Geom.* **34** (2005), 507–521.
- [57] D.-N. Verma, Möbius inversion for the Bruhat order on a Weyl group, *Ann. Sci. École Norm. Sup.* **4** (1971), 393–398.

Order and Disorder in Energy Minimization

Henry Cohn*

Abstract

How can we understand the origins of highly symmetrical objects? One way is to characterize them as the solutions of natural optimization problems from discrete geometry or physics. In this paper, we explore how to prove that exceptional objects, such as regular polytopes or the E_8 root system, are optimal solutions to packing and potential energy minimization problems.

Mathematics Subject Classification (2010). Primary 05B40, 52C17; Secondary 11H31.

Keywords. Symmetry, potential energy minimization, sphere packing, E_8 , Leech lattice, regular polytopes, universal optimality.

1. Introduction

1.1. Genetics of the regular figures. Symmetry is all around us, both in the physical world and in mathematics. Of course, only a few of the many possible symmetries are ever actually realized, but we see more of them than we seemingly have any right to expect: symmetry is by its very nature delicate, and easily disturbed by perturbations. It is no great surprise to see carefully designed, symmetrical artifacts, but it is remarkable that nature can ever produce similar effects robustly, for example in snowflakes. Any occurrence of symmetry not deliberately imposed demands an explanation.

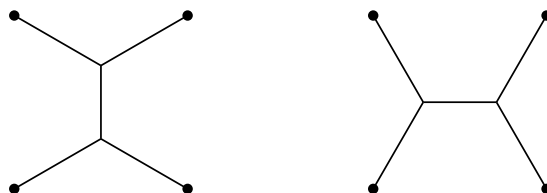
László Fejes Tóth proposed to seek the origins of symmetry in optimization problems. He referred to the *genetics of the regular figures*, in which “regular arrangements are generated from unarranged, chaotic sets by the ordering effect of an economy principle, in the widest sense of the word” [28]. It is not enough

*Microsoft Research New England, One Memorial Drive, Cambridge, MA 02142, USA.
E-mail: cohn@microsoft.com.

simply to classify the possible symmetries; we must go further and identify the circumstances in which they arise naturally.

Over the last century mathematicians have made enormous progress in identifying possible symmetry groups. We have classified the simple Lie algebras and finite simple groups, and although there is much left to learn about group theory and representation theory, our collective knowledge is both extensive and broadly applicable. Unfortunately, our understanding of the genetics of the regular figures lags behind. Much is known, but far more remains to be discovered, and many natural questions seem totally intractable.

Optimization provides a framework for this problem. How much symmetry and order should we expect in the solution of an optimization problem? It is natural to guess that the solutions of a highly symmetric problem will inherit the symmetry of the problem, but that is not always the case. For a toy example, consider the Steiner tree problem for a square, i.e., how to connect all four vertices of a square to each other via curves with minimal total length. The most obvious guess connects the vertices by an X, which displays all the symmetries of the square, but it is suboptimal. Instead, in the optimal solutions the branches meet in threes at 120° angles (this is a two-dimensional analogue of the behavior of soap films):



Note that the symmetry of the square is broken in each individual solution, but of course the set of both solutions retains the full symmetry group.

It is tempting to use symmetry to help solve problems, or at least to guess the answers, but as the Steiner tree example shows, this approach can be misleading. One of the most famous mistaken cases was the Kelvin conjecture on how to divide three-dimensional space into infinitely many equal volumes with minimal surface area between them, to create a foam of soap bubbles. In 1887 Kelvin conjectured a simple, symmetrical solution, obtained by deforming a tiling of space with truncated octahedra. (The deformation slightly curves the hexagonal facets into monkey saddles, so that the foam has the appropriate dihedral angles.) Kelvin's conjecture stood unchallenged for more than a century, but in 1994 Weaire and Phelan found a superior solution with two irregular types of bubbles¹ [54]. This shows the danger of relying too much on symmetry: sometimes it is a crucial clue as to the true optimum, but sometimes it leads in the wrong direction.

¹Their foam structure was the inspiration for the Beijing National Aquatics Center, used in the 2008 Olympics.

In many cases the symmetries that are broken are as interesting as the symmetries that are preserved. For example, crystals preserve some of the translational symmetries of space, but they dramatically break rotational symmetry, as well as most translational symmetries. This symmetry breaking is remarkable, because it entails long-range coordination: somehow widely separated pieces of the crystal nevertheless align perfectly with each other. A complete theory of crystal formation must therefore deal with how this coordination could come about. Here, however, we will focus on optimization problems and their solutions, rather than on the physical or algorithmic processes that might lead to these solutions.

1.2. Exceptional symmetry: E_8 and the Leech lattice. Certain mathematical objects, such as the icosahedron, have always fascinated mathematicians with their elegance and symmetry. These objects stand out as extraordinary and have inspired much deep mathematics (see, for example, Felix Klein's *Lectures on the Icosahedron* [34]). They are the sorts of objects one hopes to characterize and understand via the genetics of the regular figures.

These objects are often exceptional cases in classification theorems. In many different branches of mathematics, highly structured or symmetric objects can be classified into several regular, predictable families together with a handful of exceptions, such as the exceptional Lie algebras or sporadic finite simple groups. For most applications, the infinite families play the leading role, and one might be tempted to dismiss the exceptional cases as aberrations of limited importance, specific to individual problems. Instead, although they are indeed peculiar, the exceptional cases are not merely isolated examples, but rather recurring themes throughout mathematics, with the same exceptions occurring in seemingly unrelated problems. This phenomenon has not yet been fully understood, although much is known about particular cases.

For example, *ADE* classifications (i.e., simply-laced Dynkin diagrams) occur in many different mathematical areas, including finite subgroups of the rotation group $SO(3)$, representations of quivers of finite type, certain singularities of algebraic hypersurfaces, and simple critical points of multivariate functions. In each case, there are two infinite families, denoted A_n and D_n , and three exceptions E_6 , E_7 , and E_8 , with each type naturally described by a certain Dynkin diagram. See [31] for a survey. This means E_8 , for example, has a definite meaning in each of these problems. For example, among rotation groups it corresponds to the icosahedral group, and among simple critical points of functions from \mathbb{R}^n to \mathbb{R} it corresponds to the behavior of $x_1^3 + x_2^5 + x_3^2 + x_4^2 + \cdots + x_n^2$ at the origin.

In this survey, we focus primarily on two exceptional structures, namely the E_8 root lattice in \mathbb{R}^8 and the Leech lattice in \mathbb{R}^{24} . These objects bring together numerous mathematical topics, including sphere packings, finite simple

groups, combinatorial and spherical designs, error-correcting codes, lattices and quadratic forms, mathematical physics, harmonic analysis, and even hyperbolic and Lorentzian geometry. They are far too rich and well connected to do justice to here; see [24] for a much longer account as well as numerous references. Here, we will examine how to characterize E_8 and the Leech lattice, as well as some of their relatives, by optimization problems. These objects are special because they solve not just a single problem, but rather a broad range of problems. This level of breadth and robustness helps explain the widespread occurrences of these structures within mathematics. At the same time, it highlights the importance of understanding which problems have extraordinarily symmetric solutions and which do not.

1.3. Energy minimization. Much of physics is based on the idea of energy minimization, which will play a crucial role in this article. In many systems energy dissipates through forces such as friction, or more generally through heat exchange with the environment. Exact energy minimization will occur only at zero temperature; at positive temperature, a system in contact with a heat bath (a vast reservoir at a constant temperature, and with effectively infinite heat capacity) will equilibrate to the temperature of the heat bath, and its energy will fluctuate randomly, with its expected value increasing as the temperature increases.

One can describe the behavior of such a system mathematically using *Gibbs measures*, which are certain probability distributions on its states. For simplicity, imagine a system with n different states numbered 1 through n , where state i has energy E_i . For each possible expected value \bar{E} for energy, the corresponding Gibbs measure is the maximal entropy probability measure constrained to have expected energy \bar{E} . In other words, it assigns probability p_i to state i so that the entropy $\sum_{i=1}^n -p_i \log p_i$ is maximized subject to $\sum_{i=1}^n p_i E_i = \bar{E}$. (For the motivation behind the definition of entropy, see [33].)

A Lagrange multiplier argument shows that when $\min_i E_i < \bar{E} < \max_i E_i$, the probability p_i must equal $e^{-\beta E_i} / \sum_{j=1}^n e^{-\beta E_j}$ for some constant β , where β is chosen so that the expected energy equals \bar{E} . In physics terms, β is proportional to the reciprocal of temperature, and only nonnegative values of β are relevant (because energy is usually not bounded above, as it is in this toy model). As the temperature tends to infinity, β tends to zero and the system will be equidistributed among all states. As the temperature tends to zero, β tends to infinity, and the system will remain in its *ground states*, i.e., those with the lowest possible energy.

In this article, we will focus on systems of point particles interacting via a pair potential function. In other words, the energy of the system is the sum over all pairs of particles of some function depending only on the relative position of the pair (typically the distance between them). For example, in classical electrostatics, it is common to study identical charged particles interacting via

the Coulomb potential, i.e., with potential energy $1/r$ for a pair of particles at distance r .

Many other mathematical problems can be recast in this form, even sometimes in ways that are not immediately apparent. For a beautiful although tangential example, consider the distribution of eigenvalues for a random $n \times n$ unitary matrix, chosen with respect to the Haar measure on $U(n)$. These eigenvalues are unit complex numbers z_1, \dots, z_n , and the Weyl integral formula says that the induced probability measure on them has density proportional to

$$\prod_{1 \leq i < j \leq n} |z_i - z_j|^2$$

(see [27]). If we define the logarithmic potential $-\log |z_i - z_j|$ between z_i and z_j , then this measure is the Gibbs measure with $\beta = 2$ for n particles on the unit circle. The logarithmic potential is natural because it is a harmonic function on the plane (much as the Coulomb potential $x \mapsto 1/|x|$ is harmonic in three dimensions). Thus, the eigenvalues of a random unitary matrix repel each other through harmonic interactions, and the Weyl integral formula specifies the temperature $1/\beta$.

In the following survey, we will focus on the case of zero temperature. In the real world, all systems have positive temperature, which raises important questions about dynamics and phase transitions. However, for the purposes of understanding the role of symmetry, zero temperature is a crucial case.

1.4. Packing and information theory. The prototypical packing problem is sphere packing: how can one arrange non-overlapping, congruent balls in Euclidean space to fill as large a fraction of space as possible? The fraction of space filled is the *density*. Of course, it must be defined by a limiting process, by looking at the fraction of a large ball or cube that can be covered.

Packing problems fit naturally into the energy minimization framework via *hard-core potentials*, which are potentials that are infinite up to a certain radius r and zero at or beyond it. In other words, there is an infinite energy penalty for points that are too close together, but otherwise there is no effect. Under such a potential function, a collection of particles has finite energy if and only if the particles are positioned at the centers of non-overlapping balls of radius $r/2$. Note that every packing (not just the densest) minimizes energy, but knowing the minimal energy for all densities solves the packing problem.

From this perspective, one can formulate questions that are even deeper than densest packing questions. For example, at any fixed density, one can ask for a random packing at that density (i.e., a sample from the Gibbs measure at zero temperature). For which densities is there long-range order, i.e., nontrivial correlations between distant particles? In two or three dimensions, the densest packings are crystalline, and there appears to be considerable order even below the maximal density, with a phase transition between order and disorder as the density decreases. (See [41] and the references cited therein for more details.)

It is far from clear what happens in high dimensions, and the densest packings might be disordered [51].

Packings of less than maximal density are of great importance for modeling granular materials, because most such materials will be somewhat loose. The fact that long-range order seemingly persists over a range of densities means it can potentially be observed in the real world, where even under high pressure no packing is ever truly perfect. (Of course, for realistic models there are many other important refinements, such as variation in particle sizes and shapes.)

In addition to being models for granular materials, packings play an important role in information theory, as error-correcting codes for noisy communication channels. Suppose, for a simplified example, that we wish to communicate by radio. We can measure the signal strength at n different frequencies and represent it as an n -dimensional vector. Note that n may be quite large, so high-dimensional packings are especially important here. The power required to transmit a signal $x \in \mathbb{R}^n$ will be proportional to $|x|^2$, so we must restrict our attention to signals that lie within a ball of radius r centered at the origin, where r depends on the power level of our transmitter.

If we transmit a signal, then the received signal will be slightly perturbed due to noise. We can measure the noise level of the channel by ε , so that when x is transmitted, with high probability the received signal x' will satisfy $|x - x'| < \varepsilon$. In other words, if the open balls of radius ε about signals x and y do not overlap, then with high probability the received signals x' and y' cannot be confused.

To ensure error-free communication, we will rely on a restricted vocabulary of possible signals that cannot be confused with each other (i.e., an error-correcting code). That means they must be the centers of non-overlapping balls of radius ε . For efficient communication, we wish to maximize the number of signals available for use, i.e., the number of such balls whose centers lie within a ball of radius r . In the limit as r/ε tends to infinity, that is the sphere packing problem.

1.5. Outline. The remainder of the paper is organized as follows. Sections 2 and 3 survey packing and energy minimization problems in more depth. Sections 4 and 5 outline the proofs that certain exceptional objects solve these problems. Finally, Section 6 offers areas for future investigation.

2. Packings and Codes

2.1. Sphere packing in low and high dimensions. One can study the sphere packing problem in any dimension. In \mathbb{R}^1 it is trivial, because the line can be completely covered with intervals. In \mathbb{R}^2 , it is easy to guess that a hexagonal arrangement of circles is optimal, with each circle tangent to six others, but giving a rigorous proof of optimality is not completely

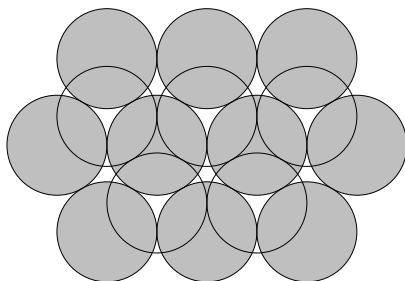


Figure 1. Two layers in a three-dimensional sphere packing, one denoted by shaded circles and the other by unshaded circles. Notice that the unshaded layer sits above half of the holes in the shaded layer.

straightforward and was first achieved in 1892 by Thue [50] (see [29] for a short, modern proof). In \mathbb{R}^3 , the usual way oranges are stacked in grocery stores is optimal, but the proof is extraordinarily difficult. Hales completed a proof in 1998, with a lengthy combination of human reasoning and computer calculations [30]. One conceptual difficulty is that the solution is not at all unique in \mathbb{R}^3 . In a technical sense, it is not unique in any dimension (even up to isometries), because density is a global property that is unchanged by, for example, removing a ball. However, in three dimensions there is a much deeper sort of non-uniqueness. One can form an optimal packing by stacking hexagonal layers, with each layer nestled into the gaps in the layer beneath it. As shown in Figure 1, the holes in a hexagonal lattice consist of two translates of the original lattice, and the next layer will sit above one of these two translates. For each layer, a binary choice must be made, and there are uncountably many ways to make these choices. (Each will be isometric to countably many others, but there remain uncountably many geometrically distinct packings, with many different symmetry groups.) All these packings are equally dense and perfectly natural. See [22] for a discussion of this issue in higher dimensions.

In four or more dimensions, no sharp density bounds are known. Instead, we merely have upper and lower bounds, which differ by a substantial factor. For example, in \mathbb{R}^{36} , the best upper bound known is more than 58 times the density of the best packing known [16]. This factor grows exponentially with the dimension: the best lower bound known is a constant times $n2^{-n}$ in \mathbb{R}^n (see [7] and [52]), while the upper bound is $(1.514724\dots + o(1))^{-n}$ (see [32]).

It may be surprising that these densities are so low. One way to think about it is in terms of volume growth in high dimensions. An ε -neighborhood of a ball in \mathbb{R}^n has volume $(1 + \varepsilon)^n$ times that of the ball, so when n is large, there is far more volume near the surface of the ball than actually inside it. In low-dimensional sphere packings, most volume is contained within the balls, with a narrow fringe of gaps between them. In high-dimensional packings, the gaps occupy far more volume.

It is easy to prove a lower bound of 2^{-n} for the sphere packing density in \mathbb{R}^n . In fact, this lower bound holds for every *saturated packing* (i.e., one in which there is no room for any additional spheres):

Lemma 2.1. *Every saturated sphere packing in \mathbb{R}^n has density at least 2^{-n} .*

Proof. Suppose the packing uses spheres of radius r . No point in space can be distance $2r$ or further from the nearest sphere center, since otherwise there would be room to center another sphere of radius r at that point. This means we can cover space completely by doubling the radius of each sphere. Doubling the radius multiplies the volume by 2^n , and hence multiplies the density by at most 2^n (in fact, exactly 2^n if we count overlaps with multiplicity). Because the enlarged spheres cover all of space, the original spheres must cover at least a 2^{-n} fraction. \square

This argument sounds highly constructive (simply add more spheres to a packing until it becomes saturated), and indeed it is constructive in the logical sense. However, in practice it offers almost no insight into what dense packings look like, because it is difficult even to tell whether a high-dimensional packing is saturated.

In fact, it is completely unclear how to construct dense packings in high dimensions. One might expect the sphere packing problem to have a simple, uniform solution that would work in all dimensions. Instead, each dimension has its own charming idiosyncrasies, as we will see in Section 2.2. There is little hope of a systematic solution to the sphere packing problem in all dimensions. Even achieving density 2^{-n} through a simple, explicit construction is an unsolved problem.

2.2. Lattices and periodic packings. The simplest sorts of packings are lattice packings. Recall that a *lattice* in \mathbb{R}^n is the integral span of a basis (i.e., it is a grid, possibly skewed). To form a sphere packing, one can center a sphere at each lattice point. The radius should be half the minimal distance between lattice points, so that the nearest spheres are tangent to each other.

There is no reason to expect that lattice packings should be the densest sphere packings, and they are probably not optimal in sufficiently high dimensions (for example, ten dimensions). However, lattices are very likely optimal in \mathbb{R}^n for $n \leq 9$ and for some higher values of n (including 12, 16, and 24). See [24] for more details about lattices and packings in general.

For $n \leq 8$ and $n = 24$, the lattice packing problem has been solved in \mathbb{R}^n . In fact, the densest lattices are unique in these dimensions (up to scaling and isometries), although that may not be true in every dimension, such as $n = 25$. For $n \leq 8$, the optimal lattices are all root lattices, the famous lattices that arise in Lie theory and are classified by Dynkin diagrams. Specifically, the densest lattices are A_1 (the integer lattice), A_2 (the hexagonal lattice), A_3 (the face-centered cubic lattice, which is also isomorphic to D_3), D_4 , D_5 , E_6 , E_7 ,

and E_8 . For $n = 24$, the Leech lattice is an optimal lattice packing; the proof will be discussed in Section 5.

The D_n lattices are particularly simple, because they are formed by a checkerboard construction as a sublattice of index 2 in \mathbb{Z}^n :

$$D_n = \{(x_1, \dots, x_n) \in \mathbb{Z}^n : x_1 + \dots + x_n \equiv 0 \pmod{2}\}.$$

To see why D_n is not optimal in high dimensions, consider the *holes* in D_n , i.e., the points in space that are local maxima for distance from the lattice. The integral points with odd coordinate sum are obvious candidates, and they are indeed holes, at distance 1 from D_n . However, there's a slightly more subtle case, namely the point $(1/2, 1/2, \dots, 1/2)$ and its translates by D_n . These points are at distance

$$\sqrt{\left(\frac{1}{2}\right)^2 + \dots + \left(\frac{1}{2}\right)^2} = \sqrt{n/4}$$

from D_n . When $n = 8$, this distance becomes $\sqrt{2}$, which is equal to the minimal distance between points in D_8 . That means these deep holes have become large enough that additional spheres can be placed in them. Doing so yields the E_8 root lattice, whose density is twice that of D_8 . (The E_6 and E_7 lattices are certain cross sections of E_8 .)

The E_8 and Leech lattices stand out among lattice packings, because all the spheres fit beautifully into place in a remarkably dense and symmetric way. There is no doubt that they are optimal packings in general, not just among lattices. Harmonic analysis ought to provide a proof, but as we will see in Section 5, a full proof has been elusive.

Periodic packings form a broader class of packings than lattice packings. A lattice can be viewed as the vertices of a tiling of space with parallelotopes (fundamental domains for the action by translation), but there's no reason to center spheres only at the vertices. More generally, one can place them in the interior, or elsewhere on the boundary, and then repeat them periodically; such a packing is called a periodic packing. Equivalently, the sphere centers in a periodic packing form the union of finitely many translates of a lattice.

The E_8 packing, as defined above, is clearly periodic (the union of two translates of D_8). It is not quite as obvious that it is actually a lattice, but that is easy to check. The Leech lattice in \mathbb{R}^{24} can be defined by a similar, but more elaborate, construction involving filling in the holes in a lattice constructed using the binary Golay code (see [38] and Section 4.4 in Chapter 4 of [24]).

Philosophically, the construction of E_8 given above is somewhat odd, because E_8 itself is extraordinarily symmetrical, but the construction is not. Instead, it builds E_8 in two pieces. This situation is actually quite common when constructing a highly symmetric object. By neglecting part of the symmetry group, one can decompose the object into simpler pieces, which can each be understood separately. However, eventually one must exhibit the extra symmetry. The symmetry group of E_8 is generated by the reflections in the hyperplanes

orthogonal to the minimal vectors of E_8 , and one can check that it acts transitively on those minimal vectors.

It is not known whether periodic packings achieve the maximal packing density in every dimension. However, they always come arbitrarily close: given any dense packing, one can take a large, cubical piece of it and repeat that piece periodically. To avoid overlaps, it may be necessary to remove some spheres near the boundary, but if the cube is large enough, then the resulting decrease in density will be small.

By contrast, it is not even known whether there exist saturated lattice packings in high dimensions. If not, then lattices cannot achieve more than half the maximal density, because one can double the density of a non-saturated lattice by filling in a hole together with all its translates by lattice vectors. It seems highly unlikely that there are saturated lattices in high dimensions, because a lattice is specified by a quadratic number of parameters, while there is an exponential volume of space in which holes could appear, so there are not enough degrees of freedom to control all the possible holes. However, this argument presumably cannot be made rigorous.

Despite all the reasons to think lattices are not the best sphere packings in high dimensions, the best asymptotic lower bounds known for sphere packing density use lattices. Ball's bound $2(n-1)2^{-n}$ in \mathbb{R}^n holds for lattice packings [7], and Vance's bound, which improves it by an asymptotic factor of $3/e$ when n is a multiple of four, uses not just lattices, but lattices that are modules over a maximal order in the quaternions [52]. Imposing algebraic structure may rule out the densest possible packings, but it makes up for that by offering powerful tools for analysis and proof.

2.3. Packing problems in other spaces. Packing problems are interesting in many metric spaces. The simplest situation is when the ambient space is compact, in which case the packing will involve only finitely many balls. The packing problem can then be formulated in terms of two different optimization problems for a finite subset of the metric space:

1. What is the largest possible minimal distance between N points?
2. What is the largest possible size of a subset whose minimal distance is at least r ?

The first fixes the number of balls and maximizes their size, while the second fixes the radius $r/2$ of the balls and maximizes the number. In Euclidean space, if we interpret the number of points as the number of points per unit volume, then both problems are the same by scaling invariance, but that does not hold in compact spaces. The two problems are equivalent, however, in the sense that a complete answer to one (for all values of r or N) yields a complete answer to the other.

Packing problems arise naturally in many compact metric spaces, including spheres, projective spaces, Grassmannians [23, 4], and the Hamming cube

$\{0,1\}^n$ (under Hamming distance, so packings are binary error-correcting codes). For a simplified example, suppose one wishes to treat a spherical tumor by beaming radiation at it. One would like to use multiple beams approaching it from different angles, so as to minimize radiation exposure outside of the tumor, and the problem of maximizing the angle between the beams is a packing problem in \mathbb{RP}^2 .

Packing problems are also important in non-compact spaces, but aside from Euclidean space we will not deal with them in this article, because defining density becomes much more subtle. See, for example, the foundational work by Bowen and Radin on defining packing density in hyperbolic space [11].

Packings on the surface of a sphere are known as spherical codes. Specifically, an *optimal spherical code* is an arrangement of points on a sphere that maximizes the minimal distance among configurations of its size. Spherical codes can be used as error-correcting codes (for example, in the toy model of radio transmission from Section 1.4, they are codes for a constant-power transmitter), and they also provide an elegant way to help characterize the many interesting spherical configurations that arise throughout mathematics.

One of the most attractive special cases of packing on a sphere is the *kissing problem*. How many non-overlapping unit balls can all be tangent to a central unit ball? The points of tangency on the central ball form a spherical code with minimal angle at least 60° , and any such code yields a kissing configuration.

In \mathbb{R}^2 , the kissing number is clearly six, but the answer is already not obvious in \mathbb{R}^3 . The twelve vertices of an icosahedron work, but the tangent balls do not touch each other and can slide around. It turns out that there is no room for a thirteenth ball, but that was first proved only in 1953 by Schütte and van der Waerden [47].

In \mathbb{R}^4 , Musin [42] showed that the kissing number is 24, but the answer is not known in \mathbb{R}^5 (it appears to be 40). In fact, the only higher dimensions for which the kissing problem has been solved are 8 and 24, independently by Levenshtein [39] and by Odlyzko and Sloane [43]. The kissing numbers are 240 in \mathbb{R}^8 and 196560 in \mathbb{R}^{24} . Furthermore, these kissing configurations are unique up to isometries [9].

The kissing number of 240 is achieved by the E_8 root lattice through its 240 minimal vectors. Specifically, there are $\binom{8}{2} \cdot 2^2 = 112$ permutations of $(\pm 1, \pm 1, 0, \dots, 0)$ and $2^7 = 128$ vectors of the form $(\pm 1/2, \dots, \pm 1/2)$ with an even number of minus signs. Thus, E_8 is not only the densest lattice packing in \mathbb{R}^8 , but it also has the highest possible kissing number. Similarly, the Leech lattice in \mathbb{R}^{24} achieves the kissing number of 196560.

In general, however, there is no reason to believe that the densest packings will also have the highest kissing numbers. The packing density is a global property, while the kissing number is purely local and might be maximized in a way that cannot be extended to a dense packing. That appears to happen in many dimensions [24]. Instead of being typical, compatibility between the optimal local and global structures is a remarkable occurrence.

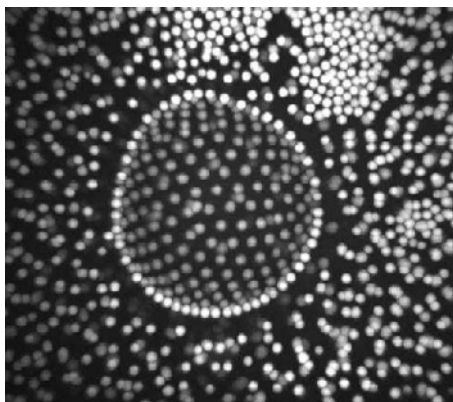


Figure 2. Energy minimization on an actual (approximate) sphere: tiny, electrically charged PMMA beads collecting on the interface between water and cyclohexyl bromide. [Courtesy of W. Irvine and P. M. Chaikin, New York University]

3. The Thomson Problem and Universal Optimality

3.1. Physics on surfaces. The Thomson problem [49, p. 255] asks for the minimal-energy configuration of N classical electrons confined to the unit sphere S^2 . In other words, the particles interact via the Coulomb potential $1/r$ at Euclidean distance r . This model was originally intended to describe atoms, before quantum mechanics or even the discovery of the nucleus. Thomson hoped it would explain the periodic table. Of course, subsequent discoveries have shown that it is a woefully inadequate atomic model, but it remains of substantial scientific interest, and its variants describe many real-world systems.

For example, imagine mixing together two immiscible liquids, such as oil and water. The oil will break up into tiny droplets, evenly dispersed within the water, but they will rapidly coalesce and the oil will separate from the water. Cooks have long known that one can prevent this separation by using emulsifiers. One type of emulsion is a Pickering emulsion, in which tiny particles collect on the boundaries of oil droplets, which prevents coalescence (the particles bounce off each other).

More generally, colloidal particles often adsorb to the interface between two different liquids. See, for example, Figure 2, which shows charged particles made of polymethyl methacrylate (i.e., plexiglas) in a mixture of water and cyclohexyl bromide. Notice that the particles on the surface of the droplet have spread out into a fairly regular arrangement due to their mutual repulsion, and they are repelling the remaining particles away from the surface.

These particles are microscopic, yet large enough that they can accurately be described using classical physics. Thus, the generalized Thomson problem is an appropriate model. See [12] for more details on these sorts of materials.

Consider the case of particles on the unit sphere in \mathbb{R}^n . Given a finite subset $\mathcal{C} \subset S^{n-1}$ and a *potential function* $f: (0, 4] \rightarrow \mathbb{R}$, define the *potential energy* by

$$E_f(\mathcal{C}) = \frac{1}{2} \sum_{\substack{x, y \in \mathcal{C} \\ x \neq y}} f(|x - y|^2).$$

For each positive integer N and each f , we seek an N -element subset $\mathcal{C} \subset S^{n-1}$ that minimizes $E_f(\mathcal{C})$ compared to all other choices of \mathcal{C} with $|\mathcal{C}| = N$. The use of squared distance instead of distance is not standard in physics, but it will prove mathematically convenient. The function f is defined only on $(0, 4]$ because no squared distance larger than 4 can occur on the unit sphere.

Typically f will be decreasing (so the force is repulsive) and convex. In fact, the most natural potential functions to use are the *completely monotonic* functions, i.e., smooth functions satisfying $(-1)^k f^{(k)} \geq 0$ for all integers $k \geq 0$. For example, inverse power laws $r \mapsto 1/r^s$ (with $s > 0$) are completely monotonic.

3.2. Varying the potential function. As we vary the potential function f above, how do the optimal configurations change? From the physics perspective, this question appears silly, because the potential is typically determined by fundamental physics. However, from a mathematical perspective it is a critical question, because it places the individual optimization problems into a richer context.

As we vary the potential function, the optimal configurations will vary in some family. This family may not be connected, because the optimum may abruptly jump as the potential function passes some threshold, and different components may have different dimensions [15]. Nevertheless, we can use the local dimension of the family as a crude measure of the complexity of an optimum: we compute the dimension of the space of perturbed configurations that minimize energy for perturbations of the potential function. Call this dimension the *parameter count* of the configuration.

Figure 3 (taken from [8]) shows the parameter counts for the configurations minimizing Coulomb energy on S^2 with 2 through 64 points. The figure is doubly conjectural: in almost all of these cases, no proof is known that the supposed optima are truly optimal or that the parameter counts are correct. However, the experimental evidence leaves little doubt.

One can see from Figure 3 that the parameter counts vary wildly. For example, for 43 points there are 21 parameters, while for 44 points there is only 1. This suggests that the 44-point optimizer will be substantially simpler and more understandable, and indeed it is (see [8]).

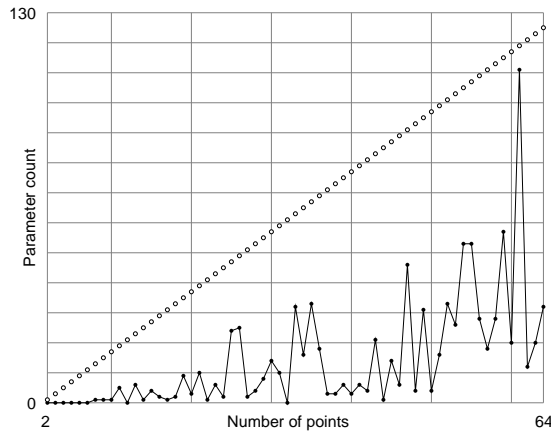


Figure 3. Parameter counts for conjectural Coulomb-energy minimizers on S^2 . For comparison, the white circles show the dimension of the space of all configurations.

3.3. Universal optimality. When one varies the potential function, the simplest case is when the optimal configuration never varies. Call a configuration *universally optimal* if it minimizes energy for all completely monotonic potential functions.

A universal optimum is automatically an optimal spherical code: for the potential function $f(r) = 1/r^s$ with s large, the energy is asymptotically determined by the minimal distance, and minimizing energy requires maximizing the minimal distance. However, optimal spherical codes are rarely universally optimal. For every number of points in every dimension, there exists some optimal code, but universal optima appear to be far less common.

In S^1 , there is an N -point universal optimum for each N , namely the vertices of a regular N -gon. In S^2 , the situation is more complicated. Aside from degenerate cases with three or fewer points, there are only three universal optima, namely the vertices of a regular tetrahedron, octahedron, or icosahedron [17]. The cube and dodecahedron are not even optimal, let alone universally optimal, since one can lower energy by rotating a facet.

The first case for which there is no universal optimum is five points in S^2 . There are two natural configurations: a triangular bipyramid, with an equilateral triangle on the equator together with the north and south poles, and a square pyramid, with its top at the north pole and its base slightly below the equator. This second family depends on one parameter, the height of the pyramid. The triangular bipyramid is known to minimize energy for several inverse power laws [48], but it is not even a local minimum when they are sufficiently steep, in which case square pyramids seem to become optimal.

Conjecture 3.1. *For every completely monotonic potential function, either the triangular bipyramid or a square pyramid minimizes energy among five-point configurations in S^2 .*

Table 1. The known N -point universal optima in S^{n-1} .

n	N	Description
n	$N \leq n + 1$	regular simplex
n	$2n$	regular cross polytope
2	N	regular N -gon
3	12	regular icosahedron
4	120	regular 600-cell
5	16	hemicube
6	27	Schläfli graph
7	56	28 equiangular lines
8	240	E_8 root system
21	112	isotropic subspaces
21	162	strongly regular graph
22	100	Higman-Sims graph
22	275	McLaughlin graph
22	891	isotropic subspaces
23	552	276 equiangular lines
23	4600	iterated kissing configuration
24	196560	Leech lattice minimal vectors
$q(q^3 + 1)/(q + 1)$	$(q + 1)(q^3 + 1)$	isotropic subspaces (q is a prime power)

For $n \geq 4$, the universal optima in S^{n-1} have not been completely classified. Table 1 shows a list of the known cases (proved in [17]). Each of them is a fascinating mathematical object. For example, the 27 points in S^5 correspond to the 27 lines on a cubic surface.

The first five lines in the table list the regular polytopes with simplicial facets. The next four lines list the E_8 root system and certain semiregular polytopes obtained as cross sections. The next eight lines list the minimal vectors of the Leech lattice and certain cross sections. If this were the complete list, it would feel reasonable, but the last line is perplexing. It describes another infinite sequence of universal optima, constructed from geometries over \mathbb{F}_q in [13] and recognized as optimal codes in [40]. How many more such cases remain to be constructed?

Another puzzling aspect of Table 1 is the gap between 8 and 21 dimensions. Are there really no universal optima in these dimensions, aside from the simplices and cross polytopes? Or do we simply lack the imagination needed to discover them? Extensive computer searches [8] suggest that the table is closer to complete than one might expect, but probably not complete. Specifically, there are a 40-point configuration in S^9 and a 64-point configuration in S^{13} that appear to be universally optimal, but these are the only conjectural cases that have been located.

Almost all of the results tabulated in Table 1 can be deduced from the following theorem. It generalizes a theorem of Levenshtein [40], which says that

these configurations are all optimal codes. The one known case not covered by the theorem is the regular 600-cell, which requires a different argument [17].

To state the theorem, we will need two definitions. A *spherical k -design* in S^{n-1} is a finite subset \mathcal{D} of the sphere such that for every polynomial $p: \mathbb{R}^n \rightarrow \mathbb{R}$ of total degree at most k , the average of p over \mathcal{D} equals its average over the entire sphere. Spherical k -designs can be thought of as sets giving quadrature rules (i.e., numerical integration schemes) that are exact for polynomials of degree up to k . An *m -distance set* is a set for which m distances occur between distinct points.

Theorem 3.2 (Cohn and Kumar [17]). *Every m -distance set that is a spherical $(2m - 1)$ -design is universally optimal.*

The proof of this theorem uses linear programming bounds, which are developed in the next section.

4. Proof Techniques: Linear Programming Bounds

4.1. Constraints on the pair correlation function. In this section, we will discuss techniques for proving lower bounds on potential energy. In particular, we will develop linear programming bounds and briefly explain how they are used to prove Theorem 3.2.

They are called “linear programming bounds” because linear programming can be used to optimize them, but no knowledge of linear programming is required to understand how the bounds work. They were originally developed by Delsarte for discrete problems in coding theory [25], extended to continuous packing problems in [26, 32], and adapted for potential energy minimization by Yudin and his collaborators [55, 35, 36, 1, 2]. In this section, we will focus on spherical configurations, although the techniques work in much greater generality.

Given a finite subset \mathcal{C} of S^{n-1} , define its *distance distribution* by

$$A_t = |\{(x, y) \in \mathcal{C}^2 : \langle x, y \rangle = t\}|,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product in \mathbb{R}^n . In physics terms, A is the pair correlation function; it measures how often each pairwise distance occurs (the inner product is a natural way to gauge distance on the sphere). Linear programming bounds are based on proving certain linear inequalities involving the numbers A_t . These inequalities are crucial because the potential energy can be expressed in terms of the distance distribution A by

$$E_f(\mathcal{C}) = \frac{1}{2} \sum_{\substack{x, y \in \mathcal{C} \\ x \neq y}} f(|x - y|^2) = \sum_{-1 \leq t < 1} \frac{f(2 - 2t)}{2} A_t, \quad (4.1)$$

since $|x - y|^2 = 2 - 2\langle x, y \rangle$. (Although (4.1) sums over uncountably many values of t , only finitely many of the summands are nonzero.) Energy is a linear function of A , and the linear programming bound is the minimum of this function subject to the linear constraints on A , which makes it the solution to a linear programming problem in infinitely many variables.

To begin, there are several obvious constraints on the distance distribution. Let $N = |\mathcal{C}|$. Then $A_t \geq 0$ for all t , $A_t = 0$ for $|t| > 1$, $A_1 = N$, and $\sum_t A_t = N^2$.

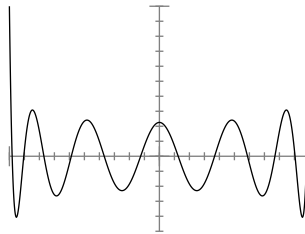
The power of linear programming bounds comes from less obvious constraints. For example, $\sum_t A_t t \geq 0$. To see why, notice that

$$\sum_t A_t t = \sum_{x,y \in \mathcal{C}} \langle x, y \rangle = \left| \sum_{x \in \mathcal{C}} x \right|^2 \geq 0.$$

More generally, there is an infinite sequence of polynomials (independent of \mathcal{C} , but depending on the dimension n) $P_0^n, P_1^n, P_2^n, \dots$, with $\deg P_k^n = k$, such that for each k ,

$$\sum_t A_t P_k^n(t) \geq 0. \tag{4.2}$$

(In fact, we can take $P_0^n(t) = 1$, $P_1^n(t) = t$, and $P_2^n(t) = t^2 - 1/n$.) This inequality is nontrivial, because these polynomials are frequently negative. For example, P_{12}^3 looks like this:



The polynomials P_k^n are called *ultraspherical polynomials*, and they are characterized by orthogonality on the interval $[-1, 1]$ with respect to the measure $(1 - t^2)^{(n-3)/2} dt$. In other words, for $i \neq j$,

$$\int_{-1}^1 P_i^n(t) P_j^n(t) (1 - t^2)^{(n-3)/2} dt = 0.$$

This relationship determines the polynomials up to scaling, as the Gram-Schmidt orthogonalization of the monomials $1, t, t^2, \dots$ with respect to this inner product. The sign of the scaling constant is determined by $P_k^n(1) > 0$, and the magnitude of the constant is irrelevant for (4.2).

In fact, these polynomials have a far stronger property than just (4.2): they are positive-definite kernels. That is, for any N and any points $x_1, \dots, x_N \in S^{n-1}$, the $N \times N$ matrix $(P_k^n(\langle x_i, x_j \rangle))_{1 \leq i, j \leq N}$ is positive semidefinite. This

implies (4.2) because the sum of the entries of a positive-semidefinite matrix is nonnegative. Schoenberg [45] proved that every continuous positive-definite kernel on S^{n-1} must be a nonnegative linear combination of ultraspherical polynomials.

4.2. Zonal spherical harmonics. As an illustration of the role of representation theory, in this section we will derive the ultraspherical polynomials as zonal spherical harmonics and verify that they satisfy (4.2). The reader who is willing to take that on faith can skip the derivation.

The orthogonal group $O(n)$ acts on S^{n-1} by isometries, and hence $L^2(S^{n-1})$ is a unitary representation of $O(n)$. To begin, we will decompose this representation into irreducibles. Let \mathcal{P}_k be the subspace of functions on S^{n-1} defined by polynomials on \mathbb{R}^n of total degree at most k . We have $\mathcal{P}_0 \subset \mathcal{P}_1 \subset \dots$, and each \mathcal{P}_k is a finite-dimensional representation of $O(n)$, with $\bigcup_k \mathcal{P}_k$ dense in $L^2(S^{n-1})$. To convert this filtration into a direct sum decomposition, let $V_0 = \mathcal{P}_0$ and define V_k to be the orthogonal complement of $V_0 \oplus V_1 \oplus \dots \oplus V_{k-1}$ within \mathcal{P}_k (with respect to the usual inner product on $L^2(S^{n-1})$). Then V_k is still preserved by $O(n)$, and the entire space breaks up as

$$L^2(S^{n-1}) = \widehat{\bigoplus_{k \geq 0} V_k}.$$

(The hat indicates the completion of the algebraic direct sum.) The functions in V_k are known as *spherical harmonics* of degree k , because V_k is an eigenspace of the spherical Laplacian, but we will not need that characterization of them.

For each $x \in S^{n-1}$, evaluating at x defines a linear map $f \mapsto f(x)$ on V_k . Thus, there exists a unique vector $v_{k,x} \in V_k$ such that for all $f \in V_k$,

$$f(x) = \langle f, v_{k,x} \rangle,$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product on V_k from $L^2(S^{n-1})$. The map $x \mapsto v_{k,x}$ is called a *reproducing kernel*.

For each $T \in O(n)$ and $f \in V_k$,

$$\langle f, v_{k,Tx} \rangle = f(Tx) = (T^{-1}f)(x) = \langle T^{-1}f, v_{k,x} \rangle = \langle f, Tv_{k,x} \rangle.$$

Thus, $Tv_{k,x} = v_{k,Tx}$, by the uniqueness of $v_{k,Tx}$. It follows that $v_{k,x}$ is invariant under the stabilizer of x in $O(n)$. In other words, it is invariant under rotations about the axis through $\pm x$, so it is effectively a function of only one variable, the inner product with x . Such a function is called a *zonal spherical harmonic*.

We can define P_k^n by

$$v_{k,x}(y) = P_k^n(\langle x, y \rangle).$$

These polynomials certainly satisfy (4.2), because

$$\sum_{x,y \in \mathcal{C}} P_k^n(\langle x, y \rangle) = \sum_{x,y \in \mathcal{C}} v_{k,x}(y) = \sum_{x,y \in \mathcal{C}} \langle v_{k,x}, v_{k,y} \rangle = \left| \sum_{x \in \mathcal{C}} v_{k,x} \right|^2 \geq 0,$$

and in fact they are positive-definite kernels because $(P_k^n(\langle x_i, x_j \rangle))_{1 \leq i, j \leq N}$ is the Gram matrix of the vectors v_{k, x_i} .

The functions $v_{0, x}, v_{1, x}, \dots$ are in orthogonal subspaces, and hence the polynomials P_0^n, P_1^n, \dots must be orthogonal with respect to the measure on $[-1, 1]$ obtained by projecting the surface measure of S^{n-1} onto the axis from $-x$ to x . The following simple calculation shows that the measure is proportional to $(1 - t^2)^{(n-3)/2} dt$. Consider the spherical shell defined by

$$1 \leq x_1^2 + \dots + x_n^2 \leq 1 + \varepsilon.$$

If we set $x_1 = t$, then the remaining coordinates satisfy

$$1 - t^2 \leq x_2^2 + \dots + x_n^2 \leq 1 - t^2 + \varepsilon,$$

and the volume is proportional to $(1 - t^2 + \varepsilon)^{(n-1)/2} - (1 - t^2)^{(n-1)/2}$. If we divide by ε to normalize, then as $\varepsilon \rightarrow 0$ we find that the density of the surface measure with $x_1 = t$ is proportional to $(1 - t^2)^{(n-3)/2}$, as desired.

The degree of P_k^n is at most k , and because $v_{k, x}$ is orthogonal to \mathcal{P}_{k-1} , the degree can be less than k only if P_k^n is identically zero. That cannot be the case (for $n > 1$), since otherwise evaluating at x would be identically zero. If it were, then it would follow from $Tv_{k, x} = v_{k, Tx}$ that evaluating at each point is identically zero, and thus that V_k is trivial. However, $\mathcal{P}_k \neq \mathcal{P}_{k-1}$, and hence V_k is nontrivial.

Thus, the polynomials P_k^n defined above have degree k , satisfy (4.2), and have the desired orthogonality relationship.

4.3. Linear programming bounds. Let $\mathcal{C} \subset S^{n-1}$ be a finite subset and let A be its distance distribution. To make use of the linear constraints on A discussed in Section 4.1, we will use the dual linear program. In other words, we will take linear combinations of the constraints so as to obtain a lower bound on energy.

We introduce new real variables α_k and β_t specifying which linear combination to take. Suppose we add α_0 times $\sum_t A_t = N^2$, α_k times

$$\sum_{-1 \leq t \leq 1} A_t P_k^n(t) \geq 0$$

(with $\alpha_k \geq 0$ for $k \geq 1$), and β_t times the constraint $A_t \geq 0$ (with $\beta_t \geq 0$ for $-1 \leq t < 1$). We find that

$$\sum_{-1 \leq t \leq 1} A_t \sum_k \alpha_k P_k^n(t) + \sum_{-1 \leq t < 1} A_t \beta_t \geq \alpha_0 N^2,$$

using the normalization $P_0^n(t) = 1$. Define $h(t) = \sum_k \alpha_k P_k^n(t)$. Then

$$\sum_{-1 \leq t < 1} A_t (h(t) + \beta_t) \geq \alpha_0 N^2 - h(1)N,$$

because $A_1 = N$. If we choose α_k and β_t so that $h(t) + \beta_t = f(2 - 2t)/2$ for $-1 \leq t < 1$, then the energy will be bounded below by $\alpha_0 N^2 - h(1)N$, by (4.1).

The equation $h(t) + \beta_t = f(2 - 2t)/2$ just means that $h(t) \leq f(2 - 2t)/2$ (because we have assumed only that $\beta_t \geq 0$). Thus, we have proved the following bound:

Theorem 4.1 (Yudin [55]). *Suppose $h(t) = \sum_k \alpha_k P_k^n(t)$ satisfies $\alpha_k \geq 0$ for $k > 0$ and $h(t) \leq f(2 - 2t)/2$ for $-1 \leq t < 1$. Then for every finite subset $\mathcal{C} \subset S^{n-1}$,*

$$E_f(\mathcal{C}) \geq \alpha_0 |\mathcal{C}|^2 - h(1)|\mathcal{C}|.$$

To prove Theorem 3.2, one can optimize the choice of the auxiliary function h in Theorem 4.1. Suppose \mathcal{C} is an m -distance set and a spherical $(2m - 1)$ -design, and f is completely monotonic. In the proof of Theorem 4.1, equality holds if and only if $h(t) = f(2 - 2t)/2$ for every inner product $t < 1$ that occurs between points in \mathcal{C} and $\sum_{x,y \in \mathcal{C}} P_k^n(\langle x, y \rangle) = 0$ whenever $\alpha_k > 0$ and $k > 0$. The latter equation automatically holds for $1 \leq k \leq 2m - 1$ because \mathcal{C} is a $(2m - 1)$ -design. Let h be the unique polynomial of degree at most $2m - 1$ that agrees with $f(2 - 2t)/2$ to order 2 at each of the m inner products between distinct points in \mathcal{C} , so that h satisfies the other condition for equality. The inequality $h(t) \leq f(2 - 2t)/2$ follows easily from a remainder theorem for Hermite interpolation (using the complete monotonicity of f). The most technical part of the proof is the verification that the coefficients α_k of h are nonnegative. For any single configuration, it can be checked directly; for the general case, see [17].

4.4. Semidefinite programming bounds. Semidefinite programming bounds, introduced by Schrijver [46] and generalized by Bachoc and Valentin [5], extend the idea of linear programming bounds by looking at triple (or even higher) correlation functions, rather than just pair correlations. Linear constraints are naturally replaced with semidefinite constraints, and the resulting bounds can be optimized by semidefinite programming.

This method is a far-reaching generalization of linear programming bounds, and it has led to several sharp bounds that could not be obtained previously [6, 21]. However, the improvement in the bounds when going from pairs to triples is often small, while the computational price is high. One of the most interesting conceptual questions in this area is the trade-off between higher correlations and improved bounds. When studying N -point configurations in S^{n-1} using k -point correlation bounds, how large does k need to be to prove a sharp bound? Clearly $k = N$ would suffice, and for the cases covered by Theorem 3.2 it is enough to take $k = 2$. Aside from a handful of cases in which $k = 3$ works, almost nothing is known in between. (Cases with $k \geq 4$ seem too difficult to handle computationally.) This question is connected more generally to the strength of LP and SDP hierarchies for relaxations of NP-hard combinatorial optimization problems [37].

It is also related to a conjecture of Torquato and Stillinger [51], who propose that for packings that are disordered (in a certain technical sense), in sufficiently high dimensions the two-point constraints are not only necessary but also sufficient for the existence of a packing with a given pair correlation function. They show that this conjecture would lead to packings of density $(1.715527\dots + o(1))^{-n}$ in \mathbb{R}^n , by exhibiting the corresponding pair correlation functions. The problem of finding a hypothetical pair correlation function that maximizes the packing density, subject to the two-point constraints, is dual to the problem of optimizing the linear programming bounds.

5. Euclidean Space

5.1. Linear programming bounds in Euclidean space. Linear programming bounds can also be applied to packing and energy minimization problems in Euclidean space, with Fourier analysis taking the role played by the ultraspherical polynomials in the spherical case. In this section, we will focus primarily on packing, before commenting on energy minimization at the end. The theory is formally analogous to that in compact spaces, but the resulting optimization problems are quite a bit deeper and more subtle, and the most exciting applications of the theory remain conjectures.

We will normalize the Fourier transform of an L^1 function $f: \mathbb{R}^n \rightarrow \mathbb{R}$ by

$$\widehat{f}(t) = \int_{\mathbb{R}^n} f(x) e^{2\pi i \langle t, x \rangle} dx.$$

(In this section, f will not denote a potential function.) The fundamental technical tool is the Poisson summation formula for a lattice Λ , which holds for all Schwartz functions (i.e., smooth functions all of whose derivatives are rapidly decreasing):

$$\sum_{x \in \Lambda} f(x) = \frac{1}{\text{vol}(\mathbb{R}^n/\Lambda)} \sum_{t \in \Lambda^*} \widehat{f}(t).$$

Here, $\text{vol}(\mathbb{R}^n/\Lambda)$ is the volume of a fundamental parallelotope, and Λ^* is the dual lattice defined by

$$\Lambda^* = \{t \in \mathbb{R}^n : \langle t, x \rangle \in \mathbb{Z} \text{ for all } x \in \Lambda\}.$$

Given any basis of Λ , the dual basis with respect to $\langle \cdot, \cdot \rangle$ is a basis of Λ^* .

Theorem 5.1 (Cohn and Elkies [16]). *Let $f: \mathbb{R}^n \rightarrow \mathbb{R}$ be a Schwartz function such that $\widehat{f}(0) \neq 0$. If $f(x) \leq 0$ for $|x| \geq 1$ and $\widehat{f}(t) \geq 0$ for all t , then the sphere packing density in \mathbb{R}^n is at most*

$$\frac{\pi^{n/2}}{2^n (n/2)!} \cdot \frac{f(0)}{\widehat{f}(0)}.$$

Of course, $(n/2)!$ means $\Gamma(n/2 + 1)$ when n is odd. The restriction to Schwartz functions can be replaced with milder assumptions [16, 17].

The hypotheses and conclusion of Theorem 5.1 are invariant under rotation about the origin, so without loss of generality we can symmetrize f and assume it is a radial function. Thus, optimizing the bound in Theorem 5.1 amounts to optimizing the choice of a function of one (radial) variable.

It is not hard to prove Theorem 5.1 for the special case of lattice packings. Suppose Λ is a lattice, and rescale so we can assume the minimal vector length is 1 (i.e., the packing uses balls of radius $1/2$). The density is the volume of a sphere of radius $1/2$, which is $\pi^{n/2}/(2^n(n/2)!)$, times the number of spheres occurring per unit volume in space. The latter factor is $1/\text{vol}(\mathbb{R}^n/\Lambda)$, because there is one sphere for each fundamental cell of the lattice, and hence the density equals

$$\frac{\pi^{n/2}}{2^n(n/2)!} \cdot \frac{1}{\text{vol}(\mathbb{R}^n/\Lambda)}.$$

Now we apply Poisson summation to see that

$$\sum_{x \in \Lambda} f(x) = \frac{1}{\text{vol}(\mathbb{R}^n/\Lambda)} \sum_{t \in \Lambda^*} \hat{f}(t).$$

The left side is bounded above by $f(0)$, because all the other terms come from $|x| \geq 1$ and are thus nonpositive by assumption. The right side is bounded below by $\hat{f}(0)/\text{vol}(\mathbb{R}^n/\Lambda)$, because all the other terms are nonnegative. Thus,

$$f(0) \geq \frac{\hat{f}(0)}{\text{vol}(\mathbb{R}^n/\Lambda)},$$

which is equivalent to the density bound in Theorem 5.1.

The proof in the general case is completely analogous. It suffices to prove the bound for all periodic packings (because they come arbitrarily close to the maximal density), and one can apply a version of Poisson summation for summing over translates of a lattice. See [16] for the details, as well as for an explanation of the analogy between these linear programming bounds and those for compact spaces.

5.2. Apparent optimality of E_8 and the Leech lattice. Theorem 5.1 does not explain how to choose the function f , and for $n > 1$ the optimal choice of f is unknown. However, one can use numerical methods to optimize the density bound, for example by choosing $f(x)$ to be $e^{-\pi|x|^2}$ times a polynomial in $|x|^2$ (so that the Fourier transform can be easily computed) and then optimizing the choice of the polynomial. For $4 \leq n \leq 36$, the results were collected in Table 3 of [16], and in each case the bound is the best one known, but they are typically nowhere near sharp. For example, when $n = 36$, the upper bound is roughly 58.2 times the best packing density known. That

was an improvement on the previous bound, which was off by a factor of 89.7, but the gap remains enormous.

However, for $n = 2, 8, \text{ or } 24$, Theorem 5.1 appears to be sharp:

Conjecture 5.2 (Cohn and Elkies [16]). *For $n = 2, 8, \text{ or } 24$, there exists a function f that proves a sharp bound in Theorem 5.1 (for the hexagonal, E_8 , or Leech lattice, respectively).*

The strongest numerical evidence comes from [18]: for $n = 24$ the bound is sharp to within a factor of $1 + 1.65 \cdot 10^{-30}$. Similar accuracy can be obtained for $n = 8$ or $n = 2$, although only 10^{-15} was reported in [18]. Of course, for $n = 2$ the sphere packing problem has already been solved, but Conjecture 5.2 is open.

This apparent sharpness is analogous to the sharpness of the linear programming bounds for the kissing number in $\mathbb{R}^2, \mathbb{R}^8, \text{ and } \mathbb{R}^{24}$. In that problem, it would have sufficed to prove any bound less than the answer plus one, because the kissing number must be an integer, but the bounds in fact turn out to be exact integers. In the case of the sphere packing problem, the analogous exactness is needed (because packing density is not quantized), and fortunately it appears to be true.

Examining the proof of Theorem 5.1 gives simple conditions for when the bound can be sharp for a lattice Λ , analogous to the conditions for Theorem 4.1: f must vanish at each nonzero point in Λ and \widehat{f} must vanish at each nonzero point in Λ^* . In fact, the same must be true for all rotations of Λ , so f and \widehat{f} must vanish at these radii (even if they are not radial functions). Unfortunately, it seems difficult to control the behavior of f and \widehat{f} simultaneously.

For the special case of lattices, however, it is possible to complete a proof.

Theorem 5.3 (Cohn and Kumar [18]). *The Leech lattice is the unique densest lattice in \mathbb{R}^{24} , up to scaling and isometries.*

The proof uses Theorem 5.1 to show that no sphere packing in \mathbb{R}^{24} can be more than slightly denser than the Leech lattice, and that every lattice as dense as the Leech lattice must be very close to it. However, the Leech lattice is a locally optimal packing among lattices, and the bounds can be made close enough to complete the proof. This approach also yields a new proof of optimality and uniqueness for E_8 (previously shown in [10] and [53]).

One noteworthy hint regarding the optimal functions f in \mathbb{R}^8 and \mathbb{R}^{24} is an observation of Cohn and Miller [20] about the Taylor series coefficients of f . It is more convenient to use the rescaled function $g(x) = f(x/r)$, where $r = \sqrt{2}$ when $n = 8$ and $r = 2$ when $n = 24$. Then $g(0) = \widehat{g}(0)$, and without loss of generality let this value be 1. Assuming g is radial, we can view g and \widehat{g} as functions of one variable and ask for their Taylor series coefficients. Only even exponents occur by radial symmetry, so the first nontrivial terms are quadratic. Cohn and Miller noticed that the quadratic coefficients appear to be rational numbers, as shown in Table 2. The quartic terms seem more subtle, and it is

Table 2. Approximate Taylor series coefficients of g and \widehat{g} about 0.

n	function	order	coefficient	conjecture
8	g	2	$-2.70000000000000000000000000000000 \dots$	$-27/10$
8	\widehat{g}	2	$-1.50000000000000000000000000000000 \dots$	$-3/2$
24	g	2	$2.6276556776556776556776556776556 \dots$	$14347/5460$
24	\widehat{g}	2	$1.3141025641025641025641025 \dots$	$205/156$
8	g	4	$4.2167501240968298210999141 \dots$?
8	\widehat{g}	4	$-1.2397969070295980026220772 \dots$?
24	g	4	$3.8619903167183007758184168 \dots$?
24	\widehat{g}	4	$0.7376727789015322303799539 \dots$?

not clear whether they are rational as well. If they are, then their denominators are probably much larger.

More generally, one can study not just the sphere packing problem, but also potential energy minimization in Euclidean space. The total energy of a periodic configuration will be infinite, because each distance occurs infinitely many times, but one can instead try to minimize the average energy per particle. Some of the densest packings minimize more general forms of energy, but others do not, and simulations lead to many intriguing structures [19].

Cohn and Kumar [17] proved linear programming bounds for energy and made a conjecture analogous to Conjecture 5.2:

Conjecture 5.4 (Cohn and Kumar [17]). *For $n = 2, 8,$ or $24,$ the linear programming bounds for potential energy minimization in \mathbb{R}^n are sharp for every completely monotonic potential function (for the hexagonal, $E_8,$ or Leech lattice, respectively).*

This universal optimality would be a dramatic strengthening of mere optimality as packings. It is not even known in the two-dimensional case.

6. Future Prospects

The most pressing question raised by this work is how to prove that the hexagonal lattice, $E_8,$ and the Leech lattice are universally optimal in Euclidean space. Linear programming bounds reduce this problem to finding certain auxiliary functions of one variable, and the optimal functions can even be computed to high precision, but so far there is no proof that they truly exist.

More generally, can we classify the universal optima in a given space? No proof is known even that the list of examples in S^3 is complete, although it very likely is. Each of the known universal optima is such a remarkable mathematical object that a classification would be highly desirable: if there are any others out there, we ought to find them.

One noteworthy case is equiangular line configurations in complex space. Do there exist n^2 unit vectors $x_1, \dots, x_{n^2} \in \mathbb{C}^n$ such that for $i \neq j$, $|\langle x_i, x_j \rangle|^2$ is independent of i and j (in which case one can show it must be $1/(n+1)$)? In other words, the complex lines through these vectors are equidistant under the Fubini-Study metric in $\mathbb{C}\mathbb{P}^{n-1}$. Zauner [56] conjectured that the answer is yes for all n , and substantial numerical evidence supports that conjecture [44], but only finitely many cases have been proved. A collection of n^2 vectors with this property gives an n^2 -point universal optimum in $\mathbb{C}\mathbb{P}^{n-1}$, by Theorem 8.2 in [17]. This case is particularly unusual, because normally the difficulty is in proving optimality for a configuration that has already been constructed, rather than constructing one that has already been proved optimal (should it exist).

These equiangular line configurations are in fact closely analogous to Hadamard matrices. They can be characterized as exactly the simplices in $\mathbb{C}\mathbb{P}^{n-1}$ that are projective 2-designs (where a simplex is simply a set of points for which all pairwise distances are equal). Similarly, Hadamard designs, which are an equivalent variant of Hadamard matrices [3], are symmetric block 2-designs that are simplices under the Hamming distance between blocks. The existence of Hadamard matrices of all orders divisible by four is a famous unsolved problem in combinatorics, and perhaps the problem of n^2 equiangular lines in \mathbb{C}^n will be equally difficult.

These two problems are finely balanced between order and disorder. Any Hadamard matrix or equiangular line configuration must have considerable structure, but in practice they frequently seem to have just enough structure to be tantalizing, without enough to guarantee a clear construction. This contrasts with many of the most symmetrical mathematical objects, which are characterized by their symmetry groups: once you know the full group and the stabilizer of a point, it is often not hard to deduce the structure of the complete object. That seems not to be possible in either of these two problems, and it stands as a challenge to find techniques that can circumvent this difficulty.

In conclusion, packing and energy minimization problems exhibit greatly varying degrees of symmetry and order in their solutions. In certain cases, the solutions are extraordinary mathematical objects such as E_8 or the Leech lattice. Sometimes this can be proved, and sometimes it comes down to simply stated yet elusive conjectures. In other cases, the solutions may contain defects or involve unexpectedly complicated structures. Numerical experiments suggest that this is the default behavior, but it is difficult to predict exactly when or how it will occur. Finally, in rare cases there appears to be order of an unusually subtle type, as in the complex equiangular line problem, and this type of order remains a mystery.

Acknowledgments

I am grateful to James Bernhard, Tom Brennan, Tzu-Yi Chen, Donald Cohn, Noam Elkies, Abhinav Kumar, Achill Schürmann, Sal Torquato, Frank Val-

lentin, Stephanie Vance, Jeechul Woo, and especially Nadia Heninger for their valuable feedback on this paper.

References

- [1] N. N. Andreev, *An extremal property of the icosahedron*, East J. Approx. **2** (1996), 459–462.
- [2] N. N. Andreev, *Location of points on a sphere with minimal energy*, Proc. Steklov Inst. Math. **219** (1997), 20–24.
- [3] E. F. Assmus, Jr. and J. D. Key, *Designs and Their Codes*, Cambridge Tracts in Mathematics **103**, Cambridge University Press, Cambridge, 1992.
- [4] C. Bachoc, *Linear programming bounds for codes in Grassmannian spaces*, IEEE Trans. Inform. Theory **52** (2006), 2111–2125.
- [5] C. Bachoc and F. Vallentin, *New upper bounds for kissing numbers from semidefinite programming*, J. Amer. Math. Soc. **21** (2008), 909–924.
- [6] C. Bachoc and F. Vallentin, *Optimality and uniqueness of the $(4, 10, 1/6)$ spherical code*, J. Combin. Theory Ser. A **116** (2009), 195–204.
- [7] K. Ball, *A lower bound for the optimal density of lattice packings*, Internat. Math. Res. Notices **1992**, 217–221.
- [8] B. Ballinger, G. Blekherman, H. Cohn, N. Giansiracusa, E. Kelly, and A. Schürmann, *Experimental study of energy-minimizing point configurations on spheres*, Experiment. Math. **18** (2009), 257–283.
- [9] E. Bannai and N. J. A. Sloane, *Uniqueness of certain spherical codes*, Canad. J. Math. **33** (1981), 437–449.
- [10] H. F. Blichfeldt, *The minimum values of positive quadratic forms in six, seven and eight variables*, Math. Z. **39** (1935), 1–15.
- [11] L. Bowen and C. Radin, *Densest packing of equal spheres in hyperbolic space*, Discrete Comput. Geom. **29** (2003), 23–39.
- [12] M. Bowick and L. Giomi, *Two-dimensional matter: order, curvature and defects*, Advances in Physics **58** (2009), 449–563.
- [13] P. J. Cameron, J. M. Goethals, and J. J. Seidel, *Strongly regular graphs having strongly regular subconstituents*, J. Algebra **55** (1978), 257–280.
- [14] H. Cohn, *New upper bounds on sphere packings II*, Geom. Topol. **6** (2002), 329–353.
- [15] H. Cohn, J. H. Conway, N. D. Elkies, and A. Kumar, *The D_4 root system is not universally optimal*, Experiment. Math. **16** (2007), 313–320.
- [16] H. Cohn and N. D. Elkies, *New upper bounds on sphere packings I*, Ann. of Math. **157** (2003), 689–714.
- [17] H. Cohn and A. Kumar, *Universally optimal distribution of points on spheres*, J. Amer. Math. Soc. **20** (2007), 99–148.
- [18] H. Cohn and A. Kumar, *Optimality and uniqueness of the Leech lattice among lattices*, Ann. of Math. **170** (2009), 1003–1050.

-
- [19] H. Cohn, A. Kumar, and A. Schürmann, *Ground states and formal duality relations in the Gaussian core model*, Phys. Rev. E **80** (2009), 061116:1–7.
- [20] H. Cohn and S. D. Miller, *Some conjectures on optimal auxiliary functions for sphere packing*, preprint, 2010.
- [21] H. Cohn and J. Woo, *Three-point bounds for energy minimization*, preprint, 2010.
- [22] J. H. Conway and N. J. A. Sloane, *What are all the best sphere packings in low dimensions?*, Discrete Comput. Geom. **13** (1995), 383–403.
- [23] J. H. Conway, R. H. Hardin, and N. J. A. Sloane, *Packing lines, planes, etc.: packings in Grassmannian spaces*, Experiment. Math. **5** (1996), 139–159.
- [24] J. H. Conway and N. J. A. Sloane, *Sphere Packings, Lattices and Groups*, third edition, Grundlehren der Mathematischen Wissenschaften **290**, Springer, New York, 1999.
- [25] P. Delsarte, *Bounds for unrestricted codes, by linear programming*, Philips Res. Rep. **27** (1972), 272–289.
- [26] P. Delsarte, J. M. Goethals, and J. J. Seidel, *Spherical codes and designs*, Geom. Dedicata **6** (1977), 363–388.
- [27] F. J. Dyson, *A Brownian-motion model for the eigenvalues of a random matrix*, J. Math. Phys. **3** (1962), 1191–1198.
- [28] L. Fejes Tóth, *Regular Figures*, Pergamon Press, Macmillan, New York, 1964.
- [29] T. C. Hales, *Cannonballs and honeycombs*, Notices Amer. Math. Soc. **47** (2000), 440–449.
- [30] T. C. Hales, *A proof of the Kepler conjecture*, Ann. of Math. **162** (2005), 1065–1185.
- [31] M. Hazewinkel, W. Hesselink, D. Siersma, and F. D. Veldkamp, *The ubiquity of Coxeter-Dynkin diagrams (an introduction to the A–D–E problem)*, Nieuw Arch. Wisk. **25** (1977), 257–307.
- [32] G. A. Kabatiansky and V. I. Levenshtein, *Bounds for packings on a sphere and in space*, Probl. Inf. Transm. **14** (1978), 1–17.
- [33] A. I. Khinchin, *Mathematical Foundations of Information Theory*, Dover Publications, Inc., New York, 1957.
- [34] F. Klein, *Lectures on the Icosahedron and the Solution of Equations of the Fifth Degree*, second edition, Dover Publications, Inc., New York, 1956.
- [35] A. V. Kolushov and V. A. Yudin, *On the Korkin-Zolotarev construction*, Discrete Math. Appl. **4** (1994), 143–146.
- [36] A. V. Kolushov and V. A. Yudin, *Extremal dispositions of points on the sphere*, Anal. Math. **23** (1997), 25–34.
- [37] M. Laurent, *A comparison of the Sherali-Adams, Lovász-Schrijver, and Lasserre relaxations for 0–1 programming*, Math. Oper. Res. **28** (2003), 470–496.
- [38] J. Leech, *Notes on sphere packings*, Canad. J. Math. **19** (1967), 251–267.
- [39] V. I. Levenshtein, *On bounds for packings in n -dimensional Euclidean space*, Soviet Math. Dokl. **20** (1979), 417–421.

-
- [40] V. I. Levenshtein, *Designs as maximum codes in polynomial metric spaces*, Acta Appl. Math. **29** (1992), 1–82.
- [41] H. Löwen, *Fun with hard spheres*, in *Statistical Physics and Spatial Statistics* (Wuppertal, 1999), 295–331, Lecture Notes in Phys. **554**, Springer, New York, 2000.
- [42] O. Musin, *The kissing number in four dimensions*, Ann. of Math. **168** (2008), 1–32.
- [43] A. M. Odlyzko and N. J. A. Sloane, *New bounds on the number of unit spheres that can touch a unit sphere in n dimensions*, J. Combin. Theory Ser. A **26** (1979), 210–214.
- [44] J. M. Renes, R. Blume-Kohout, A. J. Scott, and C. M. Caves, *Symmetric informationally complete quantum measurements*, J. Math. Phys. **45** (2004), 2171–2180.
- [45] I. J. Schoenberg, *Positive definite functions on spheres*, Duke Math. J. **9** (1942), 96–107.
- [46] A. Schrijver, *New code upper bounds from the Terwilliger algebra and semidefinite programming*, IEEE Trans. Inform. Theory **51** (2005), 2859–2866.
- [47] K. Schütte and B. L. van der Waerden, *Das Problem der dreizehn Kugeln*, Math. Ann. **125** (1953), 325–334.
- [48] R. E. Schwartz, *The 5 electron case of Thomson’s problem*, preprint, 2010, [arXiv:1001.3702](https://arxiv.org/abs/1001.3702).
- [49] J. J. Thomson, *On the structure of the atom*, Phil. Mag. **7** (1904), 237–265.
- [50] A. Thue, *Om nogle geometrisk-taltheoretiske Theoremer*, Forhandlingerne ved de Skandinaviske Naturforskere **14** (1892), 352–353.
- [51] S. Torquato and F. Stillinger, *New conjectural lower bounds on the optimal density of sphere packings*, Experiment. Math. **15** (2006), 307–331.
- [52] S. Vance, *Lattices and sphere packings in Euclidean space*, Ph.D. dissertation, University of Washington, 2009.
- [53] N. M. Vetčinkin, *Uniqueness of classes of positive quadratic forms on which values of the Hermite constant are attained for $6 \leq n \leq 8$* , Proc. Steklov Inst. Math. **152** (1982), 37–95.
- [54] D. Weaire and R. Phelan, *A counterexample to Kelvin’s conjecture on minimal surfaces*, Phil. Mag. Lett. **69** (1994), 107–110.
- [55] V. A. Yudin, *Minimum potential energy of a point system of charges*, Discrete Math. Appl. **3** (1993), 75–81.
- [56] G. Zauner, *Quantendesigns: Grundzüge einer nichtkommutativen Designtheorie*, Ph.D. dissertation, Universität Wien, 1999.

Hurwitz Numbers: On the Edge Between Combinatorics and Geometry

Sergei K. Lando*

Abstract

Hurwitz numbers were introduced by A. Hurwitz in the end of the nineteenth century. They enumerate ramified coverings of two-dimensional surfaces. They also have many other manifestations: as connection coefficients in symmetric groups, as numbers enumerating certain classes of graphs, as Gromov–Witten invariants of complex curves. Hurwitz numbers belong to a tribe of numerical sequences that penetrate the whole body of mathematics, like multinomial coefficients. They are indexed by partitions, or, more generally, by tuples of partitions, which does not allow one to overview all of them simultaneously. Instead, we usually deal with some of their specific subsequences. The Cayley numbers N^{N-1} enumerating rooted trees on N marked vertices is may be the simplest such instance. The corresponding exponential generating series has been considered by Euler and he gave it the name of Lambert function. Certain series of Hurwitz numbers can be expressed by nice explicit formulas, and the corresponding generating functions provide solutions to integrable hierarchies of mathematical physics. The paper surveys recent progress in understanding Hurwitz numbers.

Mathematics Subject Classification (2010). Primary 05A15; Secondary 14H10, 14H30, 37K10.

Keywords. Hurwitz numbers, permutations, ramified covering, Riemann surface, KP hierarchy, moduli space of curves, Gromov–Witten invariants

*State University — Higher School of Economics, Institute for System Research RAS and the Poncelet Laboratory, Independent University of Moscow, partly supported by the RFBR grants 08-01-00110, 09-01-12185, 10-01-92104, and the HSE fundamental research program.

Department of Mathematics, State University — Higher School of Economics, 7 Vavilova Moscow 117312 Russia, Independent University of Moscow, Laboratoire J.-V.Poncelet, Institute for System Research RAS. E-mail: lando@hse.ru.

1. Hurwitz Numbers

Since their introduction by A. Hurwitz in the end of the 19th century [23, 24], the numbers experienced attraction of prominent mathematicians, like H. Weyl, as well as long periods of neglect. During these periods, the efforts of A. Mednykh (see e.g., [39]) were rare attempts to improve our understanding of their nature. The situation changed dramatically in the beginning of 1990's, when the reviving of interest has been strongly supported by demands from mathematical physics, group theory, and algebraic geometry simultaneously. The present paper is devoted to a description of the progress made in the last couple of decades. This progress is a result of joint efforts of many people all over the world.

Acknowledgements. In the process of working on Hurwitz numbers, discussions (both personal and by correspondence) with B. Dubrovin, B. Eynard, C. Faber, I. Goulden, D. Jackson, A. Kokotov, D. Korotkin, J.-H. Kwak, K. Liu, A. Okounkov, N. Orantin, A. Mednykh, R. Pandharipande, S. Natanzon, S. Shadrin, V. Shramchenko, R. Vakil, H. Xu, A. Zvonkin were of great use. I am especially grateful to my coauthors T. Ekedahl, V. Goryunov, M. Kazarian, M. Shapiro, A. Vainshtein, and D. Zvonkine for sharing with me the pleasure of understanding this fascinating subject. I would also like to thank M. Kazarian and A. Zvonkin for careful proofreading and valuable suggestions.

In this section we give the definition of Hurwitz numbers and discuss some of their combinatorial aspects.

1.1. Simple and general Hurwitz numbers. Let S_N denote the symmetric group consisting of permutations of N elements $\{1, 2, \dots, N\}$. Any permutation $\sigma \in S_N$ can be represented as a product of transpositions, and there are many such representations. For a given m , we are interested in enumeration of m -tuples of transpositions η_1, \dots, η_m whose product is a given permutation σ ,

$$\sigma = \eta_m \circ \dots \circ \eta_1.$$

The following statements are clear:

- the number of such representations depends on the cyclic type of the permutation σ rather than on the permutation itself;
- there is a minimal number $m_{\min} = m_{\min}(\sigma)$ for which such a representation exists, and this minimal number is $N - c(\sigma)$, where $c(\sigma)$ is the number of cycles in σ . Indeed, the minimal number of transpositions whose product is a cycle of length l is $l - 1$;
- all values of m for which the number of representations is nonzero have the same parity, which coincides with the parity of the permutation σ .

Now we are ready to give a precise definition of a simple Hurwitz number.

Definition 1. Let μ be a partition, $\mu \vdash |\mu|$. The *simple Hurwitz number* $h_{m;\mu}^\circ$ is defined as

$$h_{m;\mu}^\circ = \frac{1}{|\mu|!} \left| \{(\eta_1, \dots, \eta_m), \eta_i \in C_2(S_{|\mu|}) \mid \eta_m \circ \dots \circ \eta_1 \in C_\mu(S_{|\mu|})\} \right|.$$

Here $C_2(S_{|\mu|})$ denotes the set of all transpositions in $S_{|\mu|}$, and $C_\mu(S_{|\mu|})$ is the set of all permutations of cyclic type $\mu \vdash |\mu|$ in $S_{|\mu|}$, so that, in particular, $C_2(S_{|\mu|}) = C_{1^{|\mu|}-2^{2^1}}(S_{|\mu|})$. The *connected simple Hurwitz number* $h_{m;\mu}$ is defined in a similar way, but we take into account only m -tuples of transpositions such that the subgroup $\langle \eta_1, \dots, \eta_m \rangle \subseteq S_{|\mu|}$ they generate acts transitively on the set $\{1, \dots, |\mu|\}$.

The terminology has a topological origin and will be explained later. Below, we denote partitions in one of the two equivalent ways: either as a sequence of decreasing parts, $\mu = (\mu_1, \mu_2, \dots)$, where $\mu_1 \geq \mu_2 \geq \dots$, with only finitely many nonzero parts, or in the multiplicative form $1^{k_1} 2^{k_2} \dots$, where k_i denotes the multiplicity of the part i in the partition, all but finitely many multiplicities being 0 (and the corresponding parts omitted in the notation).

In slightly different terms, Hurwitz numbers enumerate ordered factorizations of permutations of given cyclic type into transpositions, while connected Hurwitz numbers enumerate those factorizations that are transitive.

Hurwitz numbers are not necessarily integers. This is true even for the simplest case,

$$h_{1;2^1}^\circ = h_{1;2^1} = \frac{1}{2} \cdot 1 = \frac{1}{2}.$$

More generally, for a tuple μ_1, \dots, μ_m of partitions of N , one can consider general Hurwitz numbers enumerating representations of the identity permutation as the product of the form $\sigma_m \circ \dots \circ \sigma_1$, where each permutation σ_i has the cyclic type μ_i , $1 \leq i \leq m$. (For simple Hurwitz numbers, all the permutations but one are transpositions, and the last permutation is σ^{-1} , whose cyclic type coincides with that of σ .) The *general Hurwitz number* is defined as the number of m -tuples of permutations $\sigma_1, \dots, \sigma_m$ of given cyclic types whose product is the identity permutation, divided by $N!$. *Connected Hurwitz numbers* are defined similarly, but with the restriction that the subgroup $\langle \sigma_1, \dots, \sigma_m \rangle \subseteq S_N$ generated by the permutations σ_i must act transitively. We do not introduce notation for general Hurwitz numbers, since we are not going to use them in our survey.

It is also worth mentioning other kinds of Hurwitz numbers, like real Hurwitz numbers (see e.g., [1]) or tropical Hurwitz numbers [5], but we are not going to discuss them in detail.

1.2. Topological interpretation. Hurwitz numbers naturally arise in the enumeration problem for ramified coverings of the 2-sphere. Below, a surface

means an oriented two-dimensional manifold. A continuous mapping $\beta : E_1 \rightarrow E_2$ of two surfaces is called a *covering* if it is an orientation preserving local homeomorphism, that is, for each point $t \in E_2$ there is a disk neighborhood $U = U(t) \subset E_2$ such that its total preimage $\beta^{-1}(U) \subset E_1$ is a disjoint union of disks, and the restriction of β to each of these disks is an orientation preserving homeomorphism. If E_2 is connected, then the number of disks in the preimage of any disk neighborhood U is the same whatever is the point t , and this number (which may well be infinite) is called the *degree* of the covering.

From the point of view of topology, a smooth projective complex curve is a compact surface. Every nonconstant holomorphic mapping $\beta : E_1 \rightarrow E_2$ of two complex curves E_1, E_2 is a *ramified covering*, meaning that it becomes a covering after puncturing E_2 at finitely many points and E_1 at their preimages under β . Locally, at a neighborhood of each point in E_1 , a ramified covering looks like $z \mapsto z^k$, for an appropriate choice of complex local coordinates in the source and the target. For all but finitely many points in E_1 , the value of k is 1, and it is greater than 1 for some preimages of the punctures. It is called the *degree* of the preimage.

For any point $t \in E_2$, the sum of the degrees of all its preimages is the same, and it is called the *degree* of the ramified covering. In other words, the degrees of the preimages of any point form a partition of the degree of the covering. For a ramified covering of degree N , all partitions different from 1^N constitute the *ramification type* of the covering. We say that a ramification point in the target surface E_2 is *non-degenerate* if the corresponding partition is $1^{N-2}2^1$, that is, if there is one preimage of degree 2, and $N - 2$ preimages at which the mapping is unramified. Otherwise, the ramification point is said to be *degenerate*. Below, we shall consider finite ramified coverings of the 2-sphere S^2 by compact oriented two-dimensional surfaces.

Consider the ramified covering $z \mapsto z^k$ of the unit disk by the unit disk. As a nonzero point in the target disk goes around 0 and returns to its initial position, its k preimages experience a cyclic permutation of length k . This property allows one to associate to a ramified covering of the sphere a tuple of permutations.

Let $\beta : E \rightarrow S^2$ be a ramified covering of degree N , and let $t_1, \dots, t_m \in S^2$ be all its points of ramification. Pick a point $t \in S^2$ distinct from all t_i and connect it with the points t_i by smooth nonintersecting segments, whose cyclic order at t coincides with the numbering. Now make each segment into a narrow path γ_i around the ramification point in the positive direction. Then the path γ_i induces a permutation σ_i of the fiber $\beta^{-1}(t)$. The cyclic type of the permutation σ_i coincides with the partition given by the degrees of the preimages in $\beta^{-1}(t_i)$, and the product $\sigma_m \circ \dots \circ \sigma_1$ is the identity permutation of the fiber $\beta^{-1}(t)$, since the concatenation of the paths $\gamma_m \circ \dots \circ \gamma_1$ is contractible in the punctured sphere $S^2 \setminus \{t_1, \dots, t_m\}$.

The m -tuple of permutations of the fiber determines the covering uniquely, up to a homeomorphism of the domain. By numbering the preimages $\beta^{-1}(t)$

of the generic point from 1 to N , we can make each permutation σ_i into a permutation of the set $\{1, 2, \dots, N\}$. Since there are $N!$ possible numberings, we conclude that Hurwitz numbers enumerate ramified coverings of the 2-sphere, with prescribed ramification types. The covering surface is connected if and only if the subgroup of S_N generated by the permutations σ_i acts transitively on the fiber $\beta^{-1}(t)$, which justifies the definition of connected Hurwitz numbers.

Let $E \rightarrow S^2$ be a ramified covering. The Riemann–Hurwitz formula allows one to recover the Euler characteristic $\chi(E)$ of the covering surface E from the ramification type. We shall use this formula only for the case of simple Hurwitz numbers, where it acquires the form

$$\chi(E) = N + c(\mu) - m.$$

Here μ is a partition of $N = |\mu|$, $c(\mu)$ is the number of parts in the partition, and m is the number of transpositions. If the covering surface is connected, then its Euler characteristic is $\chi(E) = 2 - 2g$, where g is the genus of the surface. Hence the number m of points of simple ramification can be considered as a substitute for the genus of the covering surface.

1.3. Cut-and-join equation of Goulden and Jackson. Collect the simple Hurwitz numbers into two generating functions:

$$H^\circ(u; p_1, p_2, \dots) = \sum_{m=1}^{\infty} \sum_{\mu} h_{m;\mu}^\circ p_{\mu_1} p_{\mu_2} \dots \frac{u^m}{m!}; \quad (1)$$

$$H(u; p_1, p_2, \dots) = \sum_{m=1}^{\infty} \sum_{\mu} h_{m;\mu} p_{\mu_1} p_{\mu_2} \dots \frac{u^m}{m!}, \quad (2)$$

where in each case μ runs over the set of all partitions of all numbers. These generating functions depend on infinitely many variables and are formal: we do not put any convergence requirements on them.

A very general combinatorial construction relating connected and disconnected objects justifies the following relationship between these two generating functions:

$$\text{We have } H^\circ = \exp(H).$$

This assertion allows one to translate statements about simple Hurwitz numbers into statements about connected simple Hurwitz numbers and vice versa.

The following result explains many properties of the Hurwitz numbers.

Theorem 1.1 (cut-and-join equation, [14]). *The generating function H° for simple Hurwitz numbers satisfies the following partial differential equation:*

$$\frac{\partial H^\circ}{\partial u} = \frac{1}{2} \sum_{n=1}^{\infty} \sum_{i+j=n} \left((i+j)p_i p_j \frac{\partial}{\partial p_{i+j}} + ij p_{i+j} \frac{\partial^2}{\partial p_i \partial p_j} \right) H^\circ. \quad (3)$$

Before explaining why the statement is true, let us note that the cut-and-join equation provides an explicit formula for the generating function H° . Expand it in a power series in u ,

$$H^\circ(u; p_1, p_2, \dots) = \sum_{m=0}^\infty H_{(m)}^\circ(p_1, p_2, \dots) \frac{u^m}{m!}.$$

Then the cut-and-join equation can be rewritten as the recurrence

$$H_{(m+1)}^\circ = \frac{1}{2} \sum_{n=1}^\infty \sum_{i+j=n} \left((i+j)p_i p_j \frac{\partial}{\partial p_{i+j}} + ij p_{i+j} \frac{\partial^2}{\partial p_i \partial p_j} \right) H_{(m)}^\circ = A H_{(m)}^\circ.$$

Note that the differential operator A on the right is well known in mathematical physics under the name of Calogero–Moser operator. Starting with $H_{(0)}^\circ = e^{p_1}$, we immediately obtain the first few terms of the expansion:

$$H^\circ(u; p_1, p_2, \dots) = e^{p_1} \left(1 + \frac{1}{2} p_2 \frac{u}{1!} + \left(p_1^2 + \frac{1}{2} p_2^2 + p_3 \right) \frac{u^2}{2!} + \dots \right).$$

The application of the operator A to the function $H_{(m)}^\circ$ always produces finitely many nonzero terms, although the operator itself contains infinitely many of them. The reason is that the function $H_{(m)}^\circ$ has the form e^{p_1} times a polynomial in p_1, \dots, p_m , and its derivatives over each p_k with $k > m$ vanish.

Now let us explain why the cut-and-join equation is true. It describes what happens if one of the transpositions in the decomposition of a given permutation is glued with the distinguished permutation, that is, we replace the representation

$$\sigma = \eta_m \circ \eta_{m-1} \circ \dots \circ \eta_1$$

by the representation

$$\eta_m \circ \sigma = \eta_{m-1} \circ \dots \circ \eta_1$$

(here we make use of the fact that η_m^2 is the identity permutation). Decreasing of the number of transpositions on the right by one reflects the derivation with respect to u on the left of the cut-and-join equation (3), since this procedure diminishes the degree of u by 1.

Multiplication by a transposition η_m can affect the permutation σ in one of the two different ways: either η_m exchanges two elements belonging to the same cycle of σ , or the elements it exchanges belong to distinct cycles. In the first case, a cycle in σ is split into two cycles the sum of whose lengths coincides with the length of the initial one. In the second case, conversely, two cycles are glued into a single cycle of length equal to the sum of the lengths of the two. Each of the two summands on the right of the cut-and-join equation is in charge of the corresponding possibility. The coefficients reflect the number of ways to choose two elements to be transposed by η_m : for each of the $i + j$ elements in a cycle of length $i + j$ an appropriate pair can be chosen in a unique way (if we fix the cyclic order), while in two cycles, of length i and j , respectively, there are ij choices for a pair whose transposition glues them together.

1.4. Certain formulas for rational Hurwitz numbers. Hurwitz numbers are said to be *rational* if the number of transpositions in the decomposition is the minimal possible one. The terminology comes from the fact that these numbers enumerate ramified coverings of the sphere by the sphere, that is, rational functions. Thus, rational Hurwitz numbers are, in a sense, the simplest species of Hurwitz numbers, and there are a number of explicit formulas for them.

The first such formula is the one due to Hurwitz (1891), for rational connected simple Hurwitz numbers.

Theorem 1.2 ([23]). *We have*

$$h_{|\mu|+n-2;\mu} = \frac{(|\mu| + n - 2)!}{|\text{Aut}(\mu)|} \prod_{i=1}^n \frac{\mu_i^{\mu_i}}{\mu_i!} |\mu|^{n-3},$$

where $\mu = (\mu_1, \dots, \mu_n)$ is a partition of $|\mu| = \mu_1 + \dots + \mu_n$, and $|\text{Aut}(\mu)|$ is the order of the automorphism group of the partition (for $\mu = 1^{k_1} \dots N^{k_N}$, we have $|\text{Aut}(\mu)| = k_1! \dots k_N!$).

Here $|\mu| + n - 2$ is the minimal number of transpositions (generating a permutation group acting transitively) in a product that can produce a permutation of cyclic type μ . In fact, Hurwitz did not publish the proof of his formula stating that it is too long for a journal paper. The formula was rediscovered in [14], after the problem has been revived in quantum chromodynamics models [7, 22]. A reconstruction of Hurwitz's presumable proof is given in [50]. The ELSV formula, see below, provides an alternative geometric proof [10].

Another instance of formulas for rational Hurwitz numbers is the following

Theorem 1.3 ([14]). *The number of factorizations of a cyclic permutation in S_N into a product of permutations of cyclic types ν_1, \dots, ν_m , $\nu_i \vdash N$, is*

$$N^{m-1} \frac{(c(\nu_1) - 1)!}{|\text{Aut}(\nu_1)|} \dots \frac{(c(\nu_m) - 1)!}{|\text{Aut}(\nu_m)|},$$

where $c(\nu)$ denotes the number of parts in a partition ν .

The proof in [14] is purely combinatorial. Once again, the geometric proof was given in [34].

The formula due to Bousquet-Mélou and Schaeffer enumerates decompositions of a given permutation into a product of a given number of permutations, whatever are their types. It reads as follows.

Theorem 1.4 ([4]). *Denote by $G_\mu(m)$ the number of m -tuples of permutations whose product is a permutation of cyclic type μ , divided by $N!$, $\mu \vdash N$. We have*

$$G_\mu(m) = m \frac{((m-1)N - 1)!}{((m-1)N - c(\mu) + 2)!} \prod_i \binom{m\mu_i - 1}{\mu_i} \mu_i,$$

where $c(\mu)$ is the number of parts in μ .

The original proof got a simplification in [19]. Similarly to the previous two formulas, this one also must have a geometric proof, which is still lacking.

2. Integrable Hierarchies for Hurwitz Numbers

The Kadomtsev–Petviashvili (below, KP, for brevity) hierarchy is a completely integrable system of partial differential equations playing an important role in mathematical physics. The main goal of the present section is to discuss the following statement.

Theorem 2.1. *The generating function $H(u; p_1, p_2, \dots)$ for connected simple Hurwitz numbers is a 1-parameter family of solutions to the KP hierarchy.*

In this form the theorem was first stated in [27], but it is implicitly contained in Okounkov’s paper [41]. In fact, Okounkov proves a slightly more complicated theorem stating that the generating function for *double* Hurwitz numbers (those, enumerating ramified coverings of the sphere with two points of degenerate ramification) produces a solution to the Toda lattice integrable hierarchy, which was previously conjectured by R. Pandharipande [46].

The theorem above has numerous applications, both on combinatorial and geometric side. In particular, it produces nontrivial recurrence relations on Hurwitz numbers, which mix numbers of different genera.

A general theory of KP equations, due to Sato, interprets solutions to these equations as semi-infinite planes, that is, points in the semi-infinite Grassmannian. We present a brief overlook of Sato’s construction. Proving that a given function is a solution, is thus reduced to identification of the semi-infinite plane corresponding to this function. There is no need, in particular, to know the explicit form of the equations. We make such an identification for the function $H(u; p_1, p_2, \dots)$ from a purely combinatorial point of view, without references to their geometric nature.

2.1. Grassmannian embeddings and Plücker equations.

Consider the Grassmannian $G(2, 4)$ of vector 2-planes in the 4-space $V \equiv \mathbb{C}^4$. Any 2-plane in V can be represented by the wedge product $\beta_1 \wedge \beta_2$ of any pair β_1, β_2 of linearly independent vectors in the plane. This wedge product is well defined up to a constant factor; it determines the 2-plane uniquely and thus defines an embedding of $G(2, 4)$ into the projectivization of the wedge square of V , $G(2, 4) \hookrightarrow P\Lambda^2 V$. An immediate generalization of this construction produces an embedding of any Grassmannian $G(k, n)$ of k -planes in n -space V into the projectivization $P\Lambda^k V$.

The *Plücker equations* are the equations of the image of this embedding. Note that the dimension of $G(k, n)$ is $k(n - k)$, while the dimension of $P\Lambda^k V$ is $\binom{n}{k} - 1$, whence, generally speaking, the image of the embedding does not coincide with the whole projectivized wedge product $P\Lambda^k V$. For example, the image

of the embedding of $G(2, 4)$ into $P\Lambda^2V$ is a hypersurface in the 5-dimensional projective space.

Let us find the equation of this hypersurface. Pick a basis e_1, e_2, e_3, e_4 in V . Then Λ^2V is endowed with the natural basis $\beta_{ij} = e_i \wedge e_j$, $1 \leq i < j \leq 4$, and the corresponding natural coordinate system y_{ij} . The image of the embedding of the Grassmannian consists of decomposable vectors. By definition of the wedge product, for a pair of vectors (a_1, a_2, a_3, a_4) , (b_1, b_2, b_3, b_4) , the image of the plane spanned by these two vectors has the projective coordinates

$$y_{ij} = \begin{vmatrix} a_i & b_i \\ a_j & b_j \end{vmatrix} = a_i b_j - a_j b_i.$$

An immediate calculation shows that these coordinates satisfy the homogeneous equation

$$y_{12}y_{34} - y_{13}y_{24} + y_{14}y_{23} = 0,$$

and this is the Plücker equation of the image.

For general values of n and k , the Plücker equations still are quadratic equations. In other words, *the ideal in the ring of polynomials consisting of polynomials vanishing on the image of the Plücker embedding is generated by quadratic polynomials.*

2.2. Space of Laurent series. Take for the space V the infinite dimensional vector space of formal Laurent series in one variable. Elements of this space have the form $c_{-k}z^{-k} + c_{-k+1}z^{-k+1} + \dots$. The powers z^k , $k = \dots, -2, -1, 0, 1, 2, \dots$ form the *standard* basis in V . By definition, the *semi-infinite wedge product* $\Lambda^{\infty} V$ is the vector space freely spanned by the vectors

$$v_{\mu} = z^{m_1} \wedge z^{m_2} \wedge z^{m_3} \wedge \dots, \quad m_1 < m_2 < m_3 < \dots, \quad m_i = \mu_i - i,$$

where μ is a partition, $\mu = (\mu_1, \mu_2, \mu_3, \dots)$, $\mu_1 \geq \mu_2 \geq \mu_3 \geq \dots$, and all but finitely many parts are 0. In particular, $m_i = -i$ for all i large enough.

The *vacuum vector*

$$v_{\emptyset} = z^{-1} \wedge z^{-2} \wedge z^{-3} \wedge \dots$$

corresponds to the empty partition. Similarly, we have

$$v_{11} = z^0 \wedge z^{-2} \wedge z^{-3} \wedge \dots, \quad v_{21} = z^1 \wedge z^{-2} \wedge z^{-3} \wedge \dots, \quad v_{12} = z^0 \wedge z^{-1} \wedge z^{-3} \wedge \dots,$$

and so on.

2.3. The boson-fermion correspondence. Numbering basis vectors in the semi-infinite wedge product $\Lambda^{\infty} V$ (the space of *fermions*) by partitions establishes a natural vector space isomorphism (the *boson-fermion correspondence*) between this space and the vector space of power series in infinitely

many variables p_1, p_2, \dots (the space of *bosons*). This isomorphism takes a basis vector v_μ to the Schur polynomial $s_\mu = s_\mu(p_1, p_2, \dots)$. The latter is a quasihomogeneous polynomial, of degree $|\mu|$, in the variables p_i , with the degree of p_i set to be i .

The *Schur polynomial* corresponding to a one-part partition is defined by the expansion

$$s_0 + s_1z + s_2z^2 + s_3z^3 + s_4z^4 + \dots = e^{p_1z + p_2\frac{z^2}{2} + p_3\frac{z^3}{3} + \dots},$$

and for a general partition κ it is given by the determinant

$$s_\kappa = \det \|s_{\kappa_j - j + i}\|. \tag{4}$$

The indices i, j here run over the set $\{1, 2, \dots, n\}$ for n large enough, and since $\kappa_i = 0$ for i sufficiently large, the determinant, hence s_κ , is independent of n . Here are a few first Schur polynomials:

$$\begin{aligned} s_0 &= 1, & s_{1^1} &= p_1, & s_{2^1} &= \frac{1}{2}(p_1^2 + p_2), & s_{3^1} &= \frac{1}{6}(p_1^3 + 3p_1p_2 + 2p_3), \\ s_{1^2} &= \frac{1}{2}(p_1^2 - p_2), & s_{1^1 2^1} &= \frac{1}{3}(p_1^3 - p_3), & s_{1^3} &= \frac{1}{6}(p_1^3 - 3p_1p_2 + 2p_3). \end{aligned}$$

2.4. Semi-infinite Grassmannian and the KP equations.

The *semi-infinite Grassmannian* $G(\frac{\infty}{2}, \infty)$ consists of decomposable vectors in $PA^{\frac{\infty}{2}}V$, that is, of vectors of the form

$$\beta_1(z) \wedge \beta_2(z) \wedge \beta_3(z) \wedge \dots,$$

where each β_i is a Laurent power series in z and, for i large enough, the leading term in the expansion of β_i is z^{-i} :

$$\beta_i(z) = z^{-i} + c_{i1}z^{-i+1} + c_{i2}z^{-i+2} + \dots$$

Definition 2. The *Hirota equations* are the Plücker equations of the embedding of the semi-infinite Grassmannian in the projectivized semi-infinite wedge product $PA^{\frac{\infty}{2}}V$. Solutions to the Hirota equations (that is, semi-infinite planes) are called τ -*functions* for the KP hierarchy.

As polynomial equations for the coefficients of the expansions of τ -functions, the Hirota equations can be treated as partial differential equations for the functions themselves. Being Plücker equations, the Hirota equations are quadratic in τ .

Definition 3. The form the Hirota equations take for the logarithms of τ -functions under the boson-fermion correspondence is called the *Kadomtsev-Petviashvili*, or *KP*, equations.

In other words, any solution to the KP equations can be obtained as the result of the following procedure:

- take a semi-infinite plane $\beta_1(z) \wedge \beta_2(z) \wedge \dots$ in V ;
- by expanding, rewrite the corresponding point in the semi-infinite Grassmannian as a linear combination of the basis vectors v_κ and normalize so as the coefficient of v_\emptyset becomes 1;
- replace in this linear combination each vector v_κ by the corresponding Schur polynomial $s_\kappa(p_1, p_2, \dots)$, which produces a series in infinitely many variables p_1, p_2, \dots ;
- take the logarithm of the resulting series.

An infinite sequence of homogeneous generators can be chosen for the KP equations, involving derivatives over extending sets of variables. For example, the first KP equation for an unknown function $W = W(p_1, p_2, \dots)$ looks like

$$\frac{\partial^2 W}{\partial p_2^2} = \frac{\partial^2 W}{\partial p_1 \partial p_3} - \frac{1}{2} \left(\frac{\partial^2 W}{\partial p_1^2} \right)^2 - \frac{1}{12} \frac{\partial^4 W}{\partial p_1^4}$$

(it contains derivatives only over p_1, p_2, p_3 , and is homogeneous, in a natural sense).

2.5. Action of the diagonal matrices. Linear transformations of the vector space V of Laurent polynomials induce linear transformations of the semi-infinite wedge product $\Lambda^{\infty} V$. Since linear transformations of V take planes in V to planes, the induced transformations preserve the embedded Grassmannian. In this section we consider the action of those transformations that can be represented by diagonal matrices in the basis $\{z^k\}$ in V , $k \in \mathbb{Z}$: these are the only transformations we need in the study of simple Hurwitz numbers. By obvious reasons, the induced action on $\Lambda^{\infty} V$, written in the basis v_κ , also is diagonal.

Example 2.2. Consider the linear transformation $V \rightarrow V$ which multiplies z^{-1} by a constant a preserving all the other basis vectors. Clearly, the action of this transformation on $\Lambda^{\infty} V$, written in the basis v_κ , multiplies by a each basis vector containing z^{-1} in its decomposition (v_\emptyset, v_{1^2} , and so on), and preserves all other basis vectors (v_{1^1}, v_{2^1} , and so on). The requirement that z^{-1} enters the decomposition of a vector v_κ means that the partition κ contains a part κ_i such that $\kappa_i - i = -1$. Note that any partition can have at most one such part, since the parts κ_i follow in a decreasing order, while the sequence i grows strictly.

An important consequence of this example is that the eigenvalue of the action on $\Lambda^{\infty} V$ of a diagonal matrix on V corresponding to the eigenvector v_κ depends symmetrically on the differences $\kappa_i - i$. In other words, it belongs to the ring of so-called shifted symmetric functions.

Definition 4. A function on partitions $\kappa = (\kappa_1, \kappa_2, \dots)$ is said to be *shifted symmetric* if it is symmetric under permutations of the shifted parts $\kappa_i - i$.

Let us stress once again that the parts $\kappa_1, \kappa_2, \dots$ of the partition κ go in the nonincreasing order, $\kappa_1 \geq \kappa_2 \geq \dots$, and all but finitely many of them are 0. The definition of a shifted symmetric function bases heavily on this order.

The space of shifted-symmetric functions depending on infinitely many variables is the projective limit Γ of the spaces Γ_k of shifted symmetric functions depending on k variables. (In [42], the algebra Γ is denoted by Λ^* . We use a different notation in order to prevent confusion with the wedge products and the Hodge bundle below). The limit is taken with respect to the projections $\Gamma_{k+1} \rightarrow \Gamma_k$ obtained by setting the last argument equal to 0. All complex-valued shifted symmetric functions form an algebra. This algebra was introduced and thoroughly studied in [30]. The reason for introducing it is that the characters of certain natural elements in the centers of group algebras of symmetric groups are shifted symmetric.

Now, we have a naturally defined action on $\Lambda^{\frac{\infty}{2}} V$ of any diagonal matrix $z^k \mapsto a_k z^k$, $a_k \neq 0$, with *finitely many entries a_k with negative indices different from 1*. Indeed, were there infinitely many such coefficients, in order to compute the action of the corresponding matrix on a basis vector, say v_θ , we would have to compute the product of infinitely many entries. Fortunately, the action on the *projectivized* space $P\Lambda^{\frac{\infty}{2}} V$, which is the main subject of our interest, can be extended to the action of diagonal matrices with infinitely many entries a_k with negative indices different from 1: since we are interested in the action on the projectivized space, only the ratio of the eigenvalues of the basis eigenvectors matters, and this ratio is well defined for an arbitrary diagonal matrix.

Indeed, any two basis vectors $v_\kappa, v_\mu \in \Lambda^{\frac{\infty}{2}} V$ have a common tail: their decompositions are different in the beginning, but coincide after some position, say K . Hence the ratio of the corresponding eigenvalues is just $\frac{a_{\kappa_1-1} \dots a_{\kappa_K-K}}{a_{\mu_1-1} \dots a_{\mu_K-K}}$. That is, we must define the action of a diagonal matrix on $\Lambda^{\frac{\infty}{2}} V$ in a way that preserves this ratio of eigenvalues. Thus the result depends only on the eigenvalue of the vacuum vector v_θ , which can be chosen arbitrarily. The most natural normalization is to choose this eigenvalue to be 1. This yields the following induced action on $\Lambda^{\frac{\infty}{2}} V$ of a diagonal matrix (a_k) on V :

$$v_\kappa \mapsto \left(\prod_{i=1}^{\infty} \frac{a_{\kappa_i-i}}{a_{-i}} \right) v_\kappa.$$

The product in the brackets is well defined, since all but finitely many factors are 1. The action of the torus of diagonal matrices on the projectivized seminfinite external product of V is just the inductive limit of the actions of the tori T_K consisting of diagonal matrices with diagonal entries a_i equal to 1 for $i = -(K + 1), -(K + 2), \dots$.

Since the action of the infinite dimensional torus $\bigoplus_{i \in \mathbb{Z}} (\mathbb{C}^*)_i$ on the projectivized semi-infinite wedge product is well defined, it also defines an action of

the corresponding Lie algebra. The latter action also is diagonal, and a diagonal matrix $(\alpha_i)_{i \in \mathbb{Z}}$ (with not necessarily nonzero entries) belonging to the Lie algebra acts on a basis vector v_κ by

$$v_\kappa \mapsto \left(\sum_{j=1}^{\infty} (\alpha_{\kappa_j - j} - \alpha_{-j}) \right) v_\kappa.$$

2.6. Symmetric group representations. In this section, we prove Theorem 2.1 stating that the generating series $H(u; p_1, p_2, \dots)$ for simple Hurwitz numbers is a solution to the KP hierarchy for each value of the parameter u . This statement is true for $u = 0$, since $H(0; p_1, p_2, \dots) = p_1$. For general value of u , the statement follows from the fact that $\exp(H)$ is an integral curve of a vector field in $P\Lambda^{\infty} V$ tangent to the semi-infinite Grassmannian. This vector field is induced by a linear transformation $V \rightarrow V$, which is diagonal in the standard basis z^k . Namely, this is the transformation $z^k \mapsto (k - \frac{1}{2})^2 z^k$.

Let $\mathbb{C}[S_N]$ be the $N!$ -dimensional group algebra of the symmetric group. For each partition κ of N , denote by $C_\kappa \in \mathbb{C}[S_N]$ the sum of all permutations in S_N having the cyclic type κ . We will use a special notation C_1 for the class C_{1^N} of the unit permutation, which is the unit of the algebra $\mathbb{C}[S_N]$, and C_2 for the sum $C_{1^{N-2}2^1}$ of all transpositions. For any κ , the element C_κ is a central element in $\mathbb{C}[S_N]$. These elements span the center of $\mathbb{C}[S_N]$.

The simple Hurwitz numbers have the following natural interpretation as connection coefficients in symmetric groups. Take the m th power C_2^m of the class $C_2 \in \mathbb{C}[S_N]$ and expand it as a linear combination of the basis classes. Then the coefficient of C_μ in this expansion is equal to the number of ways to represent a given permutation of cyclic type μ as a product of m transpositions. In other words,

$$C_2^m = N! \sum_{\mu \vdash N} h_{m;\mu}^\circ \frac{C_\mu}{|C_\mu|},$$

where $|C_\mu|$ is the number of elements in the corresponding conjugacy class.

Example 2.3. For $N = 3$ and $m = 4$, we have

$$C_2^4 = 27C_1 + 27C_3,$$

whence

$$h_{4;3^1}^\circ = h_{4;3^1} = \frac{2 \cdot 27}{6} = 9.$$

(Let us explain how the coefficient 27 of the class C_1 in the above formula is obtained. Each of the 27 products of three transpositions in S_3 is a transposition. Taking for the fourth transposition one of the two transpositions different from the product we obtain 54 cyclic permutations, that is, the element C_3 , which is the sum of the two cyclic permutations, taken with multiplicity 27).

It is convenient to interpret the above relation by assigning the monomial $|C_\mu|p_\mu = |C_\mu|p_{\mu_1}p_{\mu_2}\dots$ to the element C_μ . This correspondence provides an isomorphism between the center of $\mathbb{C}[S_N]$ and the vector space of weighted homogeneous polynomials of degree N in the variables p_1, p_2, \dots . Under this isomorphism, we have

$$C_2^m = N! \sum_{\mu \vdash N} h_{m;\mu}^\circ p_\mu.$$

Therefore,

$$e^{C_2 u} = N! \sum_{m=0}^\infty \sum_{\mu \vdash N} h_{m;\mu}^\circ p_\mu \frac{u^m}{m!}.$$

In order to compute the action of the element C_2 and that of its exponent, we observe that an element of $\mathbb{C}[S_N]$ is central iff it acts as a scalar on any irreducible representation. In particular, the central elements $\chi_\mu \in \mathbb{C}[S_N]$ which act with the trace 1 in the irreducible representation V_μ and trivially in all other representations form yet another basis in the center of $\mathbb{C}[S_N]$. The elements C_2 and $e^{C_2 u}$, being central, act diagonally in this basis:

$$C_2 : \chi_\mu \mapsto f_2(\mu)\chi_\mu, \quad e^{C_2 u} : \chi_\mu \mapsto e^{f_2(\mu)u}\chi_\mu,$$

with f_2 given by

$$f_2(\mu) = \frac{1}{2} \sum_{i=1}^\infty \left(\left(\mu_i + \frac{1}{2} - i \right)^2 - \left(\frac{1}{2} - i \right)^2 \right).$$

Under the isomorphism above, the element χ_μ is taken exactly to the corresponding Schur function by (yet another) its definition. The equivalence of the two definitions of the Schur function is a standard fact known as the Frobenius theorem; the proof can be found, for example, in [47]. Expanding the function $H^\circ(0; p_1, \dots) = e^{p_1}$ in the basis of Schur polynomials,

$$e^{p_1} = \sum_{\mu} s_{\mu}(1, 0, 0, \dots) s_{\mu}(p),$$

we obtain finally

$$H^\circ(u; p_1, p_2, \dots) = \sum_{\mu} s_{\mu}(1, 0, 0, \dots) s_{\mu}(p) e^{f_2(\mu)u}.$$

This explicit formula for simple Hurwitz numbers goes back to Burnside. Similarly to formulas in Sec. 1.3 it also can be used for computation of particular simple Hurwitz numbers. Note that the above isomorphism between the center of $\mathbb{C}[S_N]$ and the space of degree N polynomials in the variables p_i takes the multiplication by C_2 to the cut-and-join (or Calogero–Moser) operator A of Sec. 1.3. We conclude that the cut-and-join operator is diagonal in the basis of Schur polynomials. The specific form of the eigenvalue function f_2 shows that this diagonal operator is induced by the diagonal operator $z^k \mapsto (k - 1/2)^2 z^k$ on the space V of Laurent polynomials. This proves Theorem 2.1.

2.7. Application: enumeration of maps and hypermaps. Informally, a map is a graph drawn on a two-dimensional surface in such a way that its edges do not intersect and self-intersect and its complement is a disjoint union of discs (faces). Maps are studied by topological graph theory, see e.g. [34]. Enumeration of maps of various kinds is a classical problem, nowadays finding numerous applications in quantum field theory. In this section we explain how the study of Hurwitz numbers helps to make enumeration results for maps more precise.

From the point of view of the present paper, the most convenient definition of a map is that in terms of permutation groups.

Definition 5. Pick a finite set D . Then a *map with the set of half-edges D on an oriented surface* is a triple of permutations α, φ, σ of D possessing the following properties:

- α is an involution without fixed points;
- the product $\varphi\alpha\sigma$ is the identity permutation.

The group $G = \langle \alpha, \varphi, \sigma \rangle$ of permutations of D generated by the permutations α, φ, σ is called the *cartographic group* of the map. A map is said to be *connected* if its cartographic group acts on the set D transitively.

For a graph drawn on an oriented surface, D is the set of half-edges, or flags, the permutation α exchanges the ends of each edge, φ rotates the half-edges along the faces in the positive direction, and σ rotates the half-edges around the vertices in the positive direction. Obviously, α is an involution without fixed points, and it is easy to check that the product of these three permutations is indeed the identity permutation. A map is connected iff the underlying surface is.

The number of edges in a map is half the number of elements in D or, which is the same, the number of cycles in the permutation α . The number of vertices in a map is the number of cycles in σ , and the degrees of the vertices are the lengths of the cycles. Similarly, the number of faces is the number of cycles in φ , and the degrees of the faces are the lengths of the cycles.

The notion of *hypermap* is a generalization of that of map. In the definition of a hypermap, we get rid of the assumption that α is an involution without fixed points, thus reestablishing the symmetry between the three permutations.

It is clear now that enumeration of maps or hypermaps of various kinds can be reduced to enumeration of triples of permutations possessing certain specific properties, and enumerative methods described above can be applied.

Denote by $R_{\kappa}^{(n,m)}$ the number of rooted connected maps with n edges, m faces, and the degrees of the vertices given by the partition κ of $2n$.

Methods close to those in the proof of Theorem 2.1 give the following statement.

Theorem 2.4 ([16]). *The generating series*

$$R(w, z; p_1, p_2, \dots) = \sum_{n, m \geq 1} \sum_{\kappa \vdash 2n} \frac{R_\kappa^{(n, m)}}{2n} p_\kappa w^m z^n,$$

(where, for a given partition $\kappa = (\kappa_1, \kappa_2, \kappa_3, \dots)$, p_κ denotes the monomial $p_\kappa = p_{\kappa_1} p_{\kappa_2} p_{\kappa_3} \dots$) is a 2-parameter family of solutions to the KP-hierarchy.

The series R in the theorem can be specialized to include only cubic maps — those whose all vertex degrees are 3. By duality, this is the same as enumerating rooted triangulations of arbitrary genus. The KP equations then can be reduced to produce recurrence relations for the number of rooted triangulations.

Denote the number of rooted triangulations of a genus g surface with $2n$ faces by $T(n, g)$. Then the recurrence relation has the following form. Introduce notation

$$S = \left\{ (n, g) \in \mathbb{Z} \times \mathbb{Z} \mid n \geq -1, \quad 0 \leq g \leq \frac{n+1}{2} \right\}.$$

Theorem 2.5 ([16]). *We have*

$$T(n, g) = \frac{1}{3n+2} t(n, g),$$

where $t(n, g)$ is defined by the quadratic recurrence

$$t(n, g) = \frac{4(3n+2)}{n+1} \left(n(3n-2)t(n-2, g-1) + \sum t(i, h)t(j, k) \right),$$

for $(n, g) \in S \setminus \{(-1, 0), (0, 0)\}$, where the summation is carried over $(i, h) \in S$, $(j, k) \in S$ with $i+j = n-2$ and $h+k = g$, subject to the initial conditions

$$t(-1, 0) = \frac{1}{2}, \quad t(n, g) = 0 \text{ for } (n, g) \notin S.$$

The recurrence relation of the theorem allowed Bender, Gao and Richmond to solve a long-standing problem of finding the exact formula for the constant factor in the leading term in the asymptotics of the number of rooted triangulations, as the number of triangles tends to infinity.

Theorem 2.6 ([3]). *The number of rooted triangulations of a genus g surface with $2n$ faces has the asymptotics*

$$T(n, g) \sim 3 \times 6^{(g-1)/2} t_g n^{5(g-1)/2} (12\sqrt{3})^n \text{ as } n \rightarrow \infty;$$

here the constant t_g has the form

$$t_g = 8 \frac{[1/5]_g [4/5]_{g-1}}{\Gamma(\frac{5g-1}{2})} \left(\frac{25}{96} \right) u_g,$$

where $[x]_k$ denotes the rising factorial $x(x+1)\dots(x+k-1)$, and the constant u_g is defined by the initial condition $u_1 = 1/10$ and the quadratic recurrence relation

$$u_g = u_{g-1} + \sum_{h=1}^{g-1} \frac{1}{R_1(g,h)R_2(g,h)} u_h u_{g-h} \text{ for } g \geq 2,$$

where

$$R_1(g,h) = \frac{[1/5]_g}{[1/5]_h} [1/5]_{g-h}, \quad R_2(g,h) = \frac{[4/5]_{g-1}}{[4/5]_{h-1}} [4/5]_{g-h-1}.$$

The first few values of the constant t_g are

$$t_0 = \frac{2}{\sqrt{\pi}}, \quad t_1 = \frac{1}{24}, \quad t_2 = \frac{7}{4320\sqrt{\pi}}.$$

This constant enters many other asymptotics as well.

3. Intersection Theory on Moduli Spaces of Complex Curves

The importance of Hurwitz numbers in modern research is mainly due to their connections with the geometry of the moduli space of curves. These connections go back to the work of A. Hurwitz in the end of the 19th century, and found numerous remarkable instances in the last decade.

3.1. The ELSV formula. Let $\overline{\mathcal{M}}_{g;n}$ denote the moduli space of stable genus g complex curves with n pairwise distinct marked points. This is the Deligne–Mumford compactification [8] of the moduli space $\mathcal{M}_{g;n}$ of stable nonsingular genus g curves with n marked points. The *stability condition* means that the group of automorphisms of the curve preserving the marked points is finite. For smooth curves, this is equivalent to the following numerical restrictions: either $g \geq 2$, or $g = 1, n \geq 1$, or $g = 0, n \geq 3$. The only singularities of the singular curves are transversal double self-intersections (nodes), and the marked points are not allowed to coincide with the nodes. Both $\mathcal{M}_{g;n}$ and $\overline{\mathcal{M}}_{g;n}$ are smooth complex orbifolds of dimension $3g - 3 + n$.

The natural “forgetting morphism” $\mathcal{M}_{g;n+1} \rightarrow \mathcal{M}_{g;n}$ extends to a forgetting morphism of the compactifications, $\overline{\mathcal{M}}_{g;n+1} \rightarrow \overline{\mathcal{M}}_{g;n}$. The composition of forgetting morphisms forgets more than one marked point.

To the i th marked point, the line bundle \mathcal{L}_i over $\overline{\mathcal{M}}_{g;n}$ is associated; the fiber of this bundle is the cotangent line to the curve at the point. Let ψ_i denote the first Chern class of \mathcal{L}_i , $\psi_i = c_1(\mathcal{L}_i) \in H^2(\overline{\mathcal{M}}_{g;n})$, $i = 1, \dots, n$. The *Hodge bundle* Λ over $\overline{\mathcal{M}}_{g;n}$ is the pull-back, under the forgetting morphism, of the rank g vector bundle over $\overline{\mathcal{M}}_{g;0}$ whose fiber is the vector space of holomorphic 1-forms over the curve. (For $g = 1$, the space $\overline{\mathcal{M}}_{g;0}$ must be replaced by $\overline{\mathcal{M}}_{g;1}$,

and for $g = 0$, the Hodge bundle is of rank zero). The characteristic classes of the Hodge bundle are denoted by $c(\Lambda) = 1 + \lambda_1 + \dots + \lambda_g$, $\lambda_i \in H^{2i}(\mathcal{M}_{g;n})$.

A formula due to Ekedahl, Lando, Shapiro, and Vainshtein, now standardly referred to as the *ELSV-formula*, expresses simple Hurwitz numbers in terms of intersection indices of the above characteristic classes over the moduli spaces of stable curves:

$$h_{m;\kappa} = \frac{m!}{|\text{Aut}(\kappa)|} \prod_{i=1}^n \frac{\kappa_i^{\kappa_i}}{\kappa_i!} \int_{\overline{\mathcal{M}}_{g;n}} \frac{c(\Lambda^\vee)}{(1 - \kappa_1 \psi_1) \dots (1 - \kappa_n \psi_n)}, \tag{5}$$

where κ is a partition of $K = |\kappa|$, $\kappa = (\kappa_1, \dots, \kappa_n)$, $m = 2g - 2 + K + n$ is the number of transpositions, and $c(\Lambda^\vee) = 1 - \lambda_1 + \lambda_2 - \dots \pm \lambda_g$ is the total Chern class of the dual Hodge bundle. This formula, together with a brief description of the idea of the proof, has been announced in [9] (with an erroneous sign in the numerator of the integrand). A complete proof was given in [10], and meanwhile another proof appeared in [21]. A special case of (5), that for $\kappa = 1^n$, has been simultaneously and independently discovered in [13].

The formula is understood in the following way: after expanding the denominator as a power series in the classes ψ_i , select the monomials of degree $\dim \overline{\mathcal{M}}_{g;n} = 3g - 3 + n$ in the product and integrate them against the fundamental class of $\overline{\mathcal{M}}_{g;n}$. The result will be a rational number.

The ELSV formula generalizes, to higher genera, Hurwitz’s formula (see Theorem 1.2) valid for $g = 0$. In its own turn, it admits a generalization known as the *Mariño–Vafa formula* conjectured in [38] and proved in [37].

In spite of the geometric nature of the ELSV formula, it produces immediate combinatorial consequences. An example is given by the following result, which has been conjectured in [17].

Theorem 3.1 ([10]). *For given g, n , the number*

$$h_{m;\kappa} \frac{|\text{Aut}(\kappa)|}{m!} \prod_{i=1}^n \frac{\kappa_i!}{\kappa_i^{\kappa_i}}$$

is a symmetric polynomial in κ_i , of degree $3g - 3 + n$, with the least monomial degree being $2g - 3 + n$.

Although the statement is purely combinatorial, no direct proof of it is known. Double Hurwitz numbers demonstrate a similar behavior. Namely, they are piecewise polynomial [18, 49].

3.2. Linear Hodge integrals as coefficients of a solution to KP. The right-hand side of the ELSV formula is a linear combination of the intersection numbers of the form

$$\ell_{j;m_1, \dots, m_n} = \int_{\overline{\mathcal{M}}_{g;n}} \lambda_j \psi_1^{m_1} \dots \psi_n^{m_n}.$$

Expressions of this kind are called *linear Hodge integrals*, meaning that they include the Chern classes λ_j of the Hodge bundle, which enter the monomial linearly. Note that the data $(j; m_1, \dots, m_n)$ determine the genus g uniquely according to the dimension count $3g - 3 + n = j + m_1 + \dots + m_n$. Similarly to the case of Hurwitz numbers, one can organize the linear Hodge integrals in the generating function

$$L(u; q_1, q_2, \dots) = \sum_{j, \mu} (-1)^j \ell_{j; m_1, \dots, m_n} u^{2j} q_{m_1} \dots q_{m_n}, \tag{6}$$

known as the *enriched Gromov–Witten potential* of a point [17].

In a recent paper, M. Kazarian has shown that this generating function can be easily transformed into a solution of the KP hierarchy. Namely, denote by $G(u; p_1, p_2, \dots)$ the result of the following substitution to the series L :

$$\begin{aligned} q_0 &= p_1, \\ q_1 &= u^2 p_1 + 2u p_2 + p_3, \\ q_2 &= u^4 p_1 + 6u^3 p_2 + 12u^2 p_3 + 10u p_4 + 3p_5, \\ q_3 &= u^6 p_1 + 14u^5 p_2 + 61u^4 p_3 + 124u^3 p_4 + 131u^2 p_5 + 70u p_6 + 15p_7, \\ &\dots \end{aligned}$$

Here the polynomials on the right-hand side are given by the recurrence

$$q_{k+1} = \sum_{m \geq 1} m(u^2 p_m + 2u p_{m+1} + p_{m+2}) \frac{\partial}{\partial p_m} q_k.$$

Theorem 3.2 ([26]). *The function $G(u; p_1, p_2, \dots)$ is a solution to the KP hierarchy (identically in u).*

The proof of the theorem uses the ELSV formula (5) and the fact that the generating series for the simple Hurwitz numbers is a solution to KP (Theorem 2.1). Note that in the present case, the infinitesimal transformation of the space V of Laurent series corresponding to the solution in question is no longer diagonal. Instead, it is three-diagonal.

3.3. Witten’s conjecture. The celebrated Witten conjecture [51] concerns computation of the intersection indices of the ψ -classes over the moduli spaces of curves. Namely, denote

$$\langle \tau_{m_1} \dots \tau_{m_n} \rangle = \int_{\overline{\mathcal{M}}_{g;n}} \psi_1^{m_1} \dots \psi_n^{m_n},$$

where the genus g can be computed from the dimensional count $\dim \overline{\mathcal{M}}_{g;n} = 3g - 3 + n = m_1 + \dots + m_n$. Collect these intersection indices into the generating

series in infinitely many variables t_i ,

$$\begin{aligned}
 F(t_0, \dots) &= \sum \frac{\langle \tau_{m_1} \dots \tau_{m_n} \rangle}{n!} t_{m_1} \dots t_{m_n} \\
 &= \frac{1}{24} t_1 + \frac{1}{6} t_0^3 + \frac{1}{48} t_1^2 + \frac{1}{24} t_0 t_2 + \frac{1}{6} t_0^3 t_1 + \frac{1}{1152} t_4 + \frac{1}{72} t_1^3 + \frac{1}{12} t_0 t_1 t_2 \\
 &\quad + \frac{1}{48} t_0^2 t_3 + \frac{1}{6} t_0^3 t_1^2 + \frac{1}{24} t_0^4 t_2 + \frac{29}{5760} t_2 t_3 + \frac{1}{384} t_1 t_4 + \frac{1}{1152} t_0 t_5 + \dots
 \end{aligned}$$

Witten’s conjecture states that

The function F satisfies the KdV hierarchy of partial differential equations. In particular, its second derivative $U = \partial^2 F / \partial t_0$ is a solution to the KdV equation,

$$\frac{\partial U}{\partial t_1} = U \frac{\partial U}{\partial t_0} + \frac{1}{12} \frac{\partial^3 U}{\partial t_0^3}. \tag{7}$$

The KdV equation (7) can be considered as a recurrence relation allowing one to compute the intersection indices of the ψ -classes for arbitrary genus recursively from their values for $g = 0$ and $g = 1$, which are known since Witten’s pioneering work [51].

Since its appearance in 1991, the conjecture has got several proofs, including those due to Kontsevich [32], Okounkov and Pandharipande [43], Mirzakhani [40], Kazarian and Lando [27], Kim and Liu [31].

Witten’s conjecture is an immediate consequence [26] of Theorem 3.2. Indeed, the solutions of the KdV hierarchy are exactly those solutions of KP that depend only on variables with odd indices. After setting $u = 0$ in G , one obtains a power series in variables p_{2i-1} with odd indices, which is therefore a solution to the KdV hierarchy. The coefficients of this series are $\ell_{0; m_1, \dots, m_n} = \langle \tau_{m_1} \dots \tau_{m_n} \rangle$. It turns into F after rescaling $p_{2i+1} = t_i / (2i - 1)!!$. In contrast to most of the other proofs, this one guarantees the whole KdV hierarchy for F , while usually one obtains only the first KdV equation and needs the additional string equation to generate the hierarchy.

4. Further Developments and Perspectives

The variety of Hurwitz numbers is not exhausted by simple and double Hurwitz numbers. Other species include general Hurwitz numbers, enumerating factorizations into permutations of arbitrary cyclic type, not necessarily transpositions, and r -Hurwitz numbers, where transpositions are replaced by certain “completed r -cycles”. In all cases, Hurwitz numbers remain closely related to the geometry of moduli spaces, and both are far from being well understood. In this section we describe briefly possible directions of further research.

4.1. Completed cycles. The center of the group algebra $\mathbb{C}[S_N]$ of the symmetric group S_N is generated by the classes $C_\kappa(S_N)$, where κ is a partition

of N . The class $C_\kappa(S_N)$ is the sum of all permutations with the cyclic type κ . For example, $C_{1^{N-2}2^1}(S_N)$ is the sum of all transpositions in S_N .

It is convenient, however, to introduce certain classes in the centers of group algebras for all symmetric groups simultaneously. Let κ be a partition. For an arbitrary integer N , choose $|\kappa|$ elements out of $\{1, \dots, N\}$ and consider in $\mathbb{C}[S_N]$ the sum of all permutations of these $|\kappa|$ elements, of cyclic type κ , all the other $N - |\kappa|$ elements being fixed. Denote by \tilde{C}_κ the element in the center of $\mathbb{C}[S_N]$ which is the sum of all such permutations, for all $\binom{N}{|\kappa|}$ choices of the $|\kappa|$ elements out of N . (If $|\kappa| > N$, then $\tilde{C}_\kappa = 0 \in \mathbb{C}[S_N]$; if $|\kappa| = N$, then $\tilde{C}_\kappa = C_\kappa(S_N)$).

For example, the class \tilde{C}_{1^1} can be understood as the sum of identity permutations, with a distinguished element in each permutation. In other words, the class \tilde{C}_{1^1} is the same as the class $N\tilde{C}_\emptyset = NC_{1^N}(S_N)$. Similarly, the class \tilde{C}_{1^2} coincides with the class $\frac{N(N-1)}{2}\tilde{C}_\emptyset$: there are $\binom{N}{2} = \frac{N(N-1)}{2}$ ways to pick two elements in the identity permutation.

The classes \tilde{C}_κ have the following advantage when compared to the classes $C_\kappa(S_N)$: the products of the classes \tilde{C}_κ can be expressed as universal linear combinations of these classes, *which are independent of the order N of the symmetric group*. For example, the equation

$$\tilde{C}_{2^1}\tilde{C}_{1^2} = \tilde{C}_{2^1} + 2\tilde{C}_{1^1 2^1} + \tilde{C}_{1^2 2^1}$$

is valid in the center of the group algebra $\mathbb{C}[S_N]$ of any symmetric group S_N , for arbitrary N .

Universality means that there is a natural inclusion of the center of $\mathbb{C}[S_N]$ into that of $\mathbb{C}[S_{N+1}]$ for any N . Tending N to infinity, we obtain a universal center of the group algebra, which can be identified with the infinite dimensional vector space freely spanned by the elements \tilde{C}_κ , for arbitrary partitions κ . This space also is endowed with an algebra structure.

This algebra is isomorphic to the algebra Γ of shifted symmetric functions defined in Sec. 2.5. As a vector space, the latter algebra is spanned by the functions f_κ indexed by partitions and defined as follows. A central element $\tilde{C}_\kappa \in \mathbb{C}[S_{|\mu|}]$ acts on the irreducible representation V_μ of the symmetric group by multiplication by a scalar; by definition, we set $f_\kappa(\mu)$ to be equal to this scalar. The *Frobenius characteristic mapping* $\tilde{C}_\kappa \mapsto f_\kappa$ establishes an isomorphism between the two algebras.

4.2. r -Hurwitz numbers and generalized Witten's conjecture. Simple Hurwitz numbers count decompositions of a given permutation into a product of transpositions. It is a natural idea to generalize them by replacing transpositions by permutations in other specific classes. For example, why not consider 3-cycles \tilde{C}_3 ? However, such a straightforward approach fails. Namely, enumerative formulas for decompositions of a given permutation into a product of 3-cycles lose elegance, when compared to that for Hurwitz numbers, and their relationship with both mathematical physics and geometry is

broken. The same is true for r -cycles for any $r \geq 3$. Fortunately, consistency can be restored by replacing r -cycles \tilde{C}_r with certain linear combinations of the classes \tilde{C}_κ , for certain partitions κ .

Definition 6 ([44]). The *completed r -cycle* \bar{C}_r is the preimage under the Frobenius characteristic mapping of the r th power function

$$(\mu_1, \mu_2, \dots) \mapsto \frac{1}{r} \sum_{i=1}^{\infty} \left(\left(\mu_i - i + \frac{1}{2} \right)^r - \left(\frac{1}{2} - i \right)^r \right).$$

We have explained the reasons why the r th power function must be of such a form in Sec. 2.5 (we use a normalization differing from that in [44] by a constant).

Let us give formulas for few first completed cycles among which we know that the completed 2-cycle simply coincides with the ordinary 2-cycle:

$$\begin{aligned} \bar{C}_1 &= \tilde{C}_{1^1} \\ \bar{C}_2 &= \tilde{C}_{2^1} \\ \bar{C}_3 &= \tilde{C}_{3^1} + \tilde{C}_{1^2} + \frac{1}{12} \tilde{C}_{1^1} \\ \bar{C}_4 &= \tilde{C}_{4^1} + 2\tilde{C}_{1^1 2^1} + \frac{5}{4} \tilde{C}_{2^1}. \end{aligned}$$

These formulas explain the origin of the term “completed cycle”: the expansion of a class \bar{C}_r as a linear combination of the classes \tilde{C}_κ starts with the class of the r -cycle \tilde{C}_{r^1} , and then terms of smaller order follow. Explicit formulas for the coefficients on the right of the expressions for all completed cycles can be found in [44].

Now we can define the generalized Hurwitz numbers.

Definition 7. The *simple r -Hurwitz number* for an integer $m \vdash N$ and a partition μ is the normalized coefficient of \tilde{C}_μ in the m th power of the completed r -cycle,

$$h_{m;\mu}^{(r)\circ} = \frac{|C_\mu|}{N!} [\tilde{C}_\mu] (\bar{C}_r)^m.$$

The simple r -Hurwitz numbers are collected into the generating function

$$H^{(r)\circ}(u; p_1, p_2, \dots) = \sum_{m=0}^{\infty} \sum_{\mu} h_{m;\mu}^{(r)\circ} p_{\mu_1} p_{\mu_2} \dots \frac{u^m}{m!},$$

and its logarithm $H^{(r)}(u; p_1, p_2, \dots) = \log H^{(r)\circ}(u; p_1, p_2, \dots)$ is the generating function for *connected simple r -Hurwitz numbers*.

The definition of the r -Hurwitz numbers and explanation in Sec. 2.5 immediately imply

Theorem 4.1. *The function $H^{(r)}(u; p_1, p_2, \dots)$ is a one-parameter family of solutions to the KP hierarchy of partial differential equations.*

Indeed, this one-parameter family is induced by the infinitesimal diagonal transformation of the vector space V of Laurent polynomials taking the vector z^k to $\frac{1}{r}(k - \frac{1}{2})^r z^k$, $k = \dots, -2, -1, 0, 1, 2, \dots$.

A similar theorem is valid for generating functions defined by any finite linear combination of completed cycles. In this case the eigenvalues $\frac{1}{r}(k - \frac{1}{2})^r$ are replaced by an appropriate polynomial in k , which can be arbitrary.

The relationship of r -Hurwitz numbers defined by means of the completed cycles to the geometry of moduli spaces of $(r - 1)$ -spin structures on algebraic curves is less clear at the moment, and this question is a subject of further investigation.

D. Zvonkine conjectured (private communication) that the simple r -Hurwitz numbers can be expressed in terms of the geometry of moduli spaces of $(r - 1)$ -spin structures on algebraic curves by an r -analogue of the ELSV-formula. Such a formula could lead, at least in principle, to an alternative proof of the generalized Witten conjecture [51], concerning intersection indices of ψ -classes on the moduli spaces of so-called r -spin curves. At the moment, only one proof of the conjecture is known, see [12], and it proceeds in a very different way.

4.3. Geometry of Hurwitz spaces and universal characteristic classes. The Hurwitz numbers are related to the geometry of moduli spaces of curves through the geometry of *Hurwitz spaces*. The latter are moduli spaces of meromorphic functions on complex curves. Without giving precise definitions, we just explain the main features of the picture. Each Hurwitz space is fibered over the corresponding moduli space of curves — the fibration proceeds by forgetting the function, and this forgetting mapping relates the geometry of the two spaces in question. In a sense, Hurwitz spaces (and, more generally, spaces of stable mappings) are more natural than moduli spaces of curves.

Each Hurwitz space is also stratified according to the degeneration of the functions. A stratum is formed by the locus of functions with prescribed singularities. On the other hand, the action of the multiplicative group \mathbb{C}^* of nonzero complex numbers on the target curve $\mathbb{C}P^1$ is lifted to the Hurwitz spaces. A Hurwitz number (either simple or a more general one) can be computed as the degree of the closure of the corresponding stratum with respect to the above action. This argument votes for the study of the stratification of Hurwitz spaces.

In the simplest case of polynomials, such a study has been carried out in [35]. In [2, 36, 28, 29], a more general case of rational functions is treated. The study applies methods of global singularity theory started by R. Thom and extended recently by M. Kazarian to the case of multisingularities (see, e.g. [32]). These methods produce universal expressions for the locus of prescribed singularities of an arbitrary generic mapping of two complex manifolds in terms of the characteristic classes of the mapping. When applied to the Hurwitz spaces, these methods yield expressions for the loci of functions with prescribed singularities, which lead to explicit formulas for the corresponding Hurwitz numbers.

The classical Thom approach, as well as its generalization by Kazarian, is applicable to the case of mappings with isolated singularities only. For spaces of stable mappings, this requirement proves to be too restrictive, since they inevitably contain mappings with nonisolated singularities, namely, those contracting certain irreducible components of the curve to a single point in the target space. Sample computations show, however, that main results can be extended to the nonisolated case as well. The corresponding construction is not elaborated yet in the desired generality.

References

- [1] A.V. Alexeevski, S.M. Natanzon, *Hurwitz numbers for regular coverings of surfaces by seamed surfaces and Cardy-Frobenius algebras of finite groups*, Geometry, topology, and mathematical physics, 1–25, Amer. Math. Soc. Transl. Ser. 2, 224, Amer. Math. Soc., Providence, RI, 2008.
- [2] V.I. Arnold, *Topological classification of complex trigonometric polynomials and the combinatorics of graphs with an identical number of vertices and edges*, (Russian) Funktsional. Anal. i Prilozhen. 30 (1996), no. 1, 1–17, 96; translation in Funct. Anal. Appl. 30 (1996), no. 1, 1–14
- [3] E.A. Bender, Z. Gao, L.B. Richmond, *The map asymptotics constant t_g* , Electron. J. Combin. 15 (2008), no. 1, Research paper 51, 8 pp.
- [4] M. Bousquet-Mélou, G. Schaeffer, *Enumeration of planar constellations*, Adv. in Apl. Math., **24**, 337–368 (2000)
- [5] R. Cavalieri, P. Johnson, H. Markwig, *Tropical Hurwitz numbers*, arXiv:0804.0579
- [6] L. Chen, Y. Li, K. Liu, *Localization, Hurwitz numbers and the Witten conjecture*. Asian J. Math. 12 (2008), no. 4, 511–518
- [7] M. Crescimanno, W. Taylor, *Large N phases of chiral QCD_2* , Nuclear Phys. B, vol. **437**, 3–24 (1995)
- [8] P. Deligne, D. Mumford, *The irreducibility of the space of curves of given genus*, Inst. Hautes Études Sci. Publ. Math. No. 36, 75–109 (1969)
- [9] T. Ekedahl, S.K. Lando, M. Shapiro, A. Vainshtein, *On Hurwitz numbers and Hodge integrals*, C. R. Acad. Sci. Paris Sér I Math., **328**, 1175–1180 (1999)
- [10] T. Ekedahl, S.K. Lando, M. Shapiro, A. Vainshtein, *Hurwitz numbers and intersections on moduli spaces of curves*, Invent. math., **146**, 297–327 (2001)
- [11] L. Euler, *De serie Lambertina plurimisque eius insignibus proprietatibus*, Acta academiae scientiarum Petropolitanae 1779: II, 1783, 29–51
- [12] C. Faber, S. Shadrin, D. Zvonkine, *Tautological relations and the r -spin Witten conjecture*, math.AG/0612510
- [13] B. Fantechi, R. Pandharipande, *Stable maps and branch divisors*, Compositio Math. 130, no. 3, 345–364 (2002).

-
- [14] I.P. Goulden, D.M. Jackson, *Transitive factorisation into transpositions and holomorphic mappings on the sphere*, Proc. Amer. Math. Soc., **125**, no. 1, 51–60 (1997)
- [15] I.P. Goulden, D.M. Jackson, *A proof of a conjecture for the number of ramified coverings of the sphere by the torus*, J. Comb. Theory, Ser. A, **88**, 246–258 (1999)
- [16] I.P. Goulden, D.M. Jackson, *The KP hierarchy, branched covers, and triangulations*, Adv. Math. 219, no. 3, 932–951 (2008).
- [17] I.P. Goulden, D.M. Jackson, R. Vakil, *The Gromov–Witten potential of a point, Hurwitz numbers, and Hodge integrals*, Proc. London Math. Soc. (3), **83**, 563–581 (2001)
- [18] I.P. Goulden, D.M. Jackson, R. Vakil, *Towards the geometry of double Hurwitz numbers*, Adv. Math. 198 (2005), no. 1, 43–92
- [19] I.P. Goulden, L.G. Serrano, *A simple recurrence for covers of the sphere with branch points of arbitrary ramification*, Annals of Combinatorics, **10**, 431–441 (2006)
- [20] A. Goupil, G. Schaeffer, *Factoring N -Cycles and Counting Maps of Given Genus*, Europ. J. Combinatorics, **19**, 819–834 (1998)
- [21] T. Graber, R. Vakil, *Hodge integrals and Hurwitz numbers via virtual localization*, Compositio Math. **135**, no. 1, 25–36 (2003)
- [22] D.J. Gross, W. Taylor, *Two-dimensional QCD is a string theory*, Nuclear Physics B, vol. **403**, 395–452 (1993)
- [23] A. Hurwitz, *Über Riemann’sche Flächen mit gegebenen Verzweigungspunkten*, Math. Ann., **39**, 1–61 (1891)
- [24] A. Hurwitz, *Über die Anzahl der Riemann’sche Flächen mit gegebenen Verzweigungspunkten*, Math. Ann., **55**, 51–60 (1902)
- [25] M. Kazarian, *Thom polynomials for Lagrange, Legendre, and critical point function singularities*. Proc. London Math. Soc. (3) 86 (2003), no. 3, 707–734.
- [26] M. Kazarian, *KP hierarchy for Hodge integrals*. Adv. Math. 221, no. 1, 1–21 (2009).
- [27] M. Kazarian, S. Lando, *An algebro-geometric proof of Witten’s conjecture*, J. Amer. Math. Soc., **20**, 1079–1089 (2007)
- [28] M. Kazarian, S. Lando, *On intersection theory on Hurwitz spaces*, Izv. Ross. Akad. Nauk Ser. Mat., **68**, no. 5, 91–122 (2004); translation in Izv. Math. **68**, no. 5, 935–964 (2004)
- [29] M. Kazarian, S. Lando, *Thom polynomials for mappings of curves with isolated singularities*, (Russian) Tr. Mat. Inst. Steklova 258 (2007), Anal. i Osob. Ch. 1, 93–106; translation in Proc. Steklov Inst. Math. 258 (2007), no. 1, 87–99
- [30] S. Kerov, G. Olshanski, *Polynomial functions on the set of Young diagrams*, C. R. Acad. Sci. Paris Sér I Math., **319**, no. 2, 121–126 (1994)
- [31] Y.-S. Kim, K. Liu, *A simple proof of Witten conjecture through localization*, preprint: math.AG/0508384.
- [32] M. Kontsevich, *Intersection theory on the moduli space of curves and the matrix Airy function*, Comm. Math. Phys., **147**, 1–23 (1992)

- [33] S.K. Lando, *Ramified coverings of the two-dimensional sphere and intersection theory in spaces of meromorphic functions on algebraic curves*, Russ. Math. Surv., **57**, no. 3, 463–533 (2002).
- [34] S.K. Lando, A.K. Zvonkin, *Graphs on surfaces and their applications*, Springer (2004)
- [35] S.K. Lando, D. Zvonkine, *On multiplicities of the Lyashko-Looijenga mapping on strata of the discriminant*, Funktsional. Anal. i Prilozhen., **33**, no. 3, 21–34 (1999); translation in Funct. Anal. Appl., **33**, no. 3, 178–188 (1999).
- [36] S.K. Lando, D. Zvonkine, *Counting Ramified Coverings and Intersection Theory on Spaces of Rational Functions I (Cohomology of Hurwitz Spaces)*, Moscow Math. J., **7** (1), 85–107 (2007)
- [37] C.-C. M. Liu, K. Liu, J. Zhou, *A proof of a conjecture of Mariño-Vafa on Hodge integrals*. J. Differential Geom. **65** (2003), no. 2, 289–340.
- [38] M. Mariño, C. Vafa, *Framed knots at large N* , in: Orbifolds in mathematics and physics (Madison, WI, 2001), 185–204, Contemp. Math., **310**, Amer. Math. Soc., Providence, RI (2002)
- [39] A.D. Mednykh, *Nonequivalent coverings of Riemann surfaces with a prescribed ramification type*, Siber. Math. J., **25**, 606–625 (1984)
- [40] M. Mirzakhani, *Weil–Petersson volumes and intersection theory on the moduli space of curves*, J. Amer. Math. Soc., **20**, no. 1, 1–23 (2007)
- [41] A. Okounkov, *Toda equations for Hurwitz numbers*, Math. Res. Lett. **7**, no. 4, 447–453 (2000).
- [42] A. Okounkov, G. Olshanski, *Shifted Schur functions*, Algebra i Analiz, **9**, no 2, 73–146 (1997); translation in St.Petersburg Math. J., **9**, no. 2, 239–300 (1998)
- [43] A. Okounkov, R. Pandharipande, *Gromov–Witten theory, Hurwitz numbers, and matrix models*. Algebraic geometry—Seattle 2005. Part 1, 325–414, Proc. Sympos. Pure Math., **80**, Part 1, Amer. Math. Soc., Providence, RI, 2009.
- [44] A. Okounkov, R. Pandharipande, *Gromov–Witten theory, Hurwitz theory, and completed cycles*, Ann. of Math. (2), **163**, no. 2, 517–560 (2006).
- [45] A. Okounkov, R. Pandharipande, *Equivariant Gromov–Witten theory of P^1* . Ann. of Math. (2) **163** (2006), no. 2, 561–605.
- [46] R. Pandharipande, *The Toda equations and the Gromov–Witten theory of the Riemann sphere*, Lett. Math. Phys. **53**, no. 1, 59–74 (2000).
- [47] B.E. Sagan, *The Symmetric Group*, Springer, 2001.
- [48] M. Sato, Y. Sato, *Soliton equations as dynamical systems on infinite dimensional Grassmann manifolds*, in: Nonlinear partial differential equations in applied science, North-Holland, Amsterdam, 259–271 (1983)
- [49] S. Shadrin, M. Shapiro, A. Vainshtein, *Chamber behavior of double Hurwitz numbers in genus 0*, Adv. Math. **217** (2008), no. 1, 79–96.
- [50] V. Strehl, *Minimal transitive products of transpositions — the reconstruction of a proof of A. Hurwitz*. Sémin. Lothar. Combin. **37** (1996), Art. S37c, 12 pp. (electronic).

- [51] E. Witten, *Two-dimensional gravity and intersection theory on moduli spaces*, Surveys in Differential Geometry, vol. 1, 243–269 (1991)
- [52] D. Zvonkine, *Counting ramified coverings and intersection theory on Hurwitz spaces. II. Local structure of Hurwitz spaces and combinatorial results*. Mosc. Math. J. 7 (2007), no. 1, 135–162.

Cluster Algebras and Representation Theory

Bernard Leclerc*

Abstract

We apply the new theory of cluster algebras of Fomin and Zelevinsky to study some combinatorial problems arising in Lie theory. This is joint work with Geiss and Schröer (§3, 4, 5, 6), and with Hernandez (§8, 9).

Mathematics Subject Classification (2010). Primary 05E10; Secondary 13F60, 16G20, 17B10, 17B37.

Keywords. Cluster algebra, canonical and semicanonical basis, preprojective algebra, quantum affine algebra.

1. Introduction: Two Problems in Lie Theory

Let \mathfrak{g} be a simple complex Lie algebra of type A , D , or E . We denote by G a simply-connected complex algebraic group with Lie algebra \mathfrak{g} , by N a maximal unipotent subgroup of G , by \mathfrak{n} its Lie algebra. In [47], Lusztig has introduced the semicanonical basis \mathcal{S} of the enveloping algebra $U(\mathfrak{n})$ of \mathfrak{n} . Using the duality between $U(\mathfrak{n})$ and the coordinate ring $\mathbb{C}[N]$ of N , one obtains a new basis \mathcal{S}^* of $\mathbb{C}[N]$ which we call the *dual semicanonical basis* [22]. This basis has remarkable properties. For example there is a natural way of realizing every irreducible finite-dimensional representation of \mathfrak{g} as a subspace $L(\lambda)$ of $\mathbb{C}[N]$, and \mathcal{S}^* is compatible with this infinite system of subspaces, that is, $\mathcal{S}^* \cap L(\lambda)$ is a basis of $L(\lambda)$ for every λ .

The definition of the semicanonical basis is geometric (see below §5). A priori, to describe an element of \mathcal{S}^* one needs to compute the Euler characteristics of certain complex algebraic varieties. Here is a simple example in type A_3 . Let $V = V_1 \oplus V_2 \oplus V_3$ be a four-dimensional graded vector space with $V_1 = \mathbb{C}e_1$,

*LMNO, Université de Caen, CNRS UMR 6139, F-14032 Caen cedex, France.
E-mail: leclerc@math.unicaen.fr.

$V_2 = \mathbb{C}e_2 \oplus \mathbb{C}e_3$, and $V_3 = \mathbb{C}e_4$. There is an element φ_X of \mathcal{S}^* attached to the nilpotent endomorphism X of V given by

$$Xe_1 = e_2, \quad Xe_2 = Xe_3 = 0, \quad Xe_4 = e_3.$$

Let \mathcal{F}_X be the variety of complete flags $F_1 \subset F_2 \subset F_3$ of subspaces of V , which are graded (i.e. $F_i = \bigoplus_j (V_j \cap F_i)$ ($1 \leq i \leq 3$)) and X -stable (i.e. $XF_i \subset F_i$). The calculation of φ_X amounts to computing the Euler characteristics of the connected components of \mathcal{F}_X . In this case there are four components, two points and two projective lines, so these Euler numbers are 1, 1, 2, 2. Unfortunately, such a direct geometric computation looks rather hopeless in general.

Problem 1.1. *Find a combinatorial algorithm for calculating \mathcal{S}^* .*

To formulate the second problem we need more notation. Let $L\mathfrak{g} = \mathfrak{g} \otimes \mathbb{C}[t, t^{-1}]$ be the loop algebra of \mathfrak{g} , and let $U_q(L\mathfrak{g})$ denote the quantum analogue of its enveloping algebra, introduced by Drinfeld and Jimbo. Here we assume that $q \in \mathbb{C}^*$ is not a root of unity. The finite-dimensional irreducible representations of $U_q(L\mathfrak{g})$ are of special importance because their tensor products give rise to trigonometric R -matrices, that is, to trigonometric solutions of the quantum Yang-Baxter equation with spectral parameters [38]. The question arises whether the tensor product of two given irreducible representations is again irreducible. Equivalently, one can ask whether a given irreducible can be factored into a tensor product of representations of strictly smaller dimensions.

For instance, if $\mathfrak{g} = \mathfrak{sl}_2$ and V_n is its $(n + 1)$ -dimensional irreducible representation, the loop algebra $L\mathfrak{sl}_2$ acts on V_n by

$$(x \otimes t^k)(v) = z^k xv, \quad (x \in \mathfrak{sl}_2, k \in \mathbb{Z}, v \in V_n).$$

Here $z \in \mathbb{C}^*$ is a fixed number called the evaluation parameter. Jimbo [37] has introduced a simple $U_q(L\mathfrak{sl}_2)$ -module $W_{n,z}$, which can be seen as a q -analogue of this evaluation representation. Chari and Pressley [7] have proved that $W_{n,z} \otimes W_{m,y}$ is an irreducible $U_q(L\mathfrak{sl}_2)$ -module if and only if

$$q^{n-m} \frac{z}{y} \notin \left\{ q^{\pm(n+m+2-2k)} \mid 0 < k \leq \min(n, m) \right\}.$$

In the other direction, they showed that every simple object in the category $\text{mod } U_q(L\mathfrak{sl}_2)$ of (type 1) finite-dimensional $U_q(L\mathfrak{sl}_2)$ -modules can be written as a tensor product of modules of the form W_{n_i, z_i} for some n_i and z_i . Thus the modules $W_{n,z}$ can be regarded as the *prime* simple objects in the tensor category $\text{mod } U_q(L\mathfrak{sl}_2)$.

Similarly, for general \mathfrak{g} one would like to ask

Problem 1.2. *Find the prime simple objects of $\text{mod } U_q(L\mathfrak{g})$, and describe the prime tensor factorization of the simple objects.*

Both problems are quite hard, and we can only offer partial solutions. An interesting feature is that, in both situations, cluster algebras provide the natural combinatorial framework to work with.

2. Cluster Algebras

Cluster algebras were invented by Fomin and Zelevinsky [16] as an abstraction of certain combinatorial structures which they had previously discovered while studying total positivity in semisimple algebraic groups. A nice introduction [14] to these ideas is given in these proceedings, with many references to the growing literature on the subject.

A cluster algebra is a commutative ring with a distinguished set of generators and a particular type of relations. Although there can be infinitely many generators and relations, they are all obtained from a finite number of them by means of an inductive procedure called *mutation*.

Let us recall the definition.¹ We start with the field of rational functions $\mathcal{F} = \mathbb{Q}(x_1, \dots, x_n)$. A *seed* in \mathcal{F} is a pair $\Sigma = (\mathbf{y}, Q)$, where $\mathbf{y} = (y_1, \dots, y_n)$ is a free generating set of \mathcal{F} , and Q is a quiver (*i.e.* an oriented graph) with vertices labelled by $\{1, \dots, n\}$. We assume that Q has neither loops nor 2-cycles. For $k = 1, \dots, n$, one defines a new seed $\mu_k(\Sigma)$ as follows. First $\mu_k(y_i) = y_i$ for $i \neq k$, and

$$\mu_k(y_k) = \frac{\prod_{i \rightarrow k} y_i + \prod_{k \rightarrow j} y_j}{y_k}, \quad (1)$$

where the first (*resp.* second) product is over all arrows of Q with target (*resp.* source) k . Next $\mu_k(Q)$ is obtained from Q by

- (a) adding a new arrow $i \rightarrow j$ for every existing pair of arrows $i \rightarrow k$ and $k \rightarrow j$;
- (b) reversing the orientation of every arrow with target or source equal to k ;
- (c) erasing every pair of opposite arrows possibly created by (a).

It is easy to check that $\mu_k(\Sigma)$ is a seed, and $\mu_k(\mu_k(\Sigma)) = \Sigma$. The *mutation class* $\mathcal{C}(\Sigma)$ is the set of all seeds obtained from Σ by a finite sequence of mutations μ_k . One can think of the elements of $\mathcal{C}(\Sigma)$ as the vertices of an n -regular tree in which every edge stands for a mutation. If $\Sigma' = ((y'_1, \dots, y'_n), Q')$ is a seed in $\mathcal{C}(\Sigma)$, then the subset $\{y'_1, \dots, y'_n\}$ is called a *cluster*, and its elements are called *cluster variables*. Now, Fomin and Zelevinsky define the *cluster algebra* \mathcal{A}_Σ as the subring of \mathcal{F} generated by all cluster variables. Some important elements of \mathcal{A}_Σ are the *cluster monomials*, *i.e.* monomials in the cluster variables supported on a single cluster.

For instance, if $n = 2$ and $\Sigma = ((x_1, x_2), Q)$, where Q is the quiver with a arrows from 1 to 2, then \mathcal{A}_Σ is the subring of $\mathbb{Q}(x_1, x_2)$ generated by the rational functions x_k defined recursively by

$$x_{k+1}x_{k-1} = 1 + x_k^a, \quad (k \in \mathbb{Z}). \quad (2)$$

¹For simplicity we only consider a particular subclass of cluster algebras: the antisymmetric cluster algebras of geometric type. This is sufficient for our purpose.

The clusters of \mathcal{A}_Σ are the subsets $\{x_k, x_{k+1}\}$, and the cluster monomials are the special elements of the form

$$x_k^l x_{k+1}^m, \quad (k \in \mathbb{Z}, l, m \in \mathbb{N}).$$

It turns out that when $a = 1$, there are only five different clusters and cluster variables, namely

$$x_{5k+1} = x_1, \quad x_{5k+2} = x_2, \quad x_{5k+3} = \frac{1+x_2}{x_1}, \quad x_{5k+4} = \frac{1+x_1+x_2}{x_1 x_2}, \quad x_{5k} = \frac{1+x_1}{x_2}.$$

For $a \geq 2$ though, the sequence (x_k) is no longer periodic and \mathcal{A}_Σ has infinitely many cluster variables.

The first deep results of this theory shown by Fomin and Zelevinsky are:

Theorem 2.1 ([16],[17]). (i) *Every cluster variable of \mathcal{A}_Σ is a Laurent polynomial with coefficients in \mathbb{Z} in the cluster variables of any single fixed cluster.*

(ii) *\mathcal{A}_Σ has finitely many clusters if and only if the mutation class $\mathcal{C}(\Sigma)$ contains a seed whose quiver is an orientation of a Dynkin diagram of type A, D, E .*

One important open problem [16] is to prove that the coefficients of the Laurent polynomials in (i) are always positive. In §9 below, we give a (conjectural) representation-theoretical explanation of this positivity for a certain class of cluster algebras. More positivity results, based on combinatorial or geometric descriptions of these coefficients, have been obtained by Musiker, Schiffler and Williams [48], and by Nakajima [52].

3. The Cluster Structure of $\mathbb{C}[N]$

To attack Problem 1.1 we adopt the following strategy. We endow $\mathbb{C}[N]$ with the structure of a cluster algebra.² Then we show that all cluster monomials belong to \mathcal{S}^* , and therefore we obtain a large family of elements of \mathcal{S}^* which can be calculated by the combinatorial algorithm of mutation.

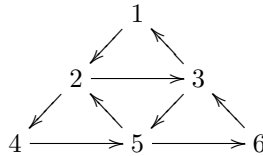
In [2, §2.6] explicit initial seeds for a cluster algebra structure in the coordinate ring of the big cell of the base affine space G/N were described. A simple modification yields initial seeds for $\mathbb{C}[N]$ (see [24]).

For instance, if $G = SL_4$ and N is the subgroup of upper unitriangular matrices, one of these seeds is

$$((D_{1,2}, D_{1,3}, D_{12,23}, D_{1,4}, D_{12,34}, D_{123,234}), Q),$$

²Here we mean that $\mathbb{C}[N] = \mathbb{C} \otimes_{\mathbb{Z}} \mathcal{A}$ for some cluster algebra \mathcal{A} contained in $\mathbb{C}[N]$.

where Q is the triangular quiver:



Here, by $D_{I,J}$ we mean the regular function on N which associates to a matrix its minor with row-set I and column-set J . Moreover, the variables

$$x_4 = D_{1,4}, \quad x_5 = D_{12,34}, \quad x_6 = D_{123,234}$$

are *frozen*, *i.e.* they cannot be mutated, and therefore they belong to every cluster. Using Theorem 2.1, it is easy to prove that this cluster algebra has finitely many clusters, namely 14 clusters and 12 cluster variables if we count the 3 frozen ones.

In general however, that is, for groups G other than SL_n with $n \leq 5$, the cluster structure of $\mathbb{C}[N]$ has infinitely many cluster variables. To relate the cluster monomials to \mathcal{S}^* we have to bring the preprojective algebra into the picture.

4. The Preprojective Algebra

Let \bar{Q} denote the quiver obtained from the Dynkin diagram of \mathfrak{g} by replacing every edge by a pair (α, α^*) of opposite arrows. Consider the element

$$\rho = \sum (\alpha\alpha^* - \alpha^*\alpha)$$

of the path algebra $\mathbb{C}\bar{Q}$ of \bar{Q} , where the sum is over all pairs of opposite arrows. Following [29, 53], we define the *preprojective algebra* Λ as the quotient of $\mathbb{C}\bar{Q}$ by the two-sided ideal generated by ρ . This is a finite-dimensional self-injective algebra, with infinitely many isomorphism classes of indecomposable modules, except if \mathfrak{g} has type A_n with $n \leq 4$. It is remarkable that these few exceptional cases coincide precisely with the cases when $\mathbb{C}[N]$ has finitely many cluster variables. Moreover, it is a nice exercise to verify that the number of indecomposable Λ -modules is then equal to the number of cluster variables.

This suggests a close relationship in general between Λ and $\mathbb{C}[N]$. To describe it we start with Lusztig’s Lagrangian construction of the enveloping algebra $U(\mathfrak{n})$ [46, 47]. This is a realization of $U(\mathfrak{n})$ as an algebra of \mathbb{C} -valued constructible functions over the varieties of representations of Λ .

To be more precise, we need to introduce more notation. Let S_i ($1 \leq i \leq n$) be the one-dimensional Λ -modules attached to the vertices i of \bar{Q} . Given a sequence $\mathbf{i} = (i_1, \dots, i_d)$ and a Λ -module X of dimension d , we introduce the variety $\mathcal{F}_{X,\mathbf{i}}$ of flags of submodules

$$\mathfrak{f} = (0 = F_0 \subset F_1 \subset \dots \subset F_d = X)$$

such that $F_k/F_{k-1} \cong S_{i_k}$ for $k = 1, \dots, d$. This is a projective variety. Denote by $\Lambda_{\mathbf{d}}$ the variety of Λ -modules X with a given dimension vector $\mathbf{d} = (d_i)$, where $\sum_i d_i = d$. Consider the constructible function $\chi_{\mathbf{i}}$ on $\Lambda_{\mathbf{d}}$ given by

$$\chi_{\mathbf{i}}(X) = \chi(\mathcal{F}_{X,\mathbf{i}})$$

where χ denotes the Euler-Poincaré characteristic. Let $\mathcal{M}_{\mathbf{d}}$ be the \mathbb{C} -vector space spanned by the functions $\chi_{\mathbf{i}}$ for all possible sequences \mathbf{i} of length d , and let

$$\mathcal{M} = \bigoplus_{\mathbf{d} \in \mathbb{N}^n} \mathcal{M}_{\mathbf{d}}.$$

Lusztig has endowed \mathcal{M} with an associative multiplication which formally resembles a convolution product, and he has shown that, if we denote by e_i the Chevalley generators of \mathfrak{n} , there is an algebra isomorphism $U(\mathfrak{n}) \xrightarrow{\sim} \mathcal{M}$ mapping the product $e_{i_1} \cdots e_{i_d}$ to $\chi_{\mathbf{i}}$ for every $\mathbf{i} = (i_1, \dots, i_d)$.

Now, following [22, 23], we dualize the picture. Every $X \in \text{mod } \Lambda$ determines a linear form δ_X on \mathcal{M} given by

$$\delta_X(f) = f(X), \quad (f \in \mathcal{M}).$$

Using the isomorphisms $\mathcal{M}^* \simeq U(\mathfrak{n})^* \simeq \mathbb{C}[N]$, the form δ_X corresponds to an element φ_X of $\mathbb{C}[N]$, and we have thus attached to every object X in $\text{mod } \Lambda$ a polynomial function φ_X on N .

For example, if \mathfrak{g} is of type A_3 , and if we denote by P_i the projective cover of S_i in $\text{mod } \Lambda$, one has

$$\varphi_{P_1} = D_{123,234}, \quad \varphi_{P_2} = D_{12,34}, \quad \varphi_{P_3} = D_{1,4}.$$

More generally, the functions φ_X corresponding to the 12 indecomposable Λ -modules are the 12 cluster variables of $\mathbb{C}[N]$.

Via the correspondence $X \mapsto \varphi_X$ the ring $\mathbb{C}[N]$ can be regarded as a kind of Hall algebra of the category $\text{mod } \Lambda$. Indeed the multiplication of $\mathbb{C}[N]$ encodes extensions in $\text{mod } \Lambda$, as shown by the following crucial result. Before stating it, we recall that $\text{mod } \Lambda$ possesses a remarkable symmetry with respect to extensions, namely, $\text{Ext}_{\Lambda}^1(X, Y)$ is isomorphic to the dual of $\text{Ext}_{\Lambda}^1(Y, X)$ functorially in X and Y (see [10, 25]). In particular $\dim \text{Ext}_{\Lambda}^1(X, Y) = \dim \text{Ext}_{\Lambda}^1(Y, X)$ for every X, Y .

Theorem 4.1 ([22, 25]). *Let $X, Y \in \text{mod } \Lambda$.*

- (i) *We have $\varphi_X \varphi_Y = \varphi_{X \oplus Y}$.*
- (ii) *Assume that $\dim \text{Ext}_{\Lambda}^1(X, Y) = 1$, and let*

$$0 \rightarrow X \rightarrow L \rightarrow Y \rightarrow 0 \quad \text{and} \quad 0 \rightarrow Y \rightarrow M \rightarrow X \rightarrow 0$$

be non-split short exact sequences. Then $\varphi_X \varphi_Y = \varphi_L + \varphi_M$.

In fact [25] contains a formula for $\varphi_X\varphi_Y$ valid for any dimension of $\text{Ext}_\Lambda^1(X, Y)$, but we will not need it here. As a simple example of (ii) in type A_2 , one can take $X = S_1$ and $Y = S_2$. Then we have the non-split short exact sequences

$$0 \rightarrow S_1 \rightarrow P_2 \rightarrow S_2 \rightarrow 0 \quad \text{and} \quad 0 \rightarrow S_2 \rightarrow P_1 \rightarrow S_1 \rightarrow 0,$$

which imply the relation $\varphi_{S_1}\varphi_{S_2} = \varphi_{P_2} + \varphi_{P_1}$, that is, the elementary determinantal relation $D_{1,2}D_{2,3} = D_{1,3} + D_{12,23}$ on the unitriangular subgroup of SL_3 . More generally, the short Plücker relations in SL_{n+1} can be obtained as instances of (ii).

We note that Theorem 4.1 is the analogue for mod Λ of a formula of Caldero and Keller [6] for the cluster categories introduced by Buan, Marsh, Reineke, Reiten and Todorov [4] to model cluster algebras with an acyclic seed. Cluster categories are not abelian, but Keller [40] has shown that they are triangulated, so in this setting exact sequences are replaced by distinguished triangles.

5. The Dual Semicanonical Basis \mathcal{S}^*

We can now introduce the basis \mathcal{S}^* of the vector space $\mathbb{C}[N]$. Let $\mathbf{d} = (d_i)$ be a dimension vector. The variety $\overline{E}_{\mathbf{d}}$ of representations of $\mathbb{C}Q$ with dimension vector \mathbf{d} is a vector space of dimension $2\sum d_i d_j$, where the sum is over all pairs $\{i, j\}$ of vertices of the Dynkin diagram which are joined by an edge. This vector space has a natural symplectic structure. Lusztig [46] has shown that $\Lambda_{\mathbf{d}}$ is a Lagrangian subvariety of $\overline{E}_{\mathbf{d}}$, and that the number of its irreducible components is equal to the dimension of the degree \mathbf{d} homogeneous component of $U(\mathfrak{n})$ (for the standard \mathbb{N}^n -grading given by the Chevalley generators). Let \mathcal{Z} be an irreducible component of $\Lambda_{\mathbf{d}}$. Since the map $\varphi : X \mapsto \varphi_X$ is a constructible map on $\Lambda_{\mathbf{d}}$, it is constant on a Zariski open subset of \mathcal{Z} . Let $\varphi_{\mathcal{Z}}$ denote this generic value of φ on \mathcal{Z} . Then, if we denote by $\mathcal{I} = \sqcup_{\mathbf{d}} \mathcal{I}_{\mathbf{d}}$ the collection of all irreducible components of all varieties $\Lambda_{\mathbf{d}}$, one can easily check that

$$\mathcal{S}^* = \{\varphi_{\mathcal{Z}} \mid \mathcal{Z} \in \mathcal{I}\}$$

is dual to the basis $\mathcal{S} = \{f_{\mathcal{Z}} \mid \mathcal{Z} \in \mathcal{I}\}$ of $\mathcal{M} \cong U(\mathfrak{n})$ constructed by Lusztig in [47], and called by him the semicanonical basis.

For example, if \mathfrak{g} is of type A_n and N is the unitriangular subgroup in SL_{n+1} , then all the matrix minors $D_{I,J}$ which do not vanish identically on N belong to \mathcal{S}^* [22]. They are of the form φ_X , where X is a subquotient of an indecomposable projective Λ -module.

More generally, suppose that X is a rigid Λ -module, *i.e.* that $\text{Ext}_\Lambda^1(X, X) = 0$. Then X is a generic point of the unique irreducible component \mathcal{Z} on which it sits, that is, $\varphi_X = \varphi_{\mathcal{Z}}$ belongs to \mathcal{S}^* , so the calculation of $\varphi_{\mathcal{Z}}$ amounts to evaluating the Euler characteristics $\chi(\mathcal{F}_{X,\mathbf{i}})$ for every \mathbf{i} (of course only finitely many

varieties $\mathcal{F}_{X,i}$ are non-empty). Thus in type A_3 , the nilpotent endomorphism X of §1 can be regarded as a rigid Λ -module with dimension vector $\mathbf{d} = (1, 2, 1)$, and the connected components of \mathcal{F}_X are just the non-trivial varieties $\mathcal{F}_{X,i}$, namely

$$\mathcal{F}_{X,(2,1,2,3)}, \quad \mathcal{F}_{X,(2,3,2,1)}, \quad \mathcal{F}_{X,(2,2,1,3)}, \quad \mathcal{F}_{X,(2,2,3,1)}.$$

Note however that if \mathfrak{g} is not of type A_n ($n \leq 4$), there exist irreducible components $\mathcal{Z} \in \mathcal{I}$ whose generic points are not rigid Λ -modules.

6. Rigid Λ -modules

Let r be the number of positive roots of \mathfrak{g} . Equivalently r is the dimension of the affine space N . This is also the number of elements of every cluster of $\mathbb{C}[N]$ (if we include the frozen variables). Geiss and Schröer have shown [28] that the number of pairwise non-isomorphic indecomposable direct summands of a rigid Λ -module is bounded above by r . A rigid module with r non-isomorphic indecomposable summands is called *maximal*. We will now see that the seeds of the cluster structure of $\mathbb{C}[N]$ come from maximal rigid Λ -modules.

Let $T = T_1 \oplus \dots \oplus T_r$ be a maximal rigid module, where every T_i is indecomposable. Define $B = \text{End}_\Lambda T$, a basic finite-dimensional algebra with simple modules s_i ($1 \leq i \leq r$). Denote by Γ_T the quiver of B , that is, the quiver with vertex set $\{1, \dots, r\}$ and d_{ij} arrows from i to j , where $d_{ij} = \dim \text{Ext}_B^1(s_i, s_j)$.

Theorem 6.1 ([23]). *The quiver Γ_T has no loops nor 2-cycles.*

Define $\Sigma(T) = ((\varphi_{T_1}, \dots, \varphi_{T_r}), \Gamma_T)$.

Theorem 6.2 ([24]). *There exists an explicit maximal rigid Λ -module U such that $\Sigma(U)$ is one of the seeds of the cluster structure of $\mathbb{C}[N]$.*

Let us now lift the notion of seed mutation to the category $\text{mod } \Lambda$.

Theorem 6.3 ([23]). *Let T_k be a non-projective indecomposable summand of T . There exists a unique indecomposable module $T_k^* \not\cong T_k$ such that $(T/T_k) \oplus T_k^*$ is maximal rigid.*

We call $(T/T_k) \oplus T_k^*$ the *mutation of T in direction k* , and denote it by $\mu_k(T)$. The proof of the next theorem relies among other things on Theorem 4.1.

Theorem 6.4 ([23]). (i) *We have $\Sigma(\mu_k(T)) = \mu_k(\Sigma(T))$, where in the right-hand side μ_k stands for the Fomin-Zelevinsky seed mutation.*

(ii) *The map $T \mapsto \Sigma(T)$ gives a one-to-one correspondence between the maximal rigid modules in the mutation class of U and the clusters of $\mathbb{C}[N]$.*

It follows immediately that the cluster monomials of $\mathbb{C}[N]$ belong to \mathcal{S}^* . Indeed, by (ii) every cluster monomial is of the form

$$\varphi_{T_1}^{a_1} \cdots \varphi_{T_r}^{a_r} = \varphi_{T_1^{a_1} \oplus \cdots \oplus T_r^{a_r}}, \quad (a_1, \dots, a_r \in \mathbb{N}),$$

for some maximal rigid module $T = T_1 \oplus \cdots \oplus T_r$, and therefore belongs to \mathcal{S}^* because $T_1^{a_1} \oplus \cdots \oplus T_r^{a_r}$ is rigid.

Thus the cluster monomials form a large subset of \mathcal{S}^* which can (in principle) be calculated algorithmically by iterating the seed mutation algorithm from an explicit initial seed. This is our partial answer to Problem 1.1.

Of course, these results also give a better understanding of the cluster structure of $\mathbb{C}[N]$. For instance they show immediately that the cluster monomials are linearly independent (a general conjecture of Fomin and Zelevinsky). Furthermore, they suggest the definition of new cluster algebra structures on the coordinate rings of unipotent radicals of parabolic subgroups of G , obtained in a similar manner from some appropriate Frobenius subcategories of $\text{mod } \Lambda$ (see [26]). One can also develop an analogous theory for finite-dimensional unipotent subgroups $N(w)$ of a Kac-Moody group attached to elements w of its Weyl group (see [3, 27]).

7. Finite-dimensional Representations of $U_q(\mathfrak{L}\mathfrak{g})$

We now turn to Problem 1.2. We need to recall some known facts about the category $\text{mod } U_q(\mathfrak{L}\mathfrak{g})^3$ of finite-dimensional modules over $U_q(\mathfrak{L}\mathfrak{g})$.

By construction, $U_q(\mathfrak{L}\mathfrak{g})$ contains a copy of $U_q(\mathfrak{g})$, so in a sense the representation theory of $U_q(\mathfrak{L}\mathfrak{g})$ is a refinement of that of $U_q(\mathfrak{g})$. Let ϖ_i ($1 \leq i \leq n$) be the fundamental weights of \mathfrak{g} , and denote by

$$P = \bigoplus_{i=1}^n \mathbb{Z}\varpi_i, \quad P_+ = \bigoplus_{i=1}^n \mathbb{N}\varpi_i,$$

the weight lattice and the monoid of dominant integral weights. It is well known that $\text{mod } U_q(\mathfrak{g})$ is a semisimple tensor category, with simple objects $L(\lambda)$ parametrized by $\lambda \in P_+$. In fact, every $M \in \text{mod } U_q(\mathfrak{g})$ has a decomposition

$$M = \bigoplus_{\mu \in P} M_\mu \tag{3}$$

into eigenspaces for a commutative subalgebra A of $U_q(\mathfrak{g})$ coming from a Cartan subalgebra of \mathfrak{g} . One shows that if M is irreducible, the highest weight occurring

³We only consider modules of type 1, a mild technical condition, see e.g. [8, §12.2 B].

in (3) is a dominant weight λ , $\dim M_\lambda = 1$, and there is a unique simple $U_q(\mathfrak{g})$ -module with these properties, hence the notation $M = L(\lambda)$. For an arbitrary $M \in \text{mod } U_q(\mathfrak{g})$, the formal sum

$$\chi(M) = \sum_{\mu \in P} \dim M_\mu e^\mu$$

is called the *character* of M , since it characterizes M up to isomorphism.

When dealing with representations of $U_q(L\mathfrak{g})$ one needs to introduce spectral parameters $z \in \mathbb{C}^*$, and therefore P and P_+ have to be replaced by

$$\widehat{P} = \bigoplus_{1 \leq i \leq n, z \in \mathbb{C}^*} \mathbb{Z}(\varpi_i, z), \quad \widehat{P}_+ = \bigoplus_{1 \leq i \leq n, z \in \mathbb{C}^*} \mathbb{N}(\varpi_i, z).$$

It was shown by Chari and Pressley [7, 9] that finite-dimensional irreducible representations of $U_q(L\mathfrak{g})$ were similarly determined by their highest l -weight $\widehat{\lambda} \in \widehat{P}_+$ (where l stands for “loop”). This comes from the existence of a large commutative subalgebra \widehat{A} of $U_q(L\mathfrak{g})$ containing A . If $M \in \text{mod } U_q(L\mathfrak{g})$ is regarded as a $U_q(\mathfrak{g})$ -module by restriction and decomposed as in (3), then every $U_q(\mathfrak{g})$ -weight-space M_μ has a finer decomposition into generalized eigenspaces for \widehat{A}

$$M_\mu = \bigoplus_{\widehat{\mu} \in \widehat{P}} M_{\widehat{\mu}}$$

where the $\widehat{\mu} = \sum_k m_{i_k}(\varpi_{i_k}, z_k)$ in the right-hand side all satisfy $\sum_k m_{i_k} \varpi_{i_k} = \mu$. The corresponding formal sum

$$\chi_q(M) = \sum_{\widehat{\mu} \in \widehat{P}} \dim M_{\widehat{\mu}} e^{\widehat{\mu}}$$

has been introduced by Frenkel and Reshetikhin [20] and called by them the q -character of M . It characterizes the class of M in the Grothendieck ring of $\text{mod } U_q(L\mathfrak{g})$, but one should be warned that this is not a semisimple category, so this is much coarser than an isomorphism class.

For instance, the 4-dimensional irreducible representation V_3 of $U_q(\mathfrak{sl}_2)$ with highest weight $\lambda = 3\varpi_1$ has character

$$\chi(V_3) = Y^3 + Y^1 + Y^{-1} + Y^{-3}$$

if we set $Y = e^{\varpi_1}$. There is a family $W_{3,z} \in \text{mod } U_q(L\mathfrak{sl}_2)$ of affine analogues of V_3 , parametrized by $z \in \mathbb{C}^*$, whose q -character is given by

$$\chi_q(W_{3,z}) = Y_z Y_{zq^2} Y_{zq^4} + Y_z Y_{zq^2} Y_{zq^6}^{-1} + Y_z Y_{zq^4}^{-1} Y_{zq^6}^{-1} + Y_{zq^2}^{-1} Y_{zq^4}^{-1} Y_{zq^6}^{-1},$$

where we write $Y_a = e^{(\varpi_1, a)}$ for $a \in \mathbb{C}^*$. Thus $W_{3,z}$ has highest l -weight

$$\widehat{\lambda} = (\varpi_1, z) + (\varpi_1, zq^2) + (\varpi_1, zq^4).$$

The reader can easily imagine what is the general expression of $\chi_q(W_{n,z})$ for any $(n, z) \in \mathbb{N} \times \mathbb{C}^*$. It follows that there is a closed formula for the q -character of every finite-dimensional irreducible $U_q(L\mathfrak{sl}_2)$ -module since, as already mentioned, every such module factorizes as a tensor product of W_{n_i, z_i} and the factors are given by a simple combinatorial rule [7].

The situation is far more complicated in general. In particular it is not always possible to endow an irreducible $U_q(\mathfrak{g})$ -module with the structure of a $U_q(L\mathfrak{g})$ -module. The only general description of q -characters of simple $U_q(L\mathfrak{g})$ -modules, due to Ginzburg and Vasserot for type A [30] and to Nakajima in general [49], uses intersection cohomology of certain moduli spaces of representations of graded preprojective algebras, called graded quiver varieties. This yields a Kazhdan-Lusztig type algorithm for calculating the irreducible q -characters [50], but this type of combinatorics does not easily reveal the possible factorizations of the q -characters.

8. The Subcategories \mathcal{C}_ℓ

It can be shown that Problem 1.2 for $\text{mod } U_q(L\mathfrak{g})$ can be reduced to the same problem for some much smaller tensor subcategories \mathcal{C}_ℓ ($\ell \in \mathbb{N}$) which we shall now introduce.

Denote by $L(\widehat{\lambda})$ the simple object of $\text{mod } U_q(L\mathfrak{g})$ with highest l -weight $\widehat{\lambda} \in \widehat{P}_+$. Since the Dynkin diagram of \mathfrak{g} is a tree, it is a bipartite graph. We denote by $I = I_0 \sqcup I_1$ the corresponding partition of the set of vertices, and we write $\xi_i = 0$ (resp. $\xi_i = 1$) if $i \in I_0$ (resp. $i \in I_1$). For $\ell \in \mathbb{N}$, let

$$\widehat{P}_{+, \ell} = \bigoplus_{1 \leq i \leq n, 0 \leq k \leq \ell} \mathbb{N}(\varpi_i, q^{\xi_i + 2k}).$$

We then define \mathcal{C}_ℓ as the full subcategory of $\text{mod } U_q(L\mathfrak{g})$ whose objects M have all their composition factors of the form $L(\widehat{\lambda})$ with $\widehat{\lambda} \in \widehat{P}_{+, \ell}$. It is not difficult to prove [33] that \mathcal{C}_ℓ is a tensor subcategory, and that its Grothendieck ring $K_0(\mathcal{C}_\ell)$ is the polynomial ring in the $n(\ell + 1)$ classes of fundamental modules

$$[L(\varpi_i, q^{\xi_i + 2k})], \quad (1 \leq i \leq n, 0 \leq k \leq \ell).$$

For example, let $W_{j,a}^{(i)}$ denote the simple object of $\text{mod } U_q(L\mathfrak{g})$ with highest l -weight

$$(\varpi_i, a) + (\varpi_i, aq^2) + \dots + (\varpi_i, aq^{2j-2}), \quad (i \in I, j \in \mathbb{N}^*, a \in \mathbb{C}^*),$$

a so-called *Kirillov-Reshetikhin module*. The q -characters of the Kirillov-Reshetikhin modules satisfy a nice system of recurrence relations, called T -system in the physics literature, which allows to calculate them inductively in terms of the q -characters of the fundamental modules $L(\varpi_i, a)$. This was

conjectured by Kuniba, Nakanishi and Suzuki [44], and proved by Nakajima [51] (see also [31] for the non simply-laced cases). The q -characters of the fundamental modules can in turn be calculated by means of the Frenkel-Mukhin algorithm [19]. One should therefore regard the Kirillov-Reshetikhin modules as the most “accessible” simple $U_q(L\mathfrak{g})$ -modules. There are $n(\ell + 1)(\ell + 2)/2$ such modules in \mathcal{C}_ℓ , namely:

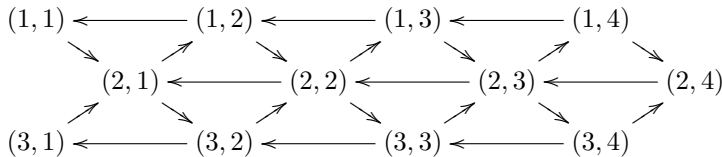
$$W_{j,q^{\varepsilon_i+2k}}^{(i)}, \quad (i \in I, 0 < j \leq \ell + 1, 0 \leq k \leq \ell + 1 - j).$$

9. The Cluster Algebras \mathcal{A}_ℓ

Let Q denote the quiver obtained by orienting the Dynkin diagram of \mathfrak{g} so that every $i \in I_0$ (resp. $i \in I_1$) is a source (resp. a sink). We define a new quiver Γ_ℓ with vertex set $\{(i, k) \mid i \in I, 1 \leq k \leq \ell + 1\}$. There are three types of arrows

- (a) arrows $(i, k) \rightarrow (j, k)$ for every arrow $i \rightarrow j$ in Q and every $1 \leq k \leq \ell + 1$;
- (b) arrows $(j, k) \rightarrow (i, k + 1)$ for every arrow $i \rightarrow j$ in Q and every $1 \leq k \leq \ell$;
- (c) arrows $(i, k) \leftarrow (i, k + 1)$ for every $i \in I$ and every $1 \leq k \leq \ell$.

For example, if \mathfrak{g} has type A_3 and $I_0 = \{1, 3\}$, the quiver Γ_3 is:



Let $\mathbf{x} = \{x_{(i,k)} \mid i \in I, 1 \leq k \leq \ell + 1\}$ be a set of indeterminates corresponding to the vertices of Γ_ℓ , and consider the seed $(\mathbf{x}, \Gamma_\ell)$ in which the n variables $x_{(i,\ell+1)}$ ($i \in I$) are frozen. This is the initial seed of a cluster algebra $\mathcal{A}_\ell \subset \mathbb{Q}(\mathbf{x})$. By Theorem 2.1, if \mathfrak{g} has type A_1 then \mathcal{A}_ℓ has finite cluster type A_ℓ . Also, if $\ell = 1$, \mathcal{A}_ℓ has finite cluster type equal to the Dynkin type of \mathfrak{g} . Otherwise, except for a few small rank cases, \mathcal{A}_ℓ has infinitely many cluster variables.

Our partial conjectural solution of Problem 1.2 can be summarized as follows (see [33] for more details):

Conjecture 9.1. *There is a ring isomorphism $\iota_\ell : \mathcal{A}_\ell \xrightarrow{\sim} K_0(\mathcal{C}_\ell)$ such that*

$$\iota_\ell(x_{(i,k)}) = \left[W_{k,q^{\varepsilon_i+2(\ell+1-k)}}^{(i)} \right], \quad (i \in I, 1 \leq k \leq \ell + 1).$$

The images by ι_ℓ of the cluster variables are classes of prime simple modules,

and the images of the cluster monomials are the classes of all real simple modules in \mathcal{C}_ℓ , i.e. those simple modules whose tensor square is simple.

Thus, if true, Conjecture 9.1 gives a combinatorial description in terms of cluster algebras of the prime tensor factorization of every real simple module. Note that, by definition, the square of a cluster monomial is again a cluster monomial. This explains why cluster monomials can only correspond to real simple modules. For $\mathfrak{g} = \mathfrak{sl}_2$, all simple $U_q(L\mathfrak{g})$ -modules are real. However for $\mathfrak{g} \neq \mathfrak{sl}_2$ there exist *imaginary* simple $U_q(L\mathfrak{g})$ -modules (i.e. simple modules whose tensor square is not simple), as shown in [45]. This is consistent with the expectation that a cluster algebra with infinitely many cluster variables is not spanned by its set of cluster monomials.

We arrived at Conjecture 9.1 by noting that the T -system equations satisfied by Kirillov-Reshetikhin modules are of the same form as the cluster exchange relations. This was inspired by the seminal work of Fomin and Zelevinsky [15], in which cluster algebra combinatorics is used to prove Zamolodchikov’s periodicity conjecture for Y -systems attached to Dynkin diagrams. Kedem [39] and Di Francesco [13], Keller [41, 42], Inoue, Iyama, Kuniba, Nakanishi and Suzuki [34], have also exploited the similarity between cluster exchange relations and other types of functional equations arising in mathematical physics (Q -systems, generalized T -systems, Y -systems attached to pairs of simply-laced Dynkin diagrams). Recently, Inoue, Iyama, Keller, Kuniba and Nakanishi [35, 36] have obtained a proof of the periodicity conjecture for all T -systems and Y -systems attached to a non simply-laced quantum affine algebra.

As evidence for Conjecture 9.1, we can easily check that for $\mathfrak{g} = \mathfrak{sl}_2$ and any $\ell \in \mathbb{N}$, it follows from the results of Chari and Pressley [7]. On the other hand, for arbitrary \mathfrak{g} we have:

Theorem 9.2 ([33, 52]). *Conjecture 9.1 holds for \mathfrak{g} of type A, D, E and $\ell = 1$.*

This was first proved in [33] for type A and D_4 by combinatorial and representation-theoretic methods, and soon after, by Nakajima [52] in the general case, by using the geometric description of the simple $U_q(L\mathfrak{g})$ -modules. In both approaches, a crucial part of the proof can be summarized in the following chart:

$$\begin{array}{ccc}
 F\text{-polynomials} & \leftrightarrow & \text{quiver Grassmannians} \\
 \updownarrow & & \updownarrow \\
 q\text{-characters} & \leftrightarrow & \text{Nakajima quiver varieties}
 \end{array}$$

Here, the F -polynomials are certain polynomials introduced by Fomin and Zelevinsky [18] which allow to calculate the cluster variables in terms of a fixed initial seed. By work of Caldero-Chapoton [5], Fu-Keller [21] and Derksen-Weyman-Zelevinsky [11, 12], F -polynomials have a geometric description via Grassmannians of subrepresentations of some quiver representations attached to cluster variables: this is the upper horizontal arrow of our diagram. The lower

horizontal arrow refers to the already mentioned relation between irreducible q -characters and perverse sheaves on quiver varieties established by Nakajima [49, 50]. In [33] we have shown that the F -polynomials for \mathcal{A}_1 are equal to certain natural truncations of the corresponding irreducible q -characters of \mathcal{C}_1 (the left vertical arrow), and we observed that this yielded an alternative geometric description of these q -characters in terms of ordinary homology of quiver Grassmannians. In [52] Nakajima used a Deligne-Fourier transform to obtain a direct relation between perverse sheaves on quiver varieties for \mathcal{C}_1 and homology of quiver Grassmannians (the right vertical arrow), and deduced from it the desired connection with the cluster algebra \mathcal{A}_1 .

The other main step in the approach of [33] is a certain tensor product theorem for the category \mathcal{C}_1 . It states that a tensor product $S_1 \otimes \cdots \otimes S_k$ of simples objects of \mathcal{C}_1 is simple if and only if $S_i \otimes S_j$ is simple for every pair $1 \leq i < j \leq k$. A generalization of this theorem to the whole category $\text{mod } U_q(L\mathfrak{g})$ has been recently proved by Hernandez [32]. Note that the theorem of Hernandez is also valid for non simply-laced Lie algebras \mathfrak{g} , and thus opens the way to a similar treatment of Problem 1.2 in this case.

Conjecture 9.1 has also been checked for \mathfrak{g} of type A_2 and $\ell = 2$ [33, §13]. In that small rank case, \mathcal{A}_2 still has finite cluster type D_4 , and this implies that \mathcal{C}_2 has only real objects. There are 18 explicit prime simple objects with respective dimensions

$$3, 3, 3, 3, 3, 3, 6, 6, 6, 6, 8, 8, 8, 10, 10, 15, 15, 35,$$

and 50 factorization patterns (corresponding to the 50 vertices of a generalized associahedron of type D_4 [17]). Our proof in this case is quite indirect and uses a lot of ingredients: the quantum affine Schur-Weyl duality, Ariki's theorem for type A affine Hecke algebras [1], the coincidence of Lusztig's dual canonical and dual semicanonical bases of $\mathbb{C}[N]$ in type A_4 [22], and Theorem 6.4.

One remarkable consequence of Theorem 9.2 from the point of view of cluster algebras is that it immediately implies the positivity conjecture of Fomin and Zelevinsky for the cluster algebras \mathcal{A}_1 with respect to any reference cluster (see [33, §2]). Conjecture 9.1 would similarly yield positivity for the whole class of cluster algebras \mathcal{A}_ℓ .

10. An Intriguing Relation

Problem 1.1 and Problem 1.2 may not be as unrelated as it would first seem. For a suggestive example, let us take \mathfrak{g} of type A_3 . In that case, the abelian category $\text{mod } \Lambda$ has 12 indecomposable objects (which are all rigid), 3 of them being projective-injective. On the other hand the tensor category \mathcal{C}_1 has 12 prime simple objects (which are all real), 3 of them having the property that their tensor product with every simple of \mathcal{C}_1 is simple. It is easy to check that $\mathbb{C}[N]$ and $\mathbb{C} \otimes_{\mathbb{Z}} K_0(\mathcal{C}_1)$ are isomorphic as (complexified) cluster algebras with

frozen variables. Therefore we have a unique one-to-one correspondence

$$X \leftrightarrow S$$

between rigid objects X of $\text{mod } \Lambda$ and simple objects S of \mathcal{C}_1 such that

$$\varphi_X \equiv [S],$$

that is, such that X and S project to the same cluster monomial. In this correspondence, direct sums $X \oplus X'$ map to tensor products $S \otimes S'$. It would be interesting to find a general framework for relating in a similar way, via cluster algebras, certain additive categories such as $\text{mod } \Lambda$ to certain tensor categories such as \mathcal{C}_1 . We refer to [43] for a very accessible survey of these ideas.

Acknowledgements

I wish to thank Christof Geiss and Jan Schröer for their very precious collaboration and friendship over the years, and David Hernandez for his stimulating enthusiasm.

References

- [1] S. Ariki, *On the decomposition numbers of the Hecke algebra of $G(n, 1, m)$* , J. Math. Kyoto Univ. **36** (1996), 789–808.
- [2] A. Berenstein, S. Fomin, A. Zelevinsky, *Cluster algebras III: Upper bounds and double Bruhat cells*, Duke Math. J. **126** (2005), 1–52.
- [3] A. Buan, O. Iyama, I. Reiten, J. Scott, *Cluster structures for 2-Calabi-Yau categories and unipotent groups*, Compos. Math. **145** (2009), 1035–1079.
- [4] A. Buan, R. Marsh, M. Reineke, I. Reiten, G. Todorov, *Tilting theory and cluster combinatorics*, Adv. Math. **204** (2006), 572–618.
- [5] P. Caldero, F. Chapoton, *Cluster algebras as Hall algebras of quiver representations*, Comment. Math. Helv. **81** (2006), 595–616.
- [6] P. Caldero, B. Keller, *From triangulated categories to cluster algebras*, Invent. Math. **172** (2008), 169–211.
- [7] V. Chari, A. Pressley, *Quantum affine algebras*, Comm. Math. Phys. **142** (1991), 261–283.
- [8] V. Chari, A. Pressley, *A guide to quantum groups*, Cambridge University Press, 1994.
- [9] V. Chari, A. Pressley, *Quantum affine algebras and their representations*, CMS Conf. Proc. **16** (1995), 59–78.
- [10] W. Crawley-Boevey, *On the exceptional fibres of Kleinian singularities*, Amer. J. Math. **122** (2000), 1027–1037.

-
- [11] H. Derksen, J. Weyman, A. Zelevinsky, *Quivers with potentials and their representations I: Mutations*, *Selecta Math. New ser.* **14** (2008), 59–119.
- [12] H. Derksen, J. Weyman, A. Zelevinsky, *Quivers with potentials and their representations II: Applications to cluster algebras*, arXiv:0904.0676, *J. Amer. Math. Soc.* (to appear).
- [13] P. Di Francesco, R. Kedem, *Q-systems as cluster algebras II: Cartan matrix of finite type and the polynomial property*, *Lett. Math. Phys.* **89** (2009), no. 3, 183–216.
- [14] S. Fomin, *Total positivity and cluster algebras*, *Proceedings of the International Congress of Mathematicians, Hyderabad*, 2010.
- [15] S. Fomin, A. Zelevinsky, *Y-systems and generalized associahedra*, *Ann. Math.* **158** (2003), 977–1018.
- [16] S. Fomin, A. Zelevinsky, *Cluster algebras I: Foundations*, *J. Amer. Math. Soc.* **15** (2002), 497–529.
- [17] S. Fomin, A. Zelevinsky, *Cluster algebras II: Finite type classification*, *Invent. Math.* **154** (2003), 63–121.
- [18] S. Fomin, A. Zelevinsky, *Cluster algebras IV: Coefficients*, *Compos. Math.* **143** (2007), 112–164.
- [19] E. Frenkel, E. Mukhin, *Combinatorics of q-characters of finite-dimensional representations of quantum affine algebras*, *Comm. Math. Phys.* **216** (2001), 23–57.
- [20] E. Frenkel, N. Reshetikhin, *The q-characters of representations of quantum affine algebras*, Recent developments in quantum affine algebras and related topics, *Contemp. Math.* **248** (1999), 163–205.
- [21] C. Fu, B. Keller, *On cluster algebras with coefficients and 2-Calabi-Yau categories*, *Trans. Amer. Math. Soc.* **362** (2010), 859–895.
- [22] C. Geiss, B. Leclerc, J. Schröer, *Semicanonical bases and preprojective algebras*, *Ann. Scient. Éc. Norm. Sup.* **38** (2005), 193–253.
- [23] C. Geiss, B. Leclerc, J. Schröer, *Rigid modules over preprojective algebras*, *Invent. Math.*, **165** (2006), 589–632.
- [24] C. Geiss, B. Leclerc, J. Schröer, *Auslander algebras and initial seeds for cluster algebras*, *J. London Math. Soc.*, **75** (2007), 718–740.
- [25] C. Geiss, B. Leclerc, J. Schröer, *Semicanonical bases and preprojective algebras II: A multiplication formula*, *Compositio Math.*, **143** (2007), 1313–1334.
- [26] C. Geiss, B. Leclerc, J. Schröer, *Partial flag varieties and preprojective algebras*, *Ann. Institut Fourier*, **58** (2008), 825–876.
- [27] C. Geiss, B. Leclerc, J. Schröer, *Kac-Moody groups and cluster algebras*, arXiv:1001.3545.
- [28] C. Geiss, J. Schröer, *Extension-orthogonal components of preprojective varieties*, *Trans. Amer. Math. Soc.* **357** (2005), 1953–1962.
- [29] I. M. Gelfand, V.A. Ponomarev, *Model algebras and representations of graphs*, *Functional Anal. Appl.* **13** (1979), 157–165.
- [30] V. Ginzburg, E. Vasserot, *Langlands reciprocity for affine quantum groups of type A_n* , *I.M.R.N.* **3** (1993), 67–85.

- [31] D. Hernandez, *The Kirillov-Reshetikhin conjecture and solutions of T -systems*, J. Reine Angew. Math. **596** (2006), 63–87.
- [32] D. Hernandez, *Simple tensor products*, arXiv:0907.3002.
- [33] D. Hernandez, B. Leclerc, *Cluster algebras and quantum affine algebras*, Duke Math. J. (to appear), arXiv:0903.1452.
- [34] R. Inoue, O. Iyama, A. Kuniba, T. Nakanishi, J. Suzuki, *Periodicities of T -systems and Y -systems*, Nagoya Math. J. (to appear), arXiv: 0812.0667.
- [35] R. Inoue, O. Iyama, B. Keller, A. Kuniba, T. Nakanishi, *Periodicities of T -systems and Y -systems, dilogarithm identities, and cluster algebras I: Type B_r* , arXiv: 1001.1880.
- [36] R. Inoue, O. Iyama, B. Keller, A. Kuniba, T. Nakanishi, *Periodicities of T -systems and Y -systems, dilogarithm identities, and cluster algebras II: Type C_r , F_4 , and G_2* , arXiv: 1001.1881.
- [37] M. Jimbo, *A q -analogue of $U(\mathfrak{gl}(N+1))$, Hecke algebra and the Yang-Baxter equation*, Lett. Math. Phys. **11** (1986), 247–252.
- [38] M. Jimbo, *Solvable lattice models and quantum groups*, Proceedings of the International Congress of Mathematicians, Kyoto, Japan, 1990, 1343–1352, Springer.
- [39] R. Kedem, *Q -systems as cluster algebras*, J. Phys. A **41** (2008), no. 19, 194011, 14 pp.
- [40] B. Keller, *On triangulated orbit categories*, Doc. Math. **10** (2005), 551–581.
- [41] B. Keller, *Cluster algebras, quiver representations and triangulated categories*, arXiv:0807.1960.
- [42] B. Keller, *The periodicity conjecture for pairs of Dynkin diagrams*, arXiv:1001.1531.
- [43] B. Keller, *Algèbres amassées et applications*, Séminaire Bourbaki Exposé 1014, november 2009.
- [44] A. Kuniba, T. Nakanishi, J. Suzuki, *Functional relations in solvable lattice models. I. Functional relations and representation theory*. Internat. J. Modern Phys. A **9** (1994), 5215–5266,
- [45] B. Leclerc, *Imaginary vectors in the dual canonical basis of $U_q(\mathfrak{n})$* , Transformation Groups, **8** (2003), 95–104.
- [46] G. Lusztig, *Quivers, perverse sheaves, and quantized enveloping algebras*, J. Amer. Math. Soc. **4** (1991), 365–421.
- [47] G. Lusztig, *Semicanonical bases arising from enveloping algebras*, Adv. Math. **151** (2000), no. 2, 129–139.
- [48] G. Musiker, R. Schiffler, L. Williams, *Positivity for cluster algebras from surfaces*, arxiv:0906.0748.
- [49] H. Nakajima, *Quiver varieties and finite-dimensional representations of quantum affine algebras*. J. Amer. Math. Soc. **14** (2001), 145–238.
- [50] H. Nakajima, *Quiver varieties and t -analogs of q -characters of quantum affine algebras*, Ann. Math. **160** (2004), 1057–1097.

- [51] H. Nakajima, *t-analogs of q-characters of Kirillov-Reshetikhin modules of quantum affine algebras*, Represent. Theory **7** (2003), 259–274.
- [52] H. Nakajima, *Quiver varieties and cluster algebras*, arXiv.0905.0002.
- [53] C.M. Ringel, *The preprojective algebra of a quiver*, Algebras and modules, II (Geiranger, 1996), 467–480, CMS Conf. Proc., **24**, Amer. Math. Soc., Providence, RI, 1998.

Subgraphs of Random Graphs with Specified Degrees

Brendan D. McKay*

Abstract

If a graph is chosen uniformly at random from all the graphs with a given degree sequence, what can be said about its subgraphs? The same can be asked of bipartite graphs, equivalently 0-1 matrices. These questions have been studied by many people. In this paper we provide a partial survey of the field, with emphasis on two general techniques: the method of switchings and the multidimensional saddle-point method.

Mathematics Subject Classification (2010). Primary 05C80; Secondary 05A16, 60B20

Keywords. Random graphs, vertex degree, subgraph, regular graph

1. Introduction

In this paper we will be concerned with *simple* graphs: those having no loops or parallel edges. Two classes of simple graphs will be considered, *generic* graphs, and *bipartite* graphs. In the former case, there are n vertices any two of which may be adjacent. In the latter case, there are two disjoint *classes* of respectively m and n vertices, and all edges must have one vertex from each set. The phrase “generic graph” is not standard, but we adopt it here for the sake of clarity. If we refer merely to “graph”, we might mean either type.

The *degree* of a vertex is the number of edges incident to it, and the *degree sequence* of a graph is a list of the degrees of the vertices. In the case of generic graphs, we will denote the degree sequence by $\mathbf{d} = (d_1, d_2, \dots, d_n)$. It satisfies the conditions that $0 \leq d_i \leq n - 1$ for each i , and $\sum_i d_i$ is even. Let $\mathcal{G}(\mathbf{d})$ be the set of all generic graphs with degree sequence \mathbf{d} .

*Supported by the Australian Research Council.

School of Computer Science, Australian National University, Canberra, ACT 0200, Australia. E-mail: bdm@cs.anu.edu.au.

In the case of bipartite graphs, we will denote the degree sequence by (\mathbf{s}, \mathbf{t}) , where $\mathbf{s} = (s_1, s_2, \dots, s_m)$ are the degrees in one class and $\mathbf{t} = (t_1, t_2, \dots, t_n)$ are the degrees in the other class. We have the conditions $0 \leq s_j \leq n$ for each j , $0 \leq t_k \leq m$ for each k , and $\sum_j s_j = \sum_k t_k$. Let $\mathcal{B}(\mathbf{s}, \mathbf{t})$ be the set of all bipartite graphs with degree sequence (\mathbf{s}, \mathbf{t}) . Examples appear in Figure 1.

In each case, stronger conditions on the degree sequence are needed before a graph with that degree sequence can be guaranteed to exist, but we will not require those conditions here.

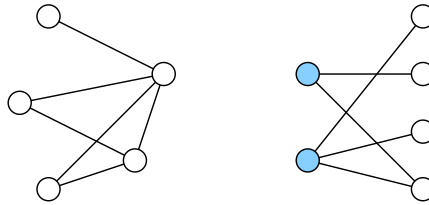


Figure 1. Members of $\mathcal{G}((1, 2, 2, 4, 3))$ and $\mathcal{B}((2, 3), (1, 1, 1, 2))$

If $\mathcal{G}(\mathbf{d}) \neq \emptyset$, which we will assume from now on, we can promote it to a probability space by assigning each element the same probability. It is this space that we refer to when we discuss a “random generic graph with degree sequence \mathbf{d} ”. Similarly, for a “random bipartite graph with degree sequence (\mathbf{s}, \mathbf{t}) ”.

There is a fair amount of literature on random graphs of these types, some of which we will cite as we go. In this incomplete survey we will focus on a particular issue: what is the probability that a specified subgraph occurs? More generally we can ask for the distribution of the number of subgraphs of a given type. Our asymptotics will be with respect to $n \rightarrow \infty$ for generic graphs, or $m, n \rightarrow \infty$ for bipartite graphs, with other parameters such as \mathbf{d} being functions of n , or (m, n) , unless otherwise specified.

Since we are dealing with uniform discrete distributions, our probability questions are just counting questions in disguise. If \mathbf{X} is a generic graph, let $\mathcal{G}(\mathbf{d}, \mathbf{X})$ denote the set of generic graphs with degree sequence \mathbf{d} and no edge in common with \mathbf{X} . Then, if \mathbf{x} is the degree sequence of \mathbf{X} , the probability that a random generic graph with degree sequence \mathbf{d} has \mathbf{X} as a subgraph is

$$\mathbf{P}_{\mathcal{G}(\mathbf{d})}(\mathbf{X}) = \frac{|\mathcal{G}(\mathbf{d}-\mathbf{x}, \mathbf{X})|}{|\mathcal{G}(\mathbf{d})|}. \tag{1}$$

Similarly, if \mathbf{X} is a bipartite graph with classes of size m and n , let $\mathcal{B}(\mathbf{s}, \mathbf{t}, \mathbf{X})$ denote the set of bipartite graphs with degree sequence (\mathbf{s}, \mathbf{t}) and no edge in common with \mathbf{X} . Then, if (\mathbf{x}, \mathbf{y}) is the degree sequence of \mathbf{X} , the probability that a random bipartite graph with degree sequence (\mathbf{s}, \mathbf{t}) has \mathbf{X} as a subgraph is

$$\mathbf{P}_{\mathcal{B}(\mathbf{s}, \mathbf{t})}(\mathbf{X}) = \frac{|\mathcal{B}(\mathbf{s}-\mathbf{x}, \mathbf{t}-\mathbf{y}, \mathbf{X})|}{|\mathcal{B}(\mathbf{s}, \mathbf{t})|}. \tag{2}$$

Largely due to the different techniques that have been fruitful, we divide the discussion into two parts. In Section 2, we consider the case where the degrees are low, such as when they are bounded. By complementation, this also applies when the degrees are almost as large as possible. In Section 3, we consider the case where the degrees are something like a constant fraction of the number of vertices.

2. Sparse Graphs

For this section we consider random generic or bipartite graphs whose degrees do not grow very quickly with the size of the graph.

Define $d_{\max} = \max\{d_1, \dots, d_n\}$, and similarly s_{\max} , t_{\max} , x_{\max} and y_{\max} . For integer $k \geq 0$, we write $(a)_k = a(a-1)\cdots(a-k+1)$.

The most celebrated technique is called the *configuration model* or *pairing model*. The popular version of it was introduced by Bollobás [2], though the concept has an older history, see [33]. We describe it for generic graphs of degree sequence \mathbf{d} ; an obvious variant works the same for bipartite graphs.

Consider n disjoint *cells* v_1, \dots, v_n , where cell v_i is a set of d_i *points*. This makes $2E = \sum_i d_i$ points in total (recall that $\sum_i d_i$ must be even). Choose a random *pairing* (partition of the points into E pairs), where each of the $(2E)!/(E! 2^E)$ possible pairings are equally likely. A pairing P is *simple* if each pair involves two different cells, and no two pairs involve the same two cells. In that case we can make a graph $\mathbf{G}(P)$ whose vertices are v_1, \dots, v_n and whose edges are those $v_j v_k$ such that there is a pair involving v_j and v_k . Clearly $\mathbf{G}(P) \in \mathcal{G}(\mathbf{d})$.

The key feature of pairings is that each graph in $\mathcal{G}(\mathbf{d})$ corresponds to exactly $\prod_j d_j!$ simple pairings. Therefore, a random simple pairing yields a random graph in $\mathcal{G}(\mathbf{d})$ (i.e., with the uniform distribution). If $P_2(\mathbf{d}, \mathbf{X})$ is the probability that a random pairing is simple and avoids the graph \mathbf{X} , then

$$|\mathcal{G}(\mathbf{d}, \mathbf{X})| = \frac{(2E)!}{E! 2^E \prod_j d_j!} P_2(\mathbf{d}, \mathbf{X}).$$

The other value $|\mathcal{G}(\mathbf{d})|$ required by (1) is just the special case of $\mathbf{X} = \emptyset$, where \emptyset is the graph with no edges. So the subgraph probability problem reduces to the sometimes easier calculation of the probabilities $P_2(\mathbf{d}, \mathbf{X})$.

If d_{\max} is at most slowly increasing, for example if $d_{\max} = O((\log n)^{1/3})$, then $P_2(\mathbf{d}, \mathbf{X})$ can be estimated under mild additional conditions on \mathbf{d} and \mathbf{X} using inclusion-exclusion or the method of moments, see Bollobás and McKay [3]. For $d_{\max} = O(1)$ (refer to Janson [14] for necessary and sufficient conditions), $P_2(\mathbf{d}, \emptyset)$ is bounded above 0, which has an immediate dramatic consequence: every event that is asymptotically unlikely or certain for random pairings is also asymptotically unlikely or certain (respectively!) for random generic graphs with degree sequence \mathbf{d} . A great many theorems are based

on this observation, and the equivalent observation for bipartite graphs, see Wormald [33] and Janson et al. [15] for summaries.

When d_{\max} increases more quickly with n , the same methods do not suffice to estimate $P_2(\mathbf{d}, \mathbf{X})$. For example, the terms of the inclusion-exclusion expansion cancel too precisely to allow estimation of their sum. An alternative method is required, which is where the method of switchings comes in.

The basic idea behind the method of switchings is the following: given two finite sets A, B and a relation R (in this context called a *switching operation*) between them, then the ratio of the average number of elements of B related to each element of A to the average number of elements of A related to each element of B is the same as the ratio of $|B|$ to $|A|$. This idea can be applied to the problem of subgraph probabilities in two different ways. In the first approach graphs with a given degree sequence are manipulated directly. In the second approach, switchings are used to analyse pairings.

We first consider the direct application of switching to subgraph probabilities. Following [20], we generalise the notation $\mathcal{G}(\mathbf{d}, \mathbf{X})$ to $\mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{Y})$, where \mathbf{Y} is a subgraph of \mathbf{X} : $\mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{Y})$ is the set of all generic graphs $\mathbf{G} \in \mathcal{G}(\mathbf{d})$ such that the intersection of \mathbf{G} and \mathbf{X} (considered as sets of edges) is exactly \mathbf{Y} . We can see that $\mathcal{G}(\mathbf{d}) = \bigcup_{\mathbf{Y} \subseteq \mathbf{X}} \mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{Y})$, $\mathcal{G}(\mathbf{d}, \mathbf{X}) = \mathcal{G}(\mathbf{d}, \mathbf{X}, \emptyset)$, and $|\mathcal{G}(\mathbf{d}-\mathbf{x}, \mathbf{X})| = |\mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{X})|$. Therefore,

$$P_{\mathcal{G}(\mathbf{d})}(\mathbf{X}) = \left(\sum_{\mathbf{Y} \subseteq \mathbf{X}} \frac{|\mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{Y})|}{|\mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{X})|} \right)^{-1}. \tag{3}$$

Let $e = ab$ be an edge of \mathbf{X} that is not an edge of \mathbf{Y} . We can define a relation between $\mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{Y} \cup ab)$ and $\mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{Y})$ using the switching operation shown in Figure 2. If the left diagram appears in a graph $\mathbf{G} \in \mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{Y} \cup ab)$, $ac, bd \notin \mathbf{G}$, and $ac, cd, bd \notin \mathbf{X}$, then replacing it by the right diagram produces a graph in $\mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{Y})$.

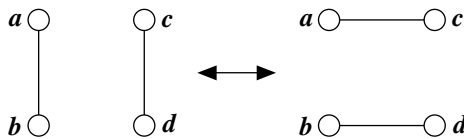


Figure 2. A simple switching operation

By bounding the number of switching operations that can apply to \mathbf{G} , and similarly bounding the number of ways of coming back from $\mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{Y})$, we obtain bounds on the ratio of $|\mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{Y} \cup ab)|$ to $|\mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{Y})|$. Combining all such ratios to obtain the relative sizes of $\mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{Y})$ for all $\mathbf{Y} \subseteq \mathbf{X}$, we can finally apply (3) to get $P_{\mathcal{G}(\mathbf{d})}(\mathbf{X})$.

The following is a consequence of Theorems 2.9 and 2.10 of McKay [20]. Define $\Delta = d_{\max}(d_{\max} + x_{\max})$ and $X = \frac{1}{2} \sum_i x_i$.

Theorem 2.1 ([20]). *If $\Delta X = o(E)$ then*

$$P_{\mathcal{G}(\mathbf{d})}(\mathbf{X}) = \frac{\prod_{j=1}^n (d_j)_{x_j}}{2^X (E)_X} (1 + O(\Delta X/E)).$$

In the bipartite case, we can use the same switching operation provided a and c are in opposite vertex classes. Define $\Delta' = (s_{\max} + t_{\max})(s_{\max} + t_{\max} + x_{\max} + y_{\max})$ and $X = \sum_j x_j = \sum_k y_k$.

Theorem 2.2 ([20]). *If $\Delta' X = o(E)$ then*

$$P_{\mathcal{B}(s,t)}(\mathbf{X}) = \frac{\prod_{j=1}^m (s_j)_{x_j} \prod_{k=1}^n (t_k)_{y_k}}{(E)_X} (1 + O(\Delta' X/E)).$$

For both Theorems 2.1 and 2.2, the exact bounds given in [20] can be useful even when the error term is not vanishing.

Since [20], two improvements to this method have been found. As first shown by McKay and Wormald [26] in a slightly different context, the counting is substantially easier if the more complex switching operation of Figure 3 is used.

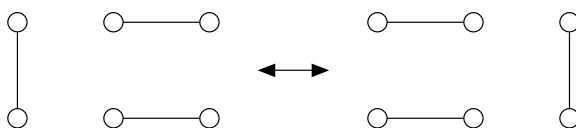


Figure 3. A better switching operation

The other improvement, introduced by Lieby, McKay, McLeod and Wanless [18], is a rearrangement of the calculation. Let the edges of \mathbf{X} be e_1, e_2, \dots, e_X and define \mathbf{X}_j to be the graph with edges $\{e_1, \dots, e_j\}$, $0 \leq j \leq X$. For $j \geq 1$, we have $\mathcal{G}(\mathbf{d}, \mathbf{X}_{j-1}, \mathbf{X}_{j-1}) = \mathcal{G}(\mathbf{d}, \mathbf{X}_j, \mathbf{X}_{j-1}) \cup \mathcal{G}(\mathbf{d}, \mathbf{X}_j, \mathbf{X}_j)$, and so

$$P_{\mathcal{G}(\mathbf{d})}(\mathbf{X}) = \frac{|\mathcal{G}(\mathbf{d}, \mathbf{X}, \mathbf{X})|}{|\mathcal{G}(\mathbf{d}, \emptyset, \emptyset)|} = \prod_{j=1}^X \left(1 + \frac{|\mathcal{G}(\mathbf{d}, \mathbf{X}_j, \mathbf{X}_{j-1})|}{|\mathcal{G}(\mathbf{d}, \mathbf{X}_j, \mathbf{X}_j)|} \right)^{-1},$$

assuming all the denominators are nonzero. The ratio of $|\mathcal{G}(\mathbf{d}, \mathbf{X}_j, \mathbf{X}_{j-1})|$ to $|\mathcal{G}(\mathbf{d}, \mathbf{X}_j, \mathbf{X}_j)|$ can be obtained by analysing a switching, as before. This method avoids the problematic sum in (3), and also allows the ordering of the edges of \mathbf{X} to be tuned to optimise the precision of the answer.

As we mentioned, the other way to apply switchings is to use them to analyse the pairing model. Recall that the task is to estimate the probability $P_2(\mathbf{d}, \mathbf{X})$ that a random pairing is simple and avoids \mathbf{X} . The basic idea is to classify pairings according to their non-simple parts (such as double pairs or pairs hitting \mathbf{X}). Then switching operations are used to estimate the relative sizes of these classes. This was first done by McKay [22] for generic graphs and McKay [21] for bipartite graphs; we summarise the main theorems below. Define E , Δ and Δ' as before.

Theorem 2.3 ([21, 22]).

(a) Suppose $d_{\max} \geq 1$ and $\Delta = o(E^{1/2})$. Then, as $n \rightarrow \infty$,

$$|\mathcal{G}(\mathbf{d}, \mathbf{X})| = \frac{(2E)!}{E! 2^E \prod_{j=1}^n d_j!} \exp\left(-\frac{\sum_{j=1}^n d_j(d_j-1)}{4E} - \frac{(\sum_{j=1}^n d_j(d_j-1))^2}{16E^2} - \frac{\sum_{jk \in \mathbf{X}} d_j d_k}{2E} + O(\Delta^2/E)\right).$$

(b) Suppose $s_{\max} \geq 1$ and $\Delta' = o(E^{1/2})$. Then, as $m, n \rightarrow \infty$,

$$|\mathcal{B}(\mathbf{s}, \mathbf{t}, \mathbf{X})| = \frac{E!}{\prod_{j=1}^m s_j! \prod_{k=1}^n t_k!} \exp\left(-\frac{\sum_{j=1}^m s_j(s_j-1) \sum_{k=1}^n t_k(t_k-1)}{2E^2} - \frac{\sum_{jk \in \mathbf{X}} s_j t_k}{E} + O(\Delta'^2/E)\right).$$

In the above, the notion $\sum_{jk \in \mathbf{X}}$ means a sum over unordered pairs $\{j, k\}$ such that jk is an edge of \mathbf{X} , with j being in the first class for the bipartite case.

The special case $P_2(\mathbf{d}, \emptyset)$, needed for estimating $|\mathcal{G}(\mathbf{d})|$ was improved McKay and Wormald [27] to cover generic graphs with $d_{\max} = o(E^{1/3})$, and by Greenhill, McKay and Wang [12] to cover bipartite graphs with $s_{\max} t_{\max} = o(E^{2/3})$. An example of a switching operation used by these papers is shown in Figure 4, where the shaded ovals represent the cells of the pairing.

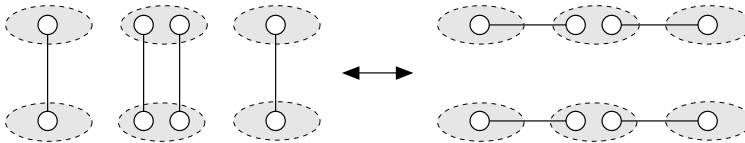


Figure 4. Removing a double pair from a pairing

The distribution of cycle counts in random regular graphs has been studied quite a lot. For fixed or very slowly increasing degree, the counts of fixed length cycles are asymptotically Poisson and independent, as shown by Bollobás [2] and Wormald [32]. Counts of longer cycles were studied by Garmó [9]. By using switching operations specifically tailored for the purpose, McKay, Wormald and Wysocka [28] found the joint distribution of the counts of cycles up to length g in a random regular graph of order n and degree d , whenever $(d-1)^{2g-1} = o(n)$. Gao and Wormald [8] found the central part of the distribution of the number of cycles of length g under the weaker condition $d = o(n^{2/(3g-2)})$, as a special case of a theory (developed in [7]) that allows asymptotic normality of the counts of many small subgraphs to be inferred from certain higher moments.

Perhaps the deepest result of this nature was that of Robinson and Wormald [30, 31] who showed that almost all regular graphs of fixed degree $d \geq 3$ are hamiltonian. The somewhat easier problem of extending this to all $d \geq 3$ was achieved later in [5, 17].

Counts of perfect matchings in the regular cases of $\mathcal{G}(\mathbf{d})$ and $\mathcal{B}(\mathbf{s}, \mathbf{t})$ for small degree were studied by Bollobás and McKay [3]. The expectation is also found in [3] in the bipartite case for extreme degrees ($m = n$ and degree at least $n - n^{1-\epsilon}$) using enumeration results for Latin rectangles [10]. A similar calculation for extreme-degree generic graphs could easily be done starting with the results in [29].

The furthest reach of the switching method to higher vertex degrees was achieved by Krivelevich, Sudakov, Vu and Wormald [17], who determined several almost-sure properties of random regular graphs of degree $o(n)$.

Ben-Shimon and Krivelevich [1] used switchings to study the number of edges spanned by a set of vertices, or between two sets of vertices, in regular graphs of degree $o(n^{1/2})$.

3. Dense Graphs

The methods of the previous section are most suitable when the graph degrees are relatively small. The exception is that Theorems 2.1 and 2.2 can provide the probability of very small subgraphs for higher degrees in some cases.

Define a new parameter $\lambda \in [0, 1]$, which we call the *density*. For generic graphs, $\lambda = E/\binom{n}{2}$. For bipartite graphs, $\lambda = E/(mn)$.

It will be worth comparing the subgraph probabilities in $\mathcal{G}(\mathbf{d})$ and $\mathcal{B}(\mathbf{s}, \mathbf{t})$ to the probabilities in similar binomial random graph models. Let $\mathcal{G}_{n,p}$ be the probability space of random generic graphs with n vertices and edge probability p (i.e., each possible edge is present with independent probability p), and let $\mathcal{B}_{m,n,p}$ be the similar space of random bipartite graphs with vertex classes of size m and n . Intuition suggests that subgraph probability $\mathbf{P}_{\mathcal{G}_{n,\lambda}}(\mathbf{X}) = \lambda^{\mathbf{X}}$ may be a rough approximation to those in $\mathbf{P}_{\mathcal{G}(\mathbf{d})}(\mathbf{X})$, and similarly for $\mathbf{P}_{\mathcal{B}_{m,n,\lambda}}(\mathbf{X}) = \lambda^{\mathbf{X}}$ versus $\mathbf{P}_{\mathcal{B}(\mathbf{s},\mathbf{t})}(\mathbf{X})$.

The strongest results of this type were proved by Greenhill and McKay [11] and McKay [24]. We will start with generic graphs and need the following additional parameters, for $\ell, m \geq 1$.

$$\begin{aligned} \bar{d} &= \frac{1}{n} \sum_{i=1}^n d_i = \lambda(n-1) = 2E/n \\ \delta_j &= d_j - \bar{d} + \lambda x_j \quad (1 \leq j \leq n), & X_\ell &= \sum_{j=1}^n x_j^\ell, \\ L &= \sum_{j,k \in \mathbf{X}} (\delta_j - x_j)(\delta_k - x_k), & C_{\ell,m} &= \sum_{j=1}^n \delta_j^\ell x_j^m. \end{aligned}$$

Theorem 3.1 ([24]). *Let $a, b > 0$ be constants such that $a + b < \frac{1}{2}$. For some $\varepsilon > 0$, suppose that $d_j - \bar{d}$ and x_j are uniformly $O(n^{1/2+\varepsilon})$ for $1 \leq j \leq n$, and that $X = O(n^{1+2\varepsilon})$. For sufficient large n , suppose that*

$$\min\{\bar{d}, n - \bar{d} - 1\} \geq \frac{n}{3a \log n}.$$

Then, provided ε is small enough, we have

$$\mathbf{P}_{\mathcal{G}(d)}(\mathbf{X}) = \lambda^X \exp\left(\frac{(1-\lambda)X}{\lambda n} - \frac{(1+\lambda)X_2}{2\lambda n} - \frac{(1+\lambda)(1+2\lambda)X_3}{6\lambda^2 n^2} + \frac{(1-\lambda)X^2}{\lambda n^2} - \frac{L}{\lambda(1-\lambda)n^2} + \frac{C_{1,1}}{\lambda n} + \frac{(1+2\lambda)C_{1,2}}{2\lambda^2 n^2} - \frac{C_{2,1}}{2\lambda^2 n^2} + O(n^{-b})\right).$$

A corollary of Theorem 3.1 is that $\mathbf{P}_{\mathcal{G}(d)}(\mathbf{X}) \sim \mathbf{P}_{\mathcal{G}_{n,\lambda}}(\mathbf{X})$ when

$$X \max_j |d_j - d| + (1 - \lambda)X_2 = o(\lambda n).$$

This sufficient condition holds, for example, if $X = O(n^{1/2-2\varepsilon})$, or if $d_j - \bar{d}$ and x_j are uniformly $O(n^\varepsilon)$ for $1 \leq j \leq n$ and $X = O(n^{1-2\varepsilon})$.

A special case of this result was proved by Krivelevich, Sudakov and Wormald [16], who determined the probability of induced subgraphs of $o(n^{1/2})$ vertices in random regular graphs of degree $(n - 1)/2$ under some conditions on the degree sequence of the subgraph.

For bipartite graphs, a similar result holds. Define the following parameters for integers $h, \ell \geq 0$.

$$\begin{aligned} \bar{s} &= \frac{1}{m} \sum_{j=1}^m s_j = E/m = \lambda n, & \bar{t} &= \frac{1}{n} \sum_{k=1}^n t_k = E/n = \lambda m, \\ \xi_j &= s_j - \bar{s} + \lambda x_j \quad (1 \leq j \leq m), & \eta_k &= t_k - \bar{t} + \lambda y_k \quad (1 \leq k \leq n) \\ Z &= \sum_{jk \in \mathbf{X}} (x_j - \xi_j)(y_k - \eta_k). & Q_{h,\ell} &= n^{1-h-\ell} \sum_{j=1}^m \xi_j^h x_j^\ell + m^{1-h-\ell} \sum_{k=1}^n \eta_k^h y_k^\ell \end{aligned}$$

Theorem 3.2 ([11]). *Let $a, b > 0$ be constants such that $a + b < \frac{1}{2}$. For some $\varepsilon > 0$, suppose that $m, n \rightarrow \infty$ with $n = o(m^{1+\varepsilon})$ and $m = o(n^{1+\varepsilon})$, and further that $s_j - \bar{s}$, x_j , $t_k - \bar{t}$ and y_k are uniformly $O(n^{1/2+\varepsilon})$ for $1 \leq j \leq m$ and $1 \leq k \leq n$, and $X = O(n^{1+2\varepsilon})$. Assume*

$$\frac{(1 - 2\lambda)^2}{4\lambda(1 - \lambda)} \left(1 + \frac{5m}{6n} + \frac{5n}{6m}\right) \leq a \log n.$$

Then, provided $\varepsilon > 0$ is small enough, we have

$$\mathbf{P}_{\mathcal{B}(\mathbf{s},\mathbf{t})}(\mathbf{X}) = \lambda^X \exp \left(\frac{(1-\lambda)X}{2\lambda} \left(\frac{1}{n} + \frac{1}{m} \right) + \frac{(1-\lambda)X^2}{2\lambda mn} + \frac{Q_{1,1}}{\lambda} - \frac{(1+\lambda)Q_{0,2}}{2\lambda} - \frac{Q_{2,1}}{2\lambda^2} + \frac{(1+2\lambda)Q_{1,2}}{2\lambda^2} - \frac{(1+\lambda)(1+2\lambda)Q_{0,3}}{6\lambda^2} - \frac{Z}{\lambda(1-\lambda)mn} + O(n^{-b}) \right).$$

A corollary of Theorem 3.2 is that $\mathbf{P}_{\mathcal{B}(\mathbf{s},\mathbf{t})}(\mathbf{X}) \sim \mathbf{P}_{\mathcal{B}_{m,n,\lambda}}(\mathbf{X})$ when

$$X \max_j |s_j - s| + (1-\lambda) \sum_j x_j^2 = o(\lambda n), \text{ and}$$

$$X \max_k |t_k - t| + (1-\lambda) \sum_k y_k^2 = o(\lambda m).$$

These extra requirements are met, for example, if $X = O(n^{1/2-2\varepsilon})$. Another interesting case is when $s_j - s, x_j, t_k - t$ and y_k are uniformly $O(n^\varepsilon)$ and $X = O(n^{1-2\varepsilon})$.

In [11] and [24] we also give the probability that \mathbf{X} is avoided, and the probability of occurrence of a specified induced subgraph. We also record the simpler formulae that are implied if the whole graph or the subgraph is regular.

We mention a sample application of Theorem 3.2. In the binomial model $\mathcal{B}_{n,n,\lambda}$, corresponding to square 0-1 matrices with each entry independently being 1 with probability λ , the expected permanent is exactly $n! \lambda^n$. It is interesting to see what the effect of specifying \mathbf{s} and \mathbf{t} (the row and column sums) is.

Corollary ([11]). *Suppose that $m = n$ and $\mathbf{s}, \mathbf{t}, \lambda$ satisfy the requirements of Theorem 3.2. Then the expected permanent of a random $n \times n$ matrix over $\{0, 1\}$ with row sums \mathbf{s} and column sums \mathbf{t} is*

$$n! \lambda^n \exp \left(\frac{1-\lambda}{2\lambda} - \frac{\sum_j (s_j - \bar{s})^2 + \sum_k (t_k - \bar{t})^2}{2\lambda^2 n^2} + O(n^{-b}) \right). \quad \square$$

Theorems 3.1 and 3.2 are proved by complex analysis, namely a multidimensional saddle-point calculation first demonstrated by McKay and Wormald [23] and McKay [25].

We will sketch the proof method for Theorem 3.1, based on [24]. Consider the n -variable generating function

$$F(\mathbf{z}) = \prod_{j,k \in \bar{\mathbf{X}}} (1 + z_j z_k),$$

where $\bar{\mathbf{X}}$ is the set of all unordered distinct pairs $\{j, k\}$ that are not edges of \mathbf{X} . This function counts n -vertex graphs disjoint from \mathbf{X} according to the degrees

of their vertices. Specifically,

$$|\mathcal{G}(\mathbf{d}, \mathbf{X})| = [z_1^{d_1} \cdots z_n^{d_n}]F(\mathbf{z}),$$

where the square bracket notation indicates coefficient extraction. By Cauchy’s theorem this implies

$$|\mathcal{G}(\mathbf{d}, \mathbf{X})| = \frac{1}{(2\pi i)^n} \oint \cdots \oint \frac{\prod_{j,k \in \bar{\mathbf{X}}} (1 + z_j z_k)}{z_1^{d_1+1} \cdots z_n^{d_n+1}} dz_1 \cdots dz_n,$$

where each integral is along a simple closed contour enclosing the origin anti-clockwise. Taking these contours to be circles, namely $z_j = r_j e^{i\theta_j}$ for each j , and changing variables gives

$$|\mathcal{G}(\mathbf{d}, \mathbf{X})| = \frac{\prod_{j,k \in \bar{\mathbf{X}}} (1 + r_j r_k)}{(2\pi)^n \prod_{j=1}^n r_j^{d_j}} \int_{-\pi}^{\pi} \cdots \int_{-\pi}^{\pi} \frac{\prod_{j,k \in \bar{\mathbf{X}}} (1 + \lambda_{jk} (e^{i(\theta_j + \theta_k)} - 1))}{\exp(i \sum_{j=1}^n d_j \theta_j)} d\boldsymbol{\theta}, \tag{4}$$

where $\boldsymbol{\theta} = (\theta_1, \dots, \theta_n)$ and

$$\lambda_{jk} = \frac{r_j r_k}{1 + r_j r_k} \quad (1 \leq j, k \leq n). \tag{5}$$

Equation (4) is valid for any positive radii $\{r_j\}$, but to facilitate estimation of the integral we choose $\{r_j\}$ so that the linear terms vanish when the logarithm of the integrand is expanded around the origin (in $\boldsymbol{\theta}$ space). This happens when

$$\sum_{k:jk \in \bar{\mathbf{X}}} \lambda_{jk} = d_j \quad (1 \leq j \leq n). \tag{6}$$

Equations 5 and 6 have a unique solution which appears to have no closed form. Under the conditions of Theorem 3.1, the solution can be expanded to sufficient accuracy in terms of λ , \mathbf{d} and \mathbf{X} . This involves summation over small subgraphs of \mathbf{X} and the expression is rather complex.

The integrand of (4) achieves its maximum modulus 1 at $\boldsymbol{\theta} = (0, 0, \dots, 0)$ and $\boldsymbol{\theta} = (\pi, \pi, \dots, \pi)$, which two points are equivalent under the symmetries of the integrand. We now define two small cubes:

$$\mathcal{R}_0 = \{\boldsymbol{\theta} : |\theta_j| \leq n^{-1/2+\epsilon}, 1 \leq j \leq n\}, \quad \mathcal{R}_\pi = \{\boldsymbol{\theta} : |\theta_j - \pi| \leq n^{-1/2+\epsilon}, 1 \leq j \leq n\},$$

where absolute value is taken modulo 2π . Within $\mathcal{R}_0 \cup \mathcal{R}_\pi$ we expand the logarithm of the integrand up to terms of order 4 and evaluate the integral by first diagonalising the quadratic part (recall that we chose the radii to eliminate the linear part) then integrating term by term. Outside $\mathcal{R}_0 \cup \mathcal{R}_\pi$ we split the region up into many pieces and show that in total the contribution to the integral is negligible.

In the case of empty \mathbf{X} , Barvinok and Hartigan [13] have identified the matrix (λ_{jk}) , with zero diagonal, as the unique symmetric matrix which satisfies (6) and maximises the entropy function

$$-\sum_{jk}(\lambda_{jk} \log \lambda_{jk} + (1 - \lambda_{jk}) \log(1 - \lambda_{jk})).$$

For the case of $\bar{d} = \Theta(n)$, they then show that an asymptotic approximation of $|\mathcal{G}(\mathbf{d})|$ can be expressed as a computable function of (λ_{jk}) whenever the values of λ_{jk} are uniformly bounded away from 0 and 1. This allows for a much larger variation amongst the degrees than Theorem 3.1 allows, but at the expense of more restricted \bar{d} and loss of explicitness. It is also shown in [13] under the same conditions that for a set S of $\Theta(n^2)$ edge-positions, a random graph in $\mathcal{G}(\mathbf{d})$ has close to $\sum_{jk \in S} \lambda_{jk}$ edges within S , with high probability.

A result similar to Theorem 3.2 for tournaments was proved by Gao, McKay and Wang [6].

4. Concluding Remarks

It is clear that many gaps still remain in our understanding of this problem. For example, there is almost nothing known about the distribution of subgraph counts in $\mathcal{G}(\mathbf{d})$ or $\mathcal{B}(\mathbf{s}, \mathbf{t})$ when the degrees are $\Theta(n)$. In the intermediate range of degrees between $n^{1/2}$ and $n/\log n$, not even the precise value of $|\mathcal{G}(\mathbf{d})|$ is known, though there is a well-tested conjecture [27]. The same is true in the bipartite case [4]. Another missing story is that of $\mathcal{B}(\mathbf{s}, \mathbf{t})$ when m and n are very different.

References

- [1] S. Ben-Shimon and M. Krivelevich, Random regular graphs of non-constant degree: concentration of the chromatic number, *Discrete Math.*, **309** (2009)x, 4149–4161.
- [2] B. Bollobás, A probabilistic proof of an asymptotic formula for the number of labelled regular graphs, *Europ. J. Combin.*, **1** (1980)x, 311–316.
- [3] B. Bollobás and B. D. McKay, The number of matchings in random regular graphs and bipartite graphs, *J. Combin. Th. Ser. B*, **41** (1986) 80–91.
- [4] E. R. Canfield, C. Greenhill and B. D. McKay, Asymptotic enumeration of dense 0-1 matrices with specified line sums, *J. Combin. Th. Ser. A*, **115** (2008) 32–66.
- [5] C. Cooper, A. Frieze and B. Reed, Random regular graphs of non-constant degree: connectivity and hamiltonicity, *Combin. Prob. Comput.*, **11** (2002) 249–261.
- [6] Z. Gao, B. D. McKay and X. Wang, Asymptotic enumeration of tournaments with a given score sequence containing a specified digraph, *Random Structures Algorithms*, **16** (2000) 47–57.

-
- [7] Z. Gao and N. C. Wormald, Asymptotic normality determined by high moments, and submap counts of random maps, *Probab. Theory Related Fields*, **130** (2004) 368–376.
- [8] Z. Gao and N. C. Wormald, Distribution of subgraphs of random regular graphs, *Random Structures Algorithms*, **32** (2008) 38–48.
- [9] H. Garmo, The asymptotic distribution of long cycles in random regular graphs. *Random Structures Algorithms*, **15** (1999) 43–92.
- [10] C. D. Godsil and B. D. McKay, Asymptotic enumeration of Latin rectangles, *J. Combin. Th. Ser. B*, **48** (1990) 19–44.
- [11] C. Greenhill and B. D. McKay, Random dense bipartite graphs and directed graphs with specified degrees, *Random Structures Algorithms*, to appear.
- [12] C. Greenhill, B. D. McKay and X. Wang, Asymptotic enumeration of sparse irregular bipartite graphs, *J. Combin. Th. Ser. A*, **113** (2006) 291–324.
- [13] A. Barvinok and J. A. Hartigan, The number of graphs and a random graph with a given degree sequence, preprint (2010), <http://arxiv.org/abs/1003.0356>.
- [14] S. Janson, The probability that a random multigraph is simple, *Combin. Prob. Comput.*, **18** (2009) 205–225.
- [15] S. Janson, T. Łuczak and A. Ruciński, *Random Graphs*, Wiley, 2000.
- [16] M. Krivelevich, B. Sudakov and N. C. Wormald, Regular induced subgraphs of a random graph, *Random Structures Algorithms*, to appear.
- [17] M. Krivelevich, B. Sudakov, V. Vu and N. C. Wormald, Random regular graphs of high degree, *Europ. J. Combin.*, **18** (2001) 346–363.
- [18] P. Lieby, B. D. McKay, J. C. McLeod and I. M. Wanless, Subgraphs of random k -edge-coloured k -regular graphs, *Combin. Probab. Comput.*, **18** (2009) 533–549.
- [19] B. D. McKay, Spanning trees in random regular graphs, *Third Caribbean Conference on Combinatorics and Computing*, (University of West Indies, 1981) 139–143.
- [20] B. D. McKay, Subgraphs of random graphs with specified degrees, *Congr. Numer.*, **33** (1981) 213–223.
- [21] B. D. McKay, Asymptotics for 0-1 matrices with prescribed line sums, in *Enumeration and Design*, (Academic Press, 1984) 225–238.
- [22] B. D. McKay, Asymptotics for symmetric 0–1 matrices with prescribed row sums, *Ars Combin.*, **19A** (1985) 15–26.
- [23] B. D. McKay, The asymptotic numbers of regular tournaments, eulerian digraphs and eulerian oriented graphs, *Combinatorica*, **10** (1990) 367–377.
- [24] B. D. McKay, Subgraphs of dense random graphs with specified degrees, submitted. <http://arxiv.org/abs/1002.3018>.
- [25] B. D. McKay and N. C. Wormald, Asymptotic enumeration by degree sequence of graphs of high degree, *European J. Combin.*, **11** (1990) 565–580.
- [26] B. D. McKay and N. C. Wormald, Uniform generation of random regular graphs of moderate degree, *J. Algorithms*, **11** (1990) 52–67.

-
- [27] B. D. McKay and N. C. Wormald, Asymptotic enumeration by degree sequence of graphs with degrees $o(n^{1/2})$, *Combinatorica*, **11** (1991) 369–382.
 - [28] B. D. McKay, N. C. Wormald and B. Wysocka, Short cycles in random regular graphs, *Electron. J. Combin.*, **11** (2004) R66, 12 pages.
 - [29] J. C. McLeod, Asymptotic enumeration of k -edge-coloured k -regular graphs, *SIAM J. Discrete Math.*, to appear.
 - [30] R. W. Robinson and N. C. Wormald, Almost all cubic graphs are Hamiltonian. *Random Structures Algorithms*, **3** (1992) 117–126.
 - [31] R. W. Robinson and N. C. Wormald, Almost all regular graphs are Hamiltonian. *Random Structures Algorithms*, **5** (1994) 363–374.
 - [32] N. C. Wormald, The asymptotic distribution of short cycles in random regular graphs, *J. Combin. Theory Ser. B*, **31** (1981) 168–182.
 - [33] N. C. Wormald, Models of random regular graphs, in *Surveys in Combinatorics, 1999* (eds. J. D. Lamb and D. A. Preece), Cambridge University Press, 1999, 239–298.

Sparse Combinatorial Structures: Classification and Applications

Jaroslav Nešetřil* and Patrice Ossona de Mendez†

Abstract

We present results of the recent research on sparse graphs and finite structures in the context of contemporary combinatorics, graph theory, model theory and mathematical logic, complexity of algorithms and probability theory. The topics include: complexity of subgraph- and homomorphism- problems; model checking problems for first order formulas in special classes; property testing in sparse classes of structures. All these problems can be studied under the umbrella of classes of structures which are Nowhere Dense and in the context of Nowhere Dense – Somewhere Dense dichotomy. This dichotomy presents the classification of the general classes of structures which proves to be very robust and stable as it can be defined alternatively by most combinatorial extremal invariants as well as by algorithmic and logical terms. We give examples from logic, geometry and extremal graph theory. Finally we characterize the existence of all restricted dualities in terms of limit objects defined on the homomorphism order of graphs.

Mathematics Subject Classification (2010). Primary 0502; Secondary 05C75, 05C15, 05C83, 05C85, 03C13, 68Q19.

Keywords. Graphs, hypergraphs, structures, homomorphism, sparsity, model checking, bounded expansion, property testing, separators, complexity, structural combinatorics.

*Supported by grant 1M0545 of the Czech Ministry of Education.

Department of Applied Mathematics and Institute of Theoretical Computer Science (ITI), Charles University, Malostranské nám.25, 11800 Praha 1, Czech Republic.
E-mail: nesetril@kam.ms.mff.cuni.cz.

†Centre d'Analyse et de Mathématiques Sociales, CNRS, UMR 8557, 54 Bd Raspail, 75006 Paris, France. E-mail: pom@ehess.fr.

1. Introduction

In this paper we survey results of the recent research on sparse graphs, hypergraphs and finite structures in the context of some of the key areas of contemporary combinatorics, graph theory, model theory and mathematical logic, complexity of algorithms and probability theory. We list the following areas as related to this paper:

- universal and generic structures of model theory;
- Constraint Satisfaction Problems in the context of descriptive complexity;
- complexity of subgraph- and homomorphism- problems;
- existence of (homomorphism) dualities in the context of the homomorphism order;
- fast model checking in first order logic;
- subgraphs statistics and local convergence;
- the existence of sublinear separators;
- property testing in sparse classes of structures;
- polynomial on-line and game-colorings of graphs;
- validity of homomorphism preservation theorems.

Although these are very distinct areas it is often easy to see that in all of these problems we have to put some restrictions on the graphs to be considered: in the full generality for finite graphs the answers to most of our questions are known to be negative, or hopelessly hard. But often the answer tend to be negative even for graphs which have many edges what is usually described by the term “dense graphs”. In the context of this paper dense graphs are not only those having $O(n^2)$ edges but even those having $O(n^{1+\epsilon})$ edges (n is the number of vertices). Even such edge sizes do not guarantee positive answers to the above problems.

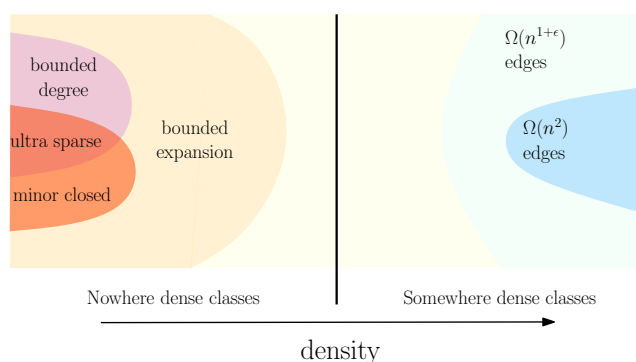
For such answers we have to look at instances with very few edges. For example subgraph problem has a positive answer for geometrically restricted (such as planar graphs [31, 30]) while the homomorphism preservation theorems hold for classes of bounded degree graphs ([10]). In contrast with this, the Separator Problem has the negative answer even for cubic (i.e. degree 3) graphs. In this case the answer is again positive for planar graphs [52], for graphs with a fixed genus [36], and for graphs excluding a minor [4, 3]. And similar diverse situations occur for the other problems and some sparsity is playing a role there.

But which structures are sparse? Sparsity seems to be an elusive and typically “fuzzy” notion and it seems that the answer to this question depends on the particular problem considered. Yet in this paper we present a classification of graph classes which clarifies the boundary between sparse and dense instances and which proved to be useful in many concrete applications and all of the above problems in particular.

How to define sparsity? Perhaps the good way to define it is by means of the stability with respect to some operations. We aim for sparsity as a self-similarity

idea, as the invariance to small changes. This approach is the one taken in this paper. We employ the mixture of geometric and combinatorial approach and define (time) resolution of a structure and of a class of structures. This in turn leads to the surprisingly general dichotomy of classes of structures — there are nowhere dense classes and classes which are somewhere dense. This is stated in Section 2.4 after the introduction in Section 2 of all the relevant notions.

In Section 4 we show how this dichotomy, which may appear on the first glance arbitrary, can be described in several very different ways. In fact *almost all* the basic extremal combinatorial parameters are suitable for the description of this dichotomy: In Section 3.1 we deal with edge densities, in Section 2.4 with clique number ω , in Section 4.1 with the chromatic number χ , in Section 4.2 with the independence number α (and of course for the space limitations we do not mention all relevant characterizations, see [74, 73]). All of this shows that the nowhere dense – somewhere dense dichotomy is not just an accident or a combinatorial curiosity but rather a natural, stable and robust dichotomy.



There is a further evidence which goes beyond the α, χ, ω . Very recently this list was complemented by the counting (densities of subgraphs) (see Section 4.3) and also by results in mathematical logic: the nowhere dense – somewhere dense dichotomy induces exactly the dividing line between (monotone) classes of graphs for which the model checking for first order logic is Fixed Parameter Tractable (FPT) and those classes for which model checking is hard, see [21, 25]. We treat this in Section 4.2 where (based on our earlier analysis of Nowhere Dense structures) we extend these result to general structures.

Some of these applications will be mentioned in Section 5 and Section 6 in a greater detail. The core of many of these applications is a possibility to approximate (with arbitrary precision) any graph in a nowhere dense class by a graph defined by finitely many data. Technically this takes form of *Low Tree Depth Decomposition* which for the illustration we formulate here for the case of a bounded expansion class \mathcal{C} of structures (defined in Section 8):

Theorem 1. (*Low Tree Depth Decomposition*) *For every bounded expansion class \mathcal{C} and for every positive integer p there exists an integer $N = N(p, \mathcal{C})$ such*

that every structure $\mathbf{A} \in \mathcal{C}$ has a decomposition $X^{\mathbf{A}} = X_1 \cup \dots \cup X_N$ with the following property:

\mathbf{A} restricted to any set $\bigcup_{i \in I} X_i$ where $|I| \leq p$, has tree depth at most $|I|$ (particularly, this substructure cannot include a path of length $2^{|I|}$).

(See Sections 4 and 8 for more details.) As there are only finitely many core graphs with tree depth at most p [65] a Low Tree Depth Decomposition can be approximated by a finite set of data and this decomposition is much in the spirit of Szemerédi regularity lemma, [88].

The research covered by this paper is related to the recent development which is based on the study of *homomorphisms* of structures. The main idea is to study the local structure of a large structure \mathbf{A} by counting the homomorphisms from various small graphs \mathbf{F} into \mathbf{A} (this relates to the area called *property testing*), and to study the global structure of \mathbf{A} by *counting* its homomorphisms into various small graphs \mathbf{C} (sometimes interpreted as *templates*). Regularity is viewed here as a structural approximation in a proper metric and also as a convergence. For a survey of this development, see [17]. This approach proved to be very fruitful and relates (among others) to the notion of *quasi-random* graphs, see e.g. [19], and to the general results characterizing testable properties, see e.g. [5, 17, 29, 82, 87, 11].

In this paper we take a similar, yet different, approach. We start our analysis with the homomorphism order. We shall see that in this setting, at a proper level of generality, some of the results for dense graphs can be extended to the world of sufficiently sparse classes of graphs. Along these lines we mention also results related to the universality problems for the homomorphism order (and we mention several results obtained jointly with Jan Hubička, [48, 46, 47]). This then naturally relates to the problems of finite dualities (which we characterized jointly with Claude Tardif [79]) and then to restricted dualities which will be characterized (in Section 9) by means of the completion of the homomorphism order. After all the existential theorem related to the homomorphism order we return to the counting and describe (in the case of graphs) the Nowhere Dense–Somewhere Dense dichotomy by means of the counting functions (see Section 4.3).

2. Preliminaries

2.1. Graphs vs Structures. Let us review some basic notions which will be used. Our graphs are finite simple undirected graphs, except when explicitly stated otherwise and we denote by *Graph* the class of all such graphs. We use standard graph theory terminology (see e.g. [59]). We find it useful to introduce the following: for a graph $G = (V, E)$, we denote by $|G|$ the *order* of G (that is: $|V|$) and by $\|G\|$ the *size* of G (that is: $|E|$). Similarly, a finite *set system* (or *hypergraph*, we shall use both notions) is a pair (X, \mathcal{M}) where \mathcal{M} is a collection of subsets of X . It is customary to call these sets *edges* again.

If all the edges have k elements then we speak simply about a k -graph. (Thus graphs are just 2-graphs.)

The *distance* in a graph G between two vertices x and y is the minimum length of a path linking x and y (or ∞ if x and y do not belong to the same connected component of G) and is denoted by $\text{dist}_G(x, y)$. Let $G = (V, E)$ be a graph and let d be an integer.

A class \mathcal{C} of graphs is *hereditary* if every induced subgraph of a graph in \mathcal{C} belongs to \mathcal{C} , and it is *monotone* if every subgraph of a graph in \mathcal{C} belongs to \mathcal{C} .

The notion of a finite relational structure is more involved and in fact it appears in two different formalisms.

One possibility is that we specify a language \mathcal{L} which accepts standard logic and sets involves relational symbols R, S, \dots each with an appropriate arity. In such case we speak about relational structures with a given signature \mathcal{L} .

Sometimes we want to be more explicit and we specify a finite sequence (*type*) of positive integers $\Delta = (\delta_i : i \in I)$ which we call *type*. A relational structure \mathbf{A} of type Δ is then a pair $(X, (R_i : i \in I))$ where $R_i \subset X^{\delta_i}$ is an d_i -ary relation on X . In this case we also put $X = X^{\mathbf{A}}$ and $R_i = R_i^{\mathbf{A}}$.

The notions of a *homomorphism* (*monomorphism*) are standardly defined as mappings (injective mappings) preserving all relations. In a difference to algebras the embeddings need a little more care: An injective mapping $f : X^{\mathbf{A}} \rightarrow X^{\mathbf{B}}$ is called an *embedding* of \mathbf{A} to \mathbf{B} if the following holds for every relation $R^{\mathbf{A}}$:

$$(x_i : i = 1, \dots, \ell) \in R^{\mathbf{A}} \quad \text{if and only if} \quad (f(x_i) : i = 1, \dots, \ell) \in R^{\mathbf{B}}.$$

The category of all finite graphs and all homomorphisms between them will be denoted again by *Graph*, the category of all finite relational structures of type Δ and all homomorphisms between them $\mathcal{Rel}(\Delta)$ and the category of all finite set systems (i.e. hypergraphs) or k -graphs and all their homomorphisms is denoted by *Hyp* or *Hyp*(k).

The relationship of very simple models (as presented by graphs) and more general relational systems is very interesting and far from trivial. Recently this connection got several new impulses. For example the connection to Constraint Satisfaction Problems, first order definability and to descriptive algorithmic complexity recently were intensively studied [51, 33, 2, 45]. We report some of this research in Section 4.2.

There are various connection between relational structures of different types. For example if the two signatures are in inclusion (i.e. if $\Delta \subset \Delta'$) then we speak about an *extension* (sometimes the name *lift* is used). This corresponds to enrichment of the original structures $\mathbf{A} \in \mathcal{Rel}(\Delta)$ by new relations, such as colors of vertices, edges, orderings, etc. The inverse construction is the *reduct* (sometimes *shadow*): We start with an object $\mathbf{A}' \in \mathcal{Rel}(\Delta')$ and define the object $\mathbf{A} \in \mathcal{Rel}(\Delta)$ by considering only those relations from type Δ (and forgetting about the others), see e.g. [44].

The reduct and extension are powerful operations (as expected; we are changing the language itself). For example, it has been proved in [51] (extending earlier works by [32, 33]) that the question whether there exists a lift with finitely many prescribed local properties is polynomially equivalent to a general problem in the class NP. Similarly reducts are related to some of the classical combinatorial (Ramsey type) statements, see [90] (using [78]).

There are other constructions which reduce one signature to another. In the logical context perhaps the most widely used is the following construction:

1. Gaifman Graph. To a relational structure \mathbf{A} we associate its *Gaifman graph* $\text{Gf}(\mathbf{A}) = (\mathbf{V}, \mathbf{E})$ by putting $V = X^{\mathbf{A}}$ and $\{x, y\} \in E$ if $x \neq y$ and x and y appear in a same tuple of \mathbf{A} . (In combinatorics this construction is known as 2-section, [15].)

Gaifman graphs allow us to translate many graph notions to general systems. This transformation has several advantages (the main one being perhaps its simplicity) but there are disadvantages too and the sparsity (which is our central theme here) is often not preserved. The relational system may be quite sparse and of a very simple form, yet the corresponding Gaifman graph may be as complicated as possible. For example the Gaifman graph of any edge is a complete graph. Other examples (with bounded arities) include Gaifman graph of any Steiner Triple System (X, \mathcal{M}) which is the complete graph on the set X . An even simpler example can be constructed on any set $X \cup \{*\}$ as follows: \mathcal{M} is formed by all triples of form $\{x, x', *\}, x \neq x' \in X$. The Gaifman graph is again a complete graph.

As a result of examples like these, more sensitive transformations were devised:

2. Block graph. The *block graph* $\text{Inc}(\mathbf{A})$ of \mathbf{A} (sometimes called *Incidence graph*) is defined as follows: The vertices are formed by $X^{\mathbf{A}}$ together with the set of all pairs $(i, (x_1, \dots, x_{\delta_i})), i \in I$. The edges are formed by all incidences between x and $(x_1, \dots, x_{\delta_i})$. This construction has many forms: edges may be directed, vertices may have colors, it is either simple graph or a multigraph, etc.

3. Path graph. The path graph $\text{Path}(\mathbf{A})$ of \mathbf{A} is defined as follows: Vertices are $X^{\mathbf{A}}$ with a tuple $(x_1, \dots, x_{\delta_i}) \in R^{\mathbf{A}}$ being replaced by a directed path $x_1 \rightarrow x_2 \rightarrow \dots \rightarrow x_{\delta_i}$ (thus in this for this is a directed graph). For example this is used in [79] to classify the dualities.

4. Star selectors. We formulate this for a hypergraph (X, \mathcal{M}) : For an edge $M \in \mathcal{M}$ a *star selector* is any star with the vertex set M . A *star selector* of a hypergraph (X, \mathcal{M}) is then the union of the edge sets of star selectors of the edges of \mathcal{M} . This is not uniquely defined as we may get several graphs (with very different properties). But perhaps because of this flexibility this is often the best transformation. Let us give an example:

For a hypergraph $\mathbf{A} \in \mathcal{C}$, we denote by $\text{Sel}(\mathbf{A})$ the set of the star selectors of \mathbf{A} and by $\mathfrak{Sel}(\mathcal{C})$ we denote the set of all the function $\zeta : \mathcal{C} \rightarrow \text{Graph}$ such that $\zeta(\mathbf{A}) \in \text{Sel}(\mathbf{A})$. We have then for example the following result (see Section 2.3 for the definition of $G \nabla d$):

Theorem 2. *Let \mathcal{C} be a class of hypergraphs. Then the following conditions are equivalent:*

1. $\forall d \in \mathbb{N} \exists \zeta_d \in \mathfrak{Sel}(\mathcal{C}) : \sup_{\mathbf{A} \in \mathcal{C}} \omega(\zeta_d(\mathbf{A}) \nabla d) < \infty;$
2. \mathcal{C} is a S -nowhere dense class of hypergraphs, meaning that there exists $\zeta \in \mathfrak{Sel}(\mathcal{C})$ such that $\zeta(\mathcal{C})$ is a nowhere dense class of graphs;
3. $\exists \zeta \in \mathfrak{Sel}(\mathcal{C}) \forall d \in \mathbb{N} : \sup_{\mathbf{A} \in \mathcal{C}} \omega(\zeta(\mathbf{A}) \nabla d) < \infty.$

We use all of these constructions according to what is most fitting for a particular result. In the context of sparse hierarchies of structures, the relationship of various models of relational structures is not yet clarified. Perhaps the situation is reminiscent to a long development of the Szemerédi regularity lemma for hypergraphs and finite structures, see e.g. [57, 89, 39, 83].

Yet another direction to extend the results for graphs to more general structures is to consider an edge version of low tree depth decomposition. This naturally generalizes to matroids [76].

2.2. Homomorphism order. The central role (and indeed the leitmotiv) in our paper is played by the simplification of the above categories. This takes the following form: Given structures \mathbf{A}, \mathbf{B} we write $\mathbf{A} \leq \mathbf{B}$ to denote the existence of a homomorphism $f : \mathbf{A} \rightarrow \mathbf{B}$. The relation \leq is clearly a quasiorder on $\text{Rel}(\Delta)$ or Graph or Hyp . The relation \leq is called *homomorphism order* which will be indicated as $(\text{Graph}, \leq), (\text{Rel}(\Delta), \leq), (\text{Hyp}, \leq)$. The homomorphism order can be reduced to a partial order in two steps:

- First, we consider cores of all structures. A *core* is any minimal retract of a structure (this term was coined in [43]);
- Then, we consider the isomorphism types of core structures. If a more precision is needed then we denote by $[\mathbf{A}]$ the isomorphism type determined by the structure \mathbf{A} .

In most of the paper there is no danger of confusion and thus we also denote by $(\text{Graph}, \leq), (\text{Rel}(\Delta), \leq), (\text{Hyp}, \leq)$ the corresponding partial orders of isomorphism types of the corresponding core structures.

The homomorphism order has spectacular properties, some of which will be reviewed here:

Theorem 3. *(Universality of the homomorphism order) For every countable partial order P there is an embedding of P into (Graph, \leq) . Not only that, but a much smaller variety of graphs suffices: For every countable partial order P there is an embedding of P into the suborder of (Graph, \leq) induced by planar graphs with all degrees bounded by 3.*

This is a classical result proved in [42]. The second part is much more recent and it presented a well known problem, see e.g. [80], which was finally proved by Jan Hubička and J. Nešetřil [47, 46].

Theorem 4. (*Density of the homomorphism order*) With “a few exceptions” the homomorphism order is dense. Explicitly, for most pairs \mathbf{A}, \mathbf{B} with $\mathbf{A} \leq \mathbf{B}$ and $\mathbf{B} \not\leq \mathbf{A}$ (with “a few exceptions”, there exists \mathbf{C} such that $\mathbf{A} \leq \mathbf{C} \leq \mathbf{B}$ and $\mathbf{B} \not\leq \mathbf{C} \not\leq \mathbf{A}$).

This result is proved in [79] (extending earlier result of [91] for undirected graphs). In fact, again with “a few exceptions”, every interval in the homomorphism order is itself universal (an unpublished result; see [61]).

What are “few exceptions”? They are important and they are completely characterized. Basically the only exceptions to the density are induced by trees. More explicitly, a pair (\mathbf{A}, \mathbf{B}) of structures is called a *gap* in the homomorphism order if $\mathbf{A} \leq \mathbf{B}, \mathbf{B} \not\leq \mathbf{A}$ and there is no \mathbf{C} strictly in between \mathbf{A} and \mathbf{B} . One of the main results of [79] is that all gaps $(\mathbf{A}, \mathbf{B}), \mathbf{B}$ connected, in $\mathcal{R}el(\Delta)$ are induced by trees. Explicitly, for every relational tree \mathbf{T} there exists (uniquely determined) *predecessor* structure $P(\mathbf{T})$ such that the pair $P(\mathbf{T}) \leq \mathbf{T}$ forms a gap. (Other gaps are not connected and they are also related to trees [79].) What is a relational tree? We can use above reductions. A relational tree is a structure such that its path graphs (as above) is (an orientation of) a tree.

The homomorphism order has the rich algebraic structure. There is a beautiful (and surprising) connection of gaps to the dual description of graph classes which goes under name *homomorphism dualities* (defined in [77]). This can be outlined as follows (compare [43]):

A *singleton duality* is a pair of objects (\mathbf{F}, \mathbf{D}) with the following property: For every object \mathbf{A} of the same type as (\mathbf{F}, \mathbf{D}) holds:

$$\mathbf{A} \longrightarrow \mathbf{D} \quad \text{if and only if} \quad \mathbf{F} \not\rightarrow \mathbf{A}.$$

For undirected graphs there are just two trivial dualities. However already for oriented graphs we have infinitely many dualities and these dualities are important as they relate to the chromatic number of graphs (by means of Gallai – Hasse – Roy – Vitaver theorem, [71]). The notion of duality is motivated by algorithmic considerations and particularly the dual description of homomorphisms into a fixed template \mathbf{D} by means of a simple obstacle \mathbf{F} . It is the more than surprising that this simple notion is in one to one correspondence with the purely order-theoretic notion of a gap:

Theorem 5. (*Gaps and Dualities [79]*) There is one to one correspondence between singleton dualities (\mathbf{F}, \mathbf{D}) and gap-pairs $(P(\mathbf{A}), \mathbf{A})$ where \mathbf{A} is a connected structure.

In fact this characterization of dualities is of a categorical nature and it can be extended to the much more general situation of Heyting algebras [35].

All these dualities are class dependent. They hold in the class of structures with a fixed signature (i.e. in classes $\mathcal{Rel}(\Delta)$). This is also clear from yet another reformulation of the duality pair (\mathbf{F}, \mathbf{D}) . First, let us define the class $\text{Forb}(\mathbf{F})$ as the class of all structures (of a given signature) \mathbf{A} for which there is no homomorphism $\mathbf{F} \rightarrow \mathbf{A}$:

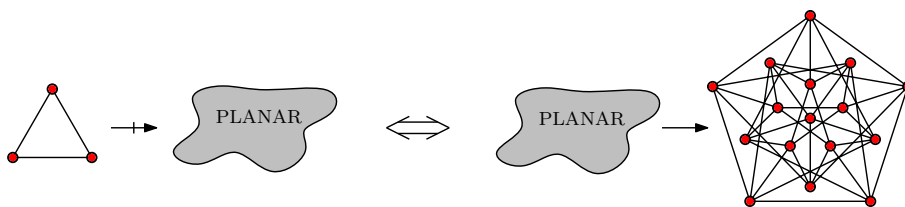
$$\text{Forb}(\mathbf{F}) = \{\mathbf{A} : \mathbf{F} \not\rightarrow \mathbf{A}\}.$$

(Similarly we define $\text{Forb}(\mathcal{F})$ for a finite set \mathcal{F} of structures.) Now (\mathbf{F}, \mathbf{D}) is a duality pair if and only if the object \mathbf{D} is the (finite) maximum of the class $\text{Forb}(\mathbf{F})$ in the homomorphism order. This point of view is taken in [63] and it allows to treat universal and generic structures together with dualities, [51, 45]. In yet another interpretation every finite duality is the equation of two classes $\text{Forb}(\mathcal{F})$ (for a finite set \mathcal{F}) and $\text{CSP}(\mathbf{D})$ defined as a principal ideal in the homomorphism order:

$$\text{CSP}(\mathbf{D}) = \{\mathbf{A} : \mathbf{A} \rightarrow \mathbf{D}\}.$$

Let us note by passing that finite dualities (i.e. equations $\text{Forb}(\mathcal{F}) = \text{CSP}(\mathbf{D})$) are exactly those Constraint Satisfaction Problems (i.e. membership problems for classes $\text{CSP}(\mathbf{D})$, in this setting \mathbf{D} is usually called *template*) which are first order definable: Only for duals of tree structures is the corresponding Constraint Satisfaction Problem decidable by a first order formula. This follows from [79] and [9].

In Section 9 we define more general notion of *restricted dualities* and prove that our sparse classes have *all restricted dualities*. We then go on by characterizing this phenomenon (see Theorem 18). Advancing this we include the schematic Figure which holds for any planar graph G :



This was (in retrospect) one of our motivating examples [62, 63]. The homomorphism orders are fascinating structures with a rich algebraic and combinatorial contents.

2.3. Sparsity via Resolution in Time. Let us start this section by considering undirected graphs.

As remarked earlier, the notion of a sparsity of graphs is a fuzzy notion. First it does not relate to any particular graph but rather to a set, or sequence, or a class of graphs. Secondly the notion should be (certainly from naive point

of view) be invariant to some small changes of a graph. Third, to be a sparse graphs is clearly a global property and the property should be hereditary.

Combining these observations and motivated by numerous particular cases we are led to the following definitions:

For any graphs H and G and any integer d , the graph H is said to be a *shallow minor* of G at *depth* d ([81] attribute this notion, called then *low depth minor* to Ch. Leiserson and S. Toledo) if there exists a subset $\{x_1, \dots, x_p\}$ of G and a collection of disjoint subsets V_1, \dots, V_p of vertices of G , each inducing a connected subgraph of G , such that $x_i \in V_i$, every vertex in V_i is at distance at most d from x_i in the subgraph of G induced by V_i , and so that H is a subgraph of the graph obtained from G by contracting each V_i into x_i and removing loops and multiple edges (see Fig. 1).

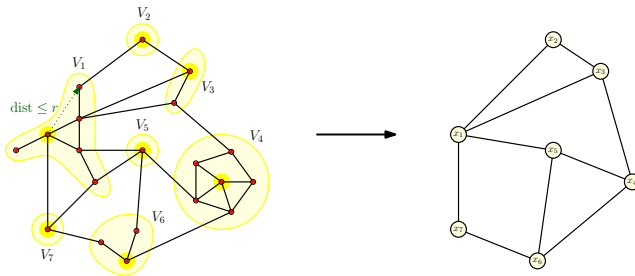


Figure 1. A shallow minor of depth r of a graph G is a simple subgraph of a minor of G obtained by contracting vertex disjoint subgraphs with radius at most r

The set of all shallow minors of G at depth d is denoted by $G \nabla d$. In particular, $G \nabla 0$ is the set of all subgraphs of G . Hence we have the following non decreasing sequence of classes (which we interpret as evolving in time):

$$G \in G \nabla 0 \subseteq G \nabla 1 \subseteq \dots \subseteq G \nabla d \subseteq \dots G \nabla \infty.$$

We extend this definition to arbitrary class of graphs \mathcal{C} by:

$$\mathcal{C} \nabla d = \bigcup_{G \in \mathcal{C}} G \nabla d.$$

We have the following (time dependent) hierarchy of classes

$$\mathcal{C} \subseteq \mathcal{C} \nabla 0 \subseteq \mathcal{C} \nabla 1 \subseteq \dots \subseteq \mathcal{C} \nabla d \subseteq \dots \mathcal{C} \nabla \infty.$$

We call this sequence *minor resolution* of the class \mathcal{C} and denote it by \mathcal{C}^∇ . Note that $\mathcal{C} \nabla 0$ is the monotone closure of \mathcal{C} and that $\mathcal{C} \nabla \infty$ is the minor closed class generated by \mathcal{C} .

2.4. The Nowhere Dense – Somewhere Dense Dichotomy.

The minor resolution of a class naturally leads to a classification of general

classes and to their interesting properties. The following are the key definitions of this paper:

Definition 1. (The Nowhere Dense – Somewhere Dense Dichotomy) An infinite class of graphs \mathcal{C} is *somewhere dense* if there exists an integer d such that $\mathcal{C} \nabla d = \text{Graph}$. Thus \mathcal{C} is somewhere dense if every graph is a bounded depth shallow minor of a graph in \mathcal{C} . In other words: we get all graphs in a fixed time.

If an infinite class is not somewhere dense, it is *nowhere dense*.

It follows directly from the definition of the minor resolution that a class \mathcal{C} is nowhere dense if and only if for every d the supremum of $\omega(G)$ for $G \in \mathcal{C} \nabla d$ is finite (here $\omega(G)$ is the the *clique number* of graph G , i.e. the maximal order of a complete graph in G). (It is perhaps surprising, as we shall see in Section 4.2, that nowhere dense classes may be defined by their independence number as well.)

For relational structures and hypergraphs we can define analogous notions.

Definition 2. (The Nowhere Dense - Somewhere Dense Dichotomy via Gaifman) An infinite class of structures \mathcal{C} is *G-somewhere dense* if the class $\text{Gf}(\mathcal{C})$ of all Gaifman graphs of structures in \mathcal{C} is somewhere dense. In other words: \mathcal{C} is somewhere dense if every graph is a bounded depth shallow minor of the Gaifman graph $\text{Gf}(\mathbf{A})$ of a structure $\mathbf{A} \in \mathcal{C}$.

If an infinite class is not G-somewhere dense, it is *G-nowhere dense*.

Definition 3. (The Nowhere Dense – Somewhere Dense Dichotomy via Incidence) An infinite class of structures \mathcal{C} is *I-somewhere dense* if the class $\text{Inc}(\mathcal{C})$ of all incidence graphs of structures in \mathcal{C} is somewhere dense. In other words: \mathcal{C} is somewhere dense if every graph is a bounded depth shallow minor of the incidence graph $\text{Inc}(\mathbf{A})$ of a structure $\mathbf{A} \in \mathcal{C}$.

If an infinite class is not I-somewhere dense, it is *I-nowhere dense*.

For path-graphs and star selectors we first observe that our resolutions are defined by means of distances and this “symmetric” neighborhoods. After that we define the dichotomy for these two constructions as well (*P-somewhere dense/P-nowhere dense* and *S-somewhere dense/S-nowhere dense*). Although in many instances are these approaches equivalent in general they differ and it is convenient to use all these definitions simultaneously.

3. Trichotomy for Binary Structures

We consider graph models in this section. For general structures the situation is more complicated and although we get analogous results we need stronger results (particularly the subgraph counting presented in Section 4.3).

3.1. Classification by Edge Densities. Let \mathcal{C} be an infinite class of graphs and let $f : \mathcal{C} \rightarrow \mathbb{R}$ be a graph invariant. Let $\text{Inj}(\mathbb{N}, \mathcal{C})$ be the set of all injective mappings from \mathbb{N} to \mathcal{C} . Then we define:

$$\limsup_{G \in \mathcal{C}} f(G) = \sup_{\phi \in \text{Inj}(\mathbb{N}, \mathcal{C})} \limsup_{i \rightarrow \infty} f(\phi(i))$$

Notice that $\limsup_{G \in \mathcal{C}} f(G)$ always exist and is either a real number or $\pm\infty$.

Theorem 6 (Trichotomy theorem). *Let \mathcal{C} be an infinite class of graphs (asymptotically not all edgeless). Then the limit*

$$\ell\text{dens}(\mathcal{C}^\nabla) = \lim_{i \rightarrow \infty} \limsup_{G \in \mathcal{C}^\nabla_i} \frac{\log \|G\|}{\log |G|}$$

may take only three values, namely 0, 1 and 2. Moreover, we have:

$$\ell\text{dens}(\mathcal{C}^\nabla) = \begin{cases} 0, & \text{iff } \sup_{G \in \mathcal{C}} \|G\| < \infty, \\ 0 \text{ or } 1, & \text{iff } \mathcal{C} \text{ is nowhere dense,} \\ 2, & \text{iff } \mathcal{C} \text{ is somewhere dense.} \end{cases}$$

For a proof see [74]. It can be seen easily that $\ell\text{dens}(\mathcal{C}^\nabla) \leq 0$ if and only if the class \mathcal{C} contains only graphs with at most k_0 edges. These *essentially finite* classes can be non-trivial. A prime example is the class of all core graphs with tree depth bounded (see Section 4.1 for the definition of the tree depth).

It is very interesting (and we feel surprising) that this theorem has a topological version which counts the edges in shallow subdivisions. (Recall that a graph G' is a *subdivision* of a graph G if G' arises from G by adding vertices (of degree 2) on edges of G .) Thus in the topological sense we have homeomorphic graphs: all edges of G are replaced by simple openly disjoint paths. If all these paths have length $\leq 2d + 1$ we say that G' is a d -shallow subdivision of G . Conversely, we say that H is *topological shallow minor at depth d* of a graph G if there exists a subgraph H' of G such that H' is a shallow subdivision of H at depth d . Having defined this we can proceed similarly as for the shallow minors and define the notion of topological minor resolution. For a proof of the topological version of Theorem 6 see [74]. (This extends work of Zdeněk Dvořák [22, 23].)

Also, the property that there exists a critical value $\tilde{\tau}(\mathcal{C})$ at which the topological resolution stabilizes to *Graph* is equivalent to the existence of a critical value $\tau(\mathcal{C})$ at which the minor resolution stabilizes to *Graph*. Notice that, according to Theorem 6, the existence of a critical value $\tilde{\tau}(\mathcal{C})$ is equivalent to the existence of a value $T(\mathcal{C})$ such that there exists $\epsilon > 0$ with

$$\limsup_{G \in \mathcal{C}^\nabla_{T(\mathcal{C})}} \frac{\log \|G\|}{\log |G|} = 1 + \epsilon.$$

Moreover, the difference between $\tilde{\tau}(\mathcal{C})$ et $T(\mathcal{C})$ is actually bounded by a function of ϵ (see Fig .2).

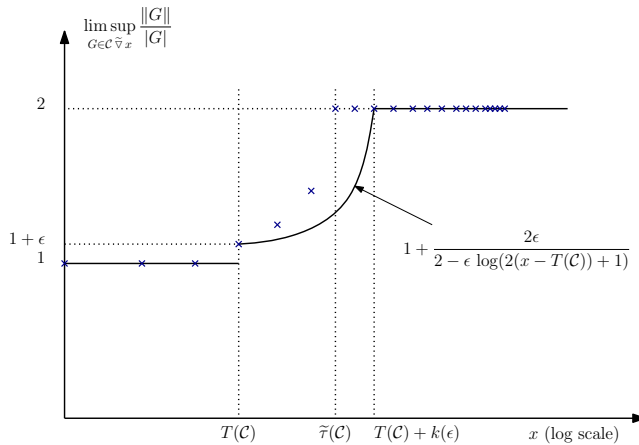


Figure 2. Evolution of the upper logarithmic density $\ell\text{dens}(\mathcal{C}^{\nabla})$ of the topological resolution of a typical somewhere dense class \mathcal{C}

However, for nowhere dense classes, the asymptotic behavior of the resolution varies. For instance, the class \mathcal{D}_3 of graphs with maximum degree at most 3 is such that $\mathcal{D}_3 \nabla \infty = \text{Graph}$ but $\mathcal{D}_3 \bar{\nabla} \infty = \mathcal{D}_3$.

Why do we state this topological variant of shallow minors, when we then claim just analogous results? The main reason is that this connection is surprising and non-trivial. The fact that minors and topological minors lead to the same classification of classes is interesting in the context of graph-minor theory where minors and topological minors lead often to very different results (as demonstrated for example by Hajós and Hadwiger’s conjectures), see [74, 69] for more details.

4. Some Alternative Characterizations

We mention just 3 characterizations. Yet they should indicate the robustness of our dichotomy classification of classes.

4.1. Classification by Decomposition — Chromatic Numbers.

First we consider the graph case. The building blocks of our decompositions will be induced by trees.

A *rooted forest* is a disjoint union of rooted trees. The *height* of a vertex x in a rooted forest F is the number of vertices of the path from the root (of the tree to which x belongs to) to x and is noted $\text{height}(x, F)$. The *height* of F is the maximum height of the vertices of F . Let x, y be vertices of F . The vertex x is an *ancestor* of y in F if x belongs to the path linking y and the root of the tree of F to which y belongs to. The *closure* $\text{clos}(F)$ of a rooted forest F is the graph

with vertex set $V(F)$ and edge set $\{\{x, y\} : x \text{ is an ancestor of } y \text{ in } F, x \neq y\}$. A rooted forest F defines a partial order on its set of vertices: $x \leq_F y$ if x is an ancestor of y in F . The comparability graph of this partial order is obviously $\text{clos}(F)$.

The *tree-depth* $\text{td}(G)$ of a graph G is the minimum height of a rooted forest F such that $G \subseteq \text{clos}(F)$ [65] (see Fig 3).

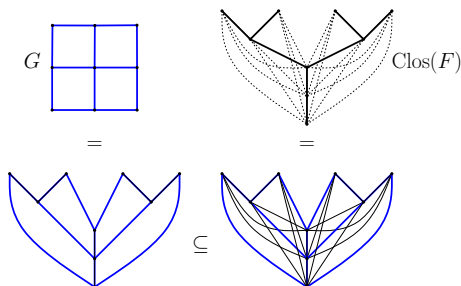


Figure 3. The tree-depth of the 3×3 grid is 4.

A principal property of the class of all graphs with $\text{td}(G) \leq k$ is that this class is finite when restricted to core graphs (or core structures). This holds more generally for colored graphs and for relational structures in general. This has also a number of consequences. For example the class of all graphs with $\text{td}(G) \leq k$ is well quasi ordered with respect to induced subgraph ordering. Nevertheless one should remark that the number of core graphs with $\text{td}(G) \leq k$ has an Ackermann growth.

In [65] we introduced the following parametrized generalization of the chromatic number: for any integer p , $\chi_p(G)$ denotes the minimum number of colors one shall use to color the vertices of G in such a way that for every subset I of at most p colors, the subgraph G_I of G induced by the vertices with color in I has tree-depth at most $|I|$. Thus χ_1 is the usual chromatic number of a graph (i.e. no edge is monochromatic) and χ_2 is minimal coloring with the property that no path with 4 vertices gets less than 3 colors.

These generalized chromatic numbers characterize nowhere dense classes ([67, 74]):

Theorem 7. *Let \mathcal{C} be an infinite class of graphs. Then the following conditions are equivalent:*

- \mathcal{C} is nowhere dense,
- for every integer p , $\limsup_{G \in \mathcal{C}} \frac{\log \chi_p(G)}{\log |G|} = 0$

Thus any graph G in a (fixed) nowhere dense class \mathcal{C} can be decomposed into a small number of classes such that the subgraphs induced by any $\leq p$ classes of the partition have components of only finitely many (homomorphism)

types. Thus p is then parameter expressing the precision of such decomposition. Moreover such a decomposition can be found in almost linear number of steps. This has a number of algorithmic consequences which are not covered here, see ([64, 67]. Such a decomposition is called *Low Tree Depth Decomposition* (LTDD).

Let us return to structures. We formulate this time the result for G -nowhere dense classes. The

Theorem 8. *Let \mathcal{C} be an infinite class of structures. Then the following conditions are equivalent:*

- \mathcal{C} is G -nowhere dense,
- for every integer p , $\limsup_{G \in \text{Gf}(\mathcal{C})} \frac{\log \chi_p(G)}{\log |G|} = 0$

Of course we can define $\chi_p(\mathbf{A})$ directly and it has a similar meaning as for graphs.

4.2. Classification by Independence. The *homomorphism preservation theorem* [58] states that a first-order formula is preserved under homomorphisms on all structures (finite and infinite) if and only if it is equivalent to an existential-positive formula. Answering a long-standing question in finite model theory [34], Ben Rossman proved [85] that the homomorphism preservation theorem remains valid when restricted to finite structures (unlike many other classical preservation theorems, including the Łoś–Tarski theorem and Lyndon’s positivity theorem). It is interesting to note that one of the main tools of Rossman’s proof is the notion of tree depth (which corresponds to the quantifier depth). In the context of relativizations of this theorem to specific classes of structures Anuj Dawar [20] introduced the following notion of quasi-wideness:

Let $d \geq 1$ be an integer. A subset A of vertices of a graph G is *d -independent* if the distance between any two distinct elements of A is strictly greater than d . Note that if we denote by $\alpha_d(G)$ the maximum size of an d -independent set of G , then $\alpha_1(G)$ is the usual independence number $\alpha(G)$ of graph G .

A graph G is *quasi-wide* if there is a function $s : \mathbb{N} \rightarrow \mathbb{N}$ such that for every integers d and m , every sufficiently big graph $G \in \mathcal{C}$ (i.e. of order at least $F(d, m)$) contains a subset S of size at most $s = s(d)$ so that the graph $\alpha_d(G - S) \geq m$.

The quasi-wide property is not hereditary. Thus we introduce the following, stronger version:

A graph G is *uniformly quasi-wide* if there is a function $s : \mathbb{N} \rightarrow \mathbb{N}$ such that for every integers d and m , every sufficiently big subset A of vertices of a graph $G \in \mathcal{C}$ (i.e. such that $|A| \geq F(d, m)$) is such that G contains a subset S of size at most $s = s(d)$ so that $G - S$ contains a d -independent set of size at least m included in A .

It appears that uniform quasi-wideness is strongly related to our classification:

Theorem 9. *Let \mathcal{C} be an infinite class of graphs. Then the following conditions are equivalent:*

- \mathcal{C} is nowhere dense,
- the hereditary closure of \mathcal{C} is quasi-wide,
- \mathcal{C} is uniformly quasi-wide.

This is a non-trivial results with several consequences, see [70]. Combined with Low Tree Depth Decomposition one deduces (via an appropriate data structure) that the model checking problems for first order formulas is Fixed Parameter Tractable for a monotone class \mathcal{C} of structures if and only if the class \mathcal{C} is nowhere dense (assuming standard hardness assumption in parametrized complexity). Thus the nowhere dense classes can be defined by the validity of *arithmetic meta-theorems*, see [21, 24] for graph case. For structures we can use I-Nowhere dense definition.

4.3. Classification by Counting. The trichotomy theorem (Theorem 2) is related to counting the numbers of copies of K_2 in a graph. This may be extended (using the decomposition theorem) if we consider homomorphism or induced copies of any non-trivial graph F . (Recall that $\text{hom}(F, G)$ denotes the number of homomorphisms from F to G and that $\#F \subseteq G$ denotes the number of induced subgraphs of G which are isomorphic to F .)

Theorem 10. *Let F be a (connected) non trivial graph (i.e. with at least one edge). Then the following limits*

$$\begin{aligned} \lim_{i \rightarrow \infty} \limsup_{G \in \mathcal{C} \nabla_i} \frac{\log \text{hom}(F, G)}{\log |G|}, & \qquad \lim_{i \rightarrow \infty} \limsup_{G \in \mathcal{C} \bar{\nabla}_i} \frac{\log \text{hom}(F, G)}{\log |G|}, \\ \lim_{i \rightarrow \infty} \limsup_{G \in \mathcal{C} \nabla_i} \frac{\log \#F \subseteq G}{\log |G|}, & \qquad \text{and} \qquad \lim_{i \rightarrow \infty} \limsup_{G \in \mathcal{C} \bar{\nabla}_i} \frac{\log \#F \subseteq G}{\log |G|} \end{aligned}$$

can only take the values $-\infty, 0, 1, \dots, \alpha(F)$ and $|F|$, where $\alpha(F)$ stands for the independence number of F . Moreover, \mathcal{C} is somewhere dense if and only if the limit is $|F|$.

For a proof, see [72]. There is more to this than meets the eye. The recent theory of graph limits developed by Laci Lovász with his coauthors, see e.g. [54, 55, 57, 56] deals with counting of homomorphisms from small graphs, or alternatively, with probabilities that a random map is a homomorphism. In this context the last alternative description is very pleasing as it may be seen as bridging the gap between these approaches and an approach based on the analysis of the homomorphism order, i.e. with existence of homomorphism (see [71]). For structures is the situation more involved and we do not state it here.

There are other alternative descriptions of nowhere dense - somewhere dense dichotomy (related to on-line colorings and game chromatic numbers). We refer to the forthcoming book [73]. From manifold applications we mention (in the next two sections) only two recent ones (and refer to e.g. [73, 64, 70, 75] instead).

5. Vertex Separators

Let G be a graph of order n . Recall that an α -vertex separator of G is a subset S of vertices such that every connected component of $G - S$ contains at most αn vertices.

5.1. Sub-exponential ω -expansion. A celebrated theorem of Lipton and Tarjan [52] states that any planar graph has a separator of size $O(\sqrt{n})$. Alon, Seymour and Thomas [3] showed that excluding K_h as a minor ensures the existence of a separator of size at most $O(h^{3/2}\sqrt{n})$. Gilbert, Hutchinson, and Tarjan [36] further proved that graphs with genus g have a separator of size $O(\sqrt{gn})$ (this result is optimal). Plotkin et al. [81] introduced the concept of *limited-depth minor* exclusion and have shown that exclusion of small limited-depth minors implies the existence of a small separator. Precisely, Plotkin et al. prove in [81] that any graph excluding K_h as a depth l minor (i.e. any graph G such that $K_h \not\in G \nabla l$) has a separator of size $O(lh^2 \log n + n/l)$ hence proving that excluding a K_h minor ensures the existence of a separator of size $O(h\sqrt{n} \log n)$.

We combine this with the following variant of expansion: The ω -expansion of a class \mathcal{C} is the mapping

$$i \mapsto \sup_{G \in \mathcal{C} \nabla i} \omega(G),$$

where $\omega(G)$ stands for the *clique number* of G , i.e. the order of the largest complete subgraph of G . Notice that a class has bounded ω -expansion if and only if it is nowhere dense.

A class \mathcal{C} has sub-exponential ω -expansion if

$$\limsup_{i \rightarrow \infty} \sup_{G \in \mathcal{C} \nabla i} \frac{\log \omega(G)}{i} = 0.$$

Theorem 11. *Let \mathcal{C} be a class of graphs with sub-exponential ω -expansion.*

Then the graphs of order n in \mathcal{C} have separators of size $s(n) = o(n)$ which may be computed in time $O(ns(n)) = o(n^2)$.

As random cubic graphs almost surely have bisection width at least $0.101n$ [50], they have almost surely no separator of size smaller than $n/20$. It follows that if $\log f(x) = (\log 2)x$, the graphs have no sublinear separators any more. This shows the optimality of Theorem 11.

6. Property Testing and Weak Hyperfiniteness

6.1. Property testing. Property testing has been introduced by Blum, Luby and Rubinfeld [16] and Rubinfeld and Sudan [86] (in the context of program testing), and by Arora, Lund, Motwani, Sudan, and M. Szegedy [7] and Arora and Safra [8] (in the context of probabilistically checkable proofs). Testing graph properties was first investigated by Goldreich, Goldwasser, and Ron [37]. From a “mathematical” point of view, the main ingredients of property testing are:

- a random sampling of the large structure,
- a suitable notion of distance between objects.

Let \mathcal{P} be a class of graphs (called *graph property* in this context). A graph G is said to have property \mathcal{P} if $G \in \mathcal{P}$; it is said to be ϵ -far for satisfying \mathcal{P} if no graph at distance at most ϵ from G satisfies \mathcal{P} . A testing algorithm (or *tester*) for graph property \mathcal{P} and accuracy ϵ is an algorithm that distinguishes with probability at least $2/3$ between graphs satisfying \mathcal{P} from graphs that are ϵ -far from satisfying it. More precisely, the property testing algorithm

- should accept with probability at least $2/3$ every input graph that belongs to \mathcal{P} ,
- should reject with probability at least $2/3$ every input graph that has distance more than ϵ from any graph in \mathcal{P} , i.e. if its ϵ -far from satisfying \mathcal{P} .

A graph property \mathcal{P} is *testable* if for any $\epsilon > 0$, there is a constant time randomized algorithm that can distinguish with high probability between graphs satisfying \mathcal{P} from those that are ϵ -far from satisfying it.

One should notice that the introduction of the parameter ϵ will make some properties impossible to distinguish. Precisely, two properties \mathcal{P} and \mathcal{Q} are *indistinguishable* if for every $\epsilon > 0$ there exists $N = N(\epsilon)$ such that:

- for every graph $G \in \mathcal{P}$ with order at least N there exists $H \in \mathcal{Q}$ with the same order such that $\text{dist}(G, H) < \epsilon$,
- for every graph $H \in \mathcal{Q}$ with order at least N there exists $G \in \mathcal{P}$ with the same order such that $\text{dist}(G, H) < \epsilon$.

As proved in [2] (in the context of dense graphs, but easily extended to the general case), if two properties are indistinguishable then either they are both testable or none of them is testable. Dense graphs (and more generally structures) seem to be well understood and we refer here to a spectacular chain of results [6, 5, 84, 82, 12, 49, 11] to name just a sample of this development (which starts with Szemerédi regularity lemma [88]). For sparse graphs the situation seems to be much less understood. Extending earlier results [38, 13, 1, 87] most general result is using the notion hyperfiniteness:

A class \mathcal{C} of (finite) graphs is *hyperfinite* if for every positive real $\epsilon > 0$ there exists a positive integer $K(\epsilon)$ such that every graph $G \in \mathcal{C}$ has a subset of at most $\epsilon|G|$ edges whose deletion leaves no connected component of order greater than $K(\epsilon)$ (see e.g. [53, 26, 27, 28, 41]).

In [14], Benjamini, Schramm and Shapira showed that every minor-closed graph property can be tested with a constant number of queries in the bounded degree model. For instance, planarity is testable in the bounded degree model. Actually, they prove a much stronger theorem:

Theorem 12 ([14]). *Every monotone hyperfinite graph property is testable.*

Using a detailed analysis of bounded expansion classes with an sub-exponential growth we can extend the range of applications of this result.

6.2. Weakly hyperfinite classes. A class \mathcal{C} of graphs is *weakly hyperfinite* if for any $\epsilon > 0$ there exists $K(\epsilon)$ such that every $G \in \mathcal{C}$ has a subset of at most $\epsilon|G|$ vertices whose deletion leaves no connected component of order greater than K .

Although it is obvious that a monotone class of graphs needs to have bounded degrees in order to be hyperfinite, weakly hyperfinite classes may have unbounded degrees. Moreover, it is straightforward that any hyperfinite class is also weakly hyperfinite.

The relation between the two notions will be made precise by the following result:

Theorem 13. *For a positive integer D , denote by Δ_D the class of the graphs having maximum degree at most D . Let \mathcal{C} be a monotone class of graphs with bounded average degree.*

The class \mathcal{C} is weakly hyperfinite if and only if for every integer D the class $\mathcal{C} \cap \Delta_D$ is hyperfinite.

A key advantage of the notion of weak hyperfinite class is its connection with the existence of sublinear vertex separators. For space limitations we leave out details and we just state the following:

Theorem 14. *Every monotone class of graphs with sublinear vertex separators is weakly hyperfinite. Consequently we have: Let \mathcal{C} be a monotone class of graphs with sublinear vertex-separators and bounded average degree and let D be a positive integer. Then the subclass of \mathcal{C} including those graphs in \mathcal{C} which have maximum degree at most D is hyperfinite.*

Combining with our results about vertex separators we arrive to the following:

Theorem 15. *Let \mathcal{P} be a monotone class of graphs with sub-exponential ω -expansion.*

Then the property $G \in \mathcal{P}$ is testable in the bounded degree model.

7. Selected Examples

1. **Classical Sparse Classes.** Fig. 4 shows the inclusion map of some important hereditary nowhere dense classes which were studied in combinatorial as well as algorithmic context.

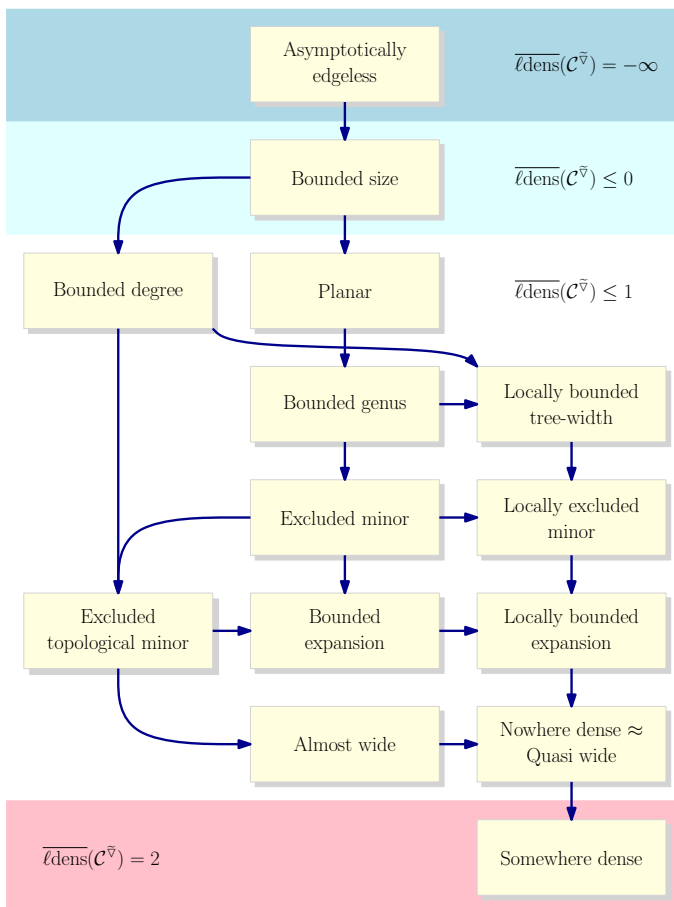


Figure 4. Inclusion map of some important hereditary nowhere dense classes.

2. **Simplicial Graphs.** A k -dimensional simplex, or k -simplex, is the convex hull of $k + 1$ affinely independent points in \mathbb{R}^d space. A d -dimensional simplicial complex is a collection of k -simplices, $k \leq d$, closed under sub-simplex and intersection. For example, a 3-dimensional simplicial complex is a collection of cells (3-simplices), faces (2-simplices), edges (1-simplices) and vertices (0-simplices). A d -dimensional simplicial graph is the collection of edges and vertices of a d -dimensional simplicial complex. The aspect ratio of a body is its

diameter divided d th root of its volume [60]. The volume of a regular d -simplex, d -cube, and d -ball of unit diameter are respectively $2^{-d/2}\sqrt{d+1}/d!$, $d^{-d/2}$ and $2^{-d}\pi^{d/2}/(d/2)!$. Hence the aspect ratios of a d -simplex, d -cube, and d -ball are respectively $\alpha_s = 2^{1/2}(d!)^{1/d}(d+1)^{-1/(2d)} \sim \sqrt{2d}/e$, $\alpha_c = \sqrt{d}$, and $\alpha_b = 2\pi^{-1/2}(d/2)^{1/d} \sim \sqrt{2d}/(e\pi)$. A simplicial graph of aspect ratio α means a simplicial graph coming from a complex in which every d -simplex has aspect ratio at most α .

Classes of simplicial graphs with bounded aspect ratio exclude big shallow complete minors as proved by Plotkin, Rao and Smith [81]. It follows that such classes are nowhere dense.

3. High Girth Graphs. A standard example of a monotone nowhere dense class of graphs is the class of the graphs whose maximum degree does not exceed some function of the girth, i.e. $\mathcal{B}_\phi = \{G : \Delta(G) \leq \phi(\text{girth}(G))\}$.

Such classes may have average degree as big as $n^{o(1)}$ as a consequence (see for instance [18]): For every positive integer n and an “expected degree” k (where $k < n/3$), there exists a graph G of order n , size $\lfloor nk/2 \rfloor$, vertex degrees in $\{k-1, k, k+1\}$ and whose girth g is such that $g > \log_k(n) + O(1)$. Hence, for any decreasing function $f : \mathbb{R}^+ \rightarrow \mathbb{R}^+$ such that $\lim_{x \rightarrow \infty} f(x) = 0$ there exists a constant C such that the class \mathcal{B}_ϕ defined by $\phi(x) = (f^{-1}(1/x) + C)^{1/x}$ contains graphs with order n , girth at least $1/f(n)$ and degrees $k \pm 1$ with $k \approx n^{f(n)}$.

8. Bounded Expansion Classes

A specific example of classes which are nowhere dense are classes with bounded expansion. These classes have been introduced in [64]. A class \mathcal{C} has *bounded expansion* if there exists a function $f : \mathbb{N} \rightarrow \mathbb{R}$ (called *expansion function*) such that

$$\forall d \in \mathbb{N} \quad \sup_{G \in \mathcal{C} \nabla d} \frac{\|G\|}{|G|} \leq f(d).$$

The value $\sup_{G \in \mathcal{C} \nabla d} \frac{\|G\|}{|G|}$ is denoted by $\nabla_d(\mathcal{C})$ and, in the particular case of a single element class $\{G\}$, $\nabla_d(G)$ is called the *greatest reduced average density* (grad) of G of rank d .

Classes with bounded expansion include [75]

- classes excluding a topological minor (this includes classes excluding a minor, like planar graphs, and also classes with bounded maximum degree),
- k -non-repetitively colorable graphs (see [40] for more details on non-repetitive colorings),
- geometrically defined classes like classes with bounded stack number and classes with bounded queue number,

- classes of highly subdivided graphs (allowing to construct examples of classes with arbitrary non-decreasing expansion function),
- sparse random graphs (in the sense that for every positive real d there exists a class \mathcal{R}_d with bounded expansion such that random graphs with edge probability d/n asymptotically almost surely belong to \mathcal{R}_d).

For an extensive study of bounded expansion classes we refer the reader to [66], [67], [68], [22], [23], [75].

(See [75] for the definition of stack and queue numbers. This paper contains further examples of bounded expansion classes.)

As for nowhere dense classes, several equivalent characterizations exist for classes with bounded expansion:

Theorem 16. *Let \mathcal{C} be a class of graphs. The following properties are equivalent:*

- \mathcal{C} has bounded expansion,
- for every integer p , $\sup_{G \in \mathcal{C}} \chi_p(G) < \infty$.

Thus any graph G in a (fixed) bounded expansion class \mathcal{C} can be decomposed into a fixed number $N_p(G)$ of classes such that the subgraphs induced by any $\leq p$ classes of the partition have components of only finitely many (homomorphism) types. Thus p is then parameter expressing the precision of such decomposition. Moreover such decomposition can be found in a linear number of steps. Not surprisingly, this has a number of algorithmic consequences ([64, 67]. Such a decomposition is called *Low Tree Depth Decomposition* and it was described explicitly in the Introduction.

9. Restricted Dualities — a Characterization

In the Introduction we described homomorphism dualities for general relational systems. Clearly if we restrict the universe of the considered structures G then we can expect more “dual phenomena”. In such cases we speak about *restricted dualities*. Explicitly, a (singleton) \mathcal{C} -restricted duality is formed by a pair (\mathbf{F}, \mathbf{D}) such that for every structure $\mathbf{A} \in \mathcal{C}$ holds:

$$\mathbf{F} \not\rightarrow \mathbf{A} \quad \iff \quad \mathbf{A} \longrightarrow \mathbf{D}.$$

. Note that we do not assume that $\mathbf{D} \in \mathcal{C}$. In the interpretation of the homomorphism order this just amounts to $\text{Forb}(\mathbf{F}) \cap \mathcal{C} = \text{CSP}(\mathbf{D}) \cap \mathcal{C}$.

In the extremal case that for every connected $F \in \mathcal{C}$ there exists D_F such that F, D_F form a \mathcal{C} -restricted duality we say that \mathcal{C} has *all restricted dualities* [68].

These two examples actually fit to a much more a general setting which has been proved by [68]:

Theorem 17. *Every class of structures with G -bounded expansion has all restricted dualities.*

Explicitly: For every bounded expansion class \mathcal{C} and for any finite set $\mathcal{F} = \{\mathbf{F}_1, \mathbf{F}_2, \dots, \mathbf{F}_t\}$ of connected graphs there exists a structure $\mathbf{D}_{\mathcal{F}}$ such that $\mathbf{D}_{\mathcal{F}} \in \text{Forb}(\mathcal{F})$ and $\mathbf{A} \rightarrow \mathbf{D}_{\mathcal{F}}$ for every $\mathbf{A} \in \mathcal{C}$ and $\mathbf{A} \in \text{Forb}(\mathcal{F})$.

A characterization of classes with all restricted dualities was not known until recently (see e.g. [69]). One can deduce such a characterization using the following notions (related to the completion of the homomorphism order): Given two structures \mathbf{A}, \mathbf{B} we define their distance $\text{dist}(\mathbf{A}, \mathbf{B})$ as 2^{-L} where L is the minimal order $|\mathbf{C}|$ of a structure \mathbf{C} which distinguishes \mathbf{A} and \mathbf{B} either from left or right. This has the following meaning: *distinguishing from left* means that either $\mathbf{C} \leq \mathbf{A}$ and $\mathbf{C} \not\leq \mathbf{B}$ or $\mathbf{C} \not\leq \mathbf{A}$ and $\mathbf{C} \leq \mathbf{B}$; similarly, *distinguishing from right* means that either $\mathbf{A} \leq \mathbf{C}$ and $\mathbf{B} \not\leq \mathbf{C}$ or $\mathbf{A} \not\leq \mathbf{C}$ and $\mathbf{B} \leq \mathbf{C}$. $\text{dist}(\mathbf{A}, \mathbf{B})$ is an ultrametric on the class $\mathcal{R}el(\Delta)$ which can be used to define the completion of the homomorphism order. This completion has interesting properties particularly with respect to dualities (see [71]). Let us just state here the following:

For a structure \mathbf{A} and a real $\epsilon > 0$, define $\phi^\epsilon(\mathbf{A})$ as a minimum order of a structure \mathbf{B} such that $\mathbf{A} \rightarrow \mathbf{B}$ and $\text{dist}(\mathbf{A}, \mathbf{B}) \leq \epsilon$ (we arbitrarily choose between those structures which have these properties, by using, for instance, some arbitrary linear order on $\mathcal{R}el(\Delta)$; such structure \mathbf{B} we can call ϵ -retract of \mathbf{A}).

Theorem 18. *Let \mathcal{C} be a class of structures. Then \mathcal{C} has all restricted dualities if and only if for every $\epsilon > 0$ we have $\sup_{\mathbf{A} \in \mathcal{C}} \phi^\epsilon(\mathbf{A}) < \infty$.*

Moreover, for every connected structure \mathbf{F} , there is a sequence $\mathbf{D}_t(\mathbf{F}) \leftarrow \mathbf{D}_{t+1}(\mathbf{F}) \leftarrow \dots$ of duals of \mathbf{F} relative to \mathcal{C} which converges to $\sup(\mathcal{C}^+ \cap \text{Forb}(\mathbf{F}))$, where \mathcal{C}^+ denotes the closure of \mathcal{C} by all finite disjoint unions of structures in \mathcal{C} .

References

- [1] D. Aldous and R. Lyons, *Processes on unimodular random networks*, arXiv:math/0603062, 2006.
- [2] N. Alon, E. Fischer, M. Krivelevich, and M. Szegedy, *Efficient testing of large graphs*, *Combinatorica* **20** (2000), 451–476.
- [3] N. Alon, P.D. Seymour, and R. Thomas, *A separator theorem for graphs with excluded minor and its applications*, *Proceedings of the 22nd Annual ACM Symposium on Theory of Computing*, 1990, pp. 293–299.
- [4] ———, *A separator theorem for nonplanar graphs*, *J. Amer. Math. Soc.* **3** (1990), 801–808.
- [5] N. Alon and A. Shapira, *A characterization of the (natural) graph properties testable with one-sided error*, *Proc. 46th IEEE FOCS*, 2005, pp. 429–438.

- [6] ———, *A characterization of the (natural) graph properties testable with one-sided error*, SIAM J. Comp. **37** (2008), no. 6, 1703–1727.
- [7] S. Arora, C. Lund, R. Motwani, M. Sudan, and M. Szegedy, *Proof verification and the hardness of approximation problems*, Journal of the ACM **45** (1998), 501–555.
- [8] S. Arora and S. Safra, *Probabilistic checking of proofs: a new characterization of NP*, Journal of the ACM **45** (1998), 70–122.
- [9] A. Atserias, *On digraph coloring problems and treewidth duality*, European J. Combin. **29** (2008), no. 4, 796–820.
- [10] A. Atserias, A. Dawar, and P.G. Kolaitis, *On preservation under homomorphisms and unions of conjunctive queries*, J. ACM **53** (2006), 208–237.
- [11] T. Austin and T. Tao, *On the testability and repair of hereditary hypergraph properties*, 2009.
- [12] C. Avart, V. Rödl, and M. Schacht, *Every monotone 3-graph property is testable*, Electronic Notes in Discrete Mathematics **22** (2005), 539–542, 7th International Colloquium on Graph Theory.
- [13] I. Benjamini and O. Schramm, *Recurrence of distributional limits of finite planar graphs*, Electron. J. Probab. **6** (2001), no. 23, 13pp.
- [14] I. Benjamini, O. Schramm, and A. Shapira, *Every minor-closed property of sparse graphs is testable*, Proceedings of the 40th annual ACM symposium on Theory of computing, 2008, pp. 393–402.
- [15] C. Berge, *Graphes*, troisième ed., Gauthier-Villars, Paris, 1983.
- [16] M. Blum, M. Luby, and R. Rubinfeld, *Self-testing/correcting with applications to numerical problems*, JCSS **47** (1993), 549–595.
- [17] C. Borgs, J. Chayes, L. Lovász, V.T. Sós, and K. Vesztegombi, *Counting graph homomorphisms*, Topics in Discrete Mathematics (M. Klazar, J. Kratochvíl, M. Loeb, J. Matoušek, R. Thomas, and P. Valtr, eds.), Algorithms and Combinatorics, vol. 26, Springer Verlag, 2006, pp. 315–371.
- [18] L.S. Chandran, *A high girth graph construction*, SIAM J. Discret. Math. **16** (2003), no. 3, 366–370.
- [19] F. R. K. Chung, R. L. Graham, and R. M. Wilson, *Quasi-random graphs*, Combinatorica **9** (1989), no. 4, 345–362.
- [20] A. Dawar, *Finite model theory on tame classes of structures*, Mathematical Foundations of Computer Science 2007 (L. Kučera and A. Kučera, eds.), Lecture Notes in Computer Science, vol. 4708, Springer, 2007, pp. 2–12.
- [21] A. Dawar and S. Kreutzer, *Parametrized complexity of first-order logic*, Tech. Report 131, Electronic Colloquium on Computational Complexity, 2009.
- [22] Z. Dvořák, *Asymptotical structure of combinatorial objects*, Ph.D. thesis, Charles University, Faculty of Mathematics and Physics, 2007.
- [23] ———, *On forbidden subdivision characterizations of graph classes*, European J. Combin. **29** (2008), no. 5, 1321–1332.
- [24] Z. Dvořák, D. Král, and R. Thomas, *Finding subgraphs and testing FOL properties in sparse graphs*, Tech. report, KAM Series, 2009.

- [25] Z. Dvořák, D. Král', and R. Thomas, *A linear time for deciding first-order properties on classes with bounded expansion*, personal communication, 2009.
- [26] G. Elek, *The combinatorial cost*, arXiv:math/0608474, 2006.
- [27] ———, *L^2 -spectral invariants and convergent sequences of finite graphs*, arXiv:0709.1261, 2007.
- [28] ———, *A regularity lemma for bounded degree graphs and its applications: Parameter testing and infinite volume limits*, arXiv:0711.2800, 2007.
- [29] ———, *Parameter testing in bounded degree graphs of subexponential growth*, arXiv:0711.2800v3[math.CO], July 2009.
- [30] D. Eppstein, *Subgraph isomorphism in planar graphs and related problems*, Proc. 6th Symp. Discrete Algorithms, ACM and SIAM, January 1995, pp. 632–640.
- [31] ———, *Subgraph isomorphism in planar graphs and related problems*, Journal of Graph Algorithms & Applications **3** (1999), no. 3, 1–27.
- [32] R. Fagin, *Generalized first-order spectra and polynomial-time recognizable sets*, Complexity of Computation (R. Karp, ed.), SIAM–AMS Proceedings, vol. 7, 1974, pp. 43–73.
- [33] T. Feder and M.Y. Vardi, *The computational structure of monotone monadic SNP and constraint satisfaction: A study through datalog and group theory.*, SIAM J. Comput. **28** (1998), no. 1, 57–104 (English).
- [34] ———, *Homomorphism closed vs. existential positive*, Logic in Computer Science, Symposium on **0** (2003), 311.
- [35] J. Foniok, J. Nešetřil, A. Pultr, and C. Tardif, *Dualities and dual pairs in heyting algebras*, Order (2010), to appear.
- [36] J.R. Gilbert, J.P. Hutchinson, and R.E. Tarjan, *A separator theorem for graphs of bounded genus*, J. Algorithms (1984), no. 5, 375–390.
- [37] O. Goldreich, S. Goldwasser, and D. Ron, *Property testing and its connection to learning and approximation*, JACM **45** (1998), no. 4, 653–750.
- [38] O. Goldreich and D. Ron, *Property testing in bounded-degree graphs*, Algorithmica **32** (2002), 302–343.
- [39] W.T. Gowers, *Hypergraph regularity and the multidimensional Szemerédi theorem*, Annals of Mathematics **166** (2007), no. 3, 897–946.
- [40] J. Grytczuk, *Nonrepetitive colorings of graphs—a survey*, Int. J. Math. Math. Sci. (2007), Art. ID 74639.
- [41] A. Hassidim, J. Kelner, H. Nguyen, and Onak; K., *Local graph partitions for approximation and testing*, Proceedings of the Fiftieth Annual Symposium on Foundations of Computer Science (FOCS), 2009.
- [42] Z. Hedrlín, *On universal partly ordered sets and classes*, J. Algebra **11** (1969), 503–509.
- [43] P. Hell and J. Nešetřil, *Graphs and homomorphisms*, Oxford Lecture Series in Mathematics and its Applications, vol. 28, Oxford University Press, 2004.
- [44] W. Hodges, *Model theory*, Cambridge University Press, 1993.

- [45] J. Hubička and J. Nešetřil, *Homomorphism and embedding universal structures for restricted classes*, submitted.
- [46] ———, *Finite paths are universal*, *Order* **22** (2005), 21–40.
- [47] ———, *Universal partial order represented by means of oriented trees and other simple graphs*, *European J. Combin.* **26** (2005), no. 5, 765–778.
- [48] ———, *A finite presentation of the rational Urysohn space*, *Topology and its applications* **155** (2008), no. 14, 1483–1492.
- [49] Y. Ishigami, *Removal lemma for infinitely-many forbidden hypergraphs and property testing*, 2008.
- [50] A.V. Kostochka and L.S. Melnikov, *On bounds of the bisection width of cubic graphs*, Fourth Czechoslovakian Symposium on Combinatorics, Graphs and Complexity (J. Nešetřil and M. Fiedler, eds.), Elsevier, 1992, pp. 151–154.
- [51] G. Kun and J. Nešetřil, *NP for combinatorialists*, *European J. Combin.* **29** (2007), no. 15, 373–381.
- [52] R. Lipton and R.E. Tarjan, *A separator theorem for planar graphs*, *SIAM Journal on Applied Mathematics* **36** (1979), no. 2, 177–189.
- [53] R.J. Lipton and R.E. Tarjan, *Applications of a planar separator theorem*, *SIAM J. Comp.* **9** (1980), no. 3, 615–627.
- [54] L. Lovász and M.L. Marx, *A forbidden subgraph characterization of Gauss codes*, *Bull. Am. Math. Soc.* **82** (1976), 121–122.
- [55] L. Lovász and V.T. Sós, *Generalized quasirandom graphs*, *J. Combin. Theory Ser. B* **98** (2008), 146–163.
- [56] L. Lovász and B. Szegedy, *Limits of dense graph sequences*, *J. Combin. Theory Ser. B* **96** (2006), 933–957.
- [57] ———, *Szemerédi lemma for the analyst*, *Geom. Func. Anal.* **17** (2007), 252–270.
- [58] R.C. Lyndon, *Properties preserved under homomorphism*, *Pacific J. Math.* **9** (1959), 129–142.
- [59] J. Matoušek and J. Nešetřil, *Invitation to discrete mathematics*, Oxford University Press, 1998 (second printing 2008).
- [60] G. L. Miller, S.-H. Teng, W. Thurston, and S. A. Vavasis, *Geometric separators for finite-element meshes.*, *SIAM J. Sci. Comput.* **19** (1998), no. 2, 364–386 (English).
- [61] J. Nešetřil, *Aspects of structural combinatorics – graph homomorphisms and their use*, *Taiwanese J. Math.* **3** (1999), no. 4, 381–424.
- [62] J. Nešetřil and P. Ossona de Mendez, *Colorings and homomorphisms of minor closed classes*, The Goodman-Pollack Festschrift (B. Aronov, S. Basu, J. Pach, and M. Sharir, eds.), Algorithms and Combinatorics, vol. 25, Discrete & Computational Geometry, 2003, pp. 651–664.
- [63] ———, *Cuts and bounds*, *Discrete Mathematics, Structural Combinatorics - Combinatorial and Computational Aspects of Optimization, Topology and Algebra* **302** (2005), no. 1–3, 211–224.

- [64] ———, *Linear time low tree-width partitions and algorithmic consequences*, STOC'06. Proceedings of the 38th Annual ACM Symposium on Theory of Computing, ACM Press, 2006, pp. 391–400.
- [65] ———, *Tree depth, subgraph coloring and homomorphism bounds*, European Journal of Combinatorics **27** (2006), no. 6, 1022–1041.
- [66] ———, *Grad and classes with bounded expansion I. decompositions*, European Journal of Combinatorics **29** (2008), no. 3, 760–776.
- [67] ———, *Grad and classes with bounded expansion II. algorithmic aspects*, European Journal of Combinatorics **29** (2008), no. 3, 777–791.
- [68] ———, *Grad and classes with bounded expansion III. restricted graph homomorphism dualities*, European Journal of Combinatorics **29** (2008), no. 4, 1012–1024.
- [69] ———, *Structural properties of sparse graphs*, Building Bridges Between Mathematics and Computer Science (Martin Grötschel and Gyula O.H. Katona, eds.), Bolyai Society Mathematical Studies, vol. 19, Springer, 2008, pp. 369–426.
- [70] ———, *First order properties on nowhere dense structures*, The Journal of Symbolic Logic (2009), accepted.
- [71] ———, *From sparse graphs to nowhere dense structures: Decompositions, independence, dualities and limits*, Proc. of the fifth European Congress of Mathematics, 2009, accepted.
- [72] ———, *How many F 's are there in G ?*, European Journal of Combinatorics (2009), submitted.
- [73] ———, *Sparsity (graphs, structures, and algorithms)*, 2009.
- [74] ———, *On nowhere dense graphs*, European Journal of Combinatorics (2010), accepted.
- [75] J. Nešetřil, P. Ossona de Mendez, and D.R. Wood, *Characterizations and examples of graph classes with bounded expansion*, European Journal of Combinatorics (2009), submitted.
- [76] J. Nešetřil, P. Ossona de Mendez, and X. Zhu, *Generalized acyclic edge colorings and generalized arboricity*, in preparation.
- [77] J. Nešetřil and A. Pultr, *On classes of relations and graphs determined by subobjects and factorobjects*, Discrete Math. **22** (1978), 287–300.
- [78] J. Nešetřil and V. Rödl, *Partitions of relational and set systems*, Journal of Combinatorial Theory, Series A **22** (1977), 289–312.
- [79] J. Nešetřil and C. Tardif, *Duality theorems for finite structures (characterizing gaps and good characterizations)*, Journal of Combinatorial Theory, Series B **80** (2000), 80–97.
- [80] J. Nešetřil and X. Zhu, *Path homomorphisms*, Proc. Cambridge Phil. Soc **120** (1996), 207–220.
- [81] S. Plotkin, S. Rao, and W.D. Smith, *Shallow excluded minors and improved graph decomposition*, 5th Symp. Discrete Algorithms, SIAM, 1994, pp. 462–470.
- [82] V. Rödl and M. Schacht, *Property testing in hypergraphs and the removable lemma*, Proceedings of the thirty-ninth annual ACM symposium on Theory of computing, 2007, pp. 488–495.

- [83] ———, *Regular partitions of hypergraphs: Regularity lemmas*, *Combin. Probab. Comput.* **16** (2007), 833–885.
- [84] ———, *Generalizations of the removal lemma*, *Combinatorica* **29** (2009), no. 4, 467–501.
- [85] B. Rossman, *Homomorphism preservation theorems*, *J. ACM* **55** (2008), no. 3, 1–53.
- [86] R. Rubinfeld and M. Sudan, *Robust characterization of polynomials with applications to program testing*, *SIAM J. Comp.* **25** (1996), 252–271.
- [87] O. Schramm, *Hyperfinitite graph limits.*, *Electron. Res. Announc. Math. Sci.* **15** (2008), 17–23.
- [88] E. Szemerédi, *Regular partitions of graphs*, *Colloq. Int. CNRS* **260** (1978), 399–401.
- [89] T. Tao, *The dichotomy between structure and randomness, arithmetic progression, and the primes*, *Proceedings of the International Congress of Mathematicians (Madrid 2006)* (European Math. Society, ed.), vol. 1, 2007, pp. 581–608.
- [90] S. Thomas, *Reducts of the random graph*, *The Journal of Symbolic Logic* **56** (1991), no. 1, 176–181.
- [91] E. Welzl, *Color families are dense*, *Theoret. Comput. Sci.* **17** (1982), 29–41.

Elliptic Analogues of the Macdonald and Koornwinder Polynomials

Eric M. Rains*

Abstract

Perhaps the nicest multivariate orthogonal polynomials are the Macdonald and Koornwinder polynomials, respectively 2-parameter deformations of Schur functions and 6-parameter deformations of orthogonal and symplectic characters, satisfying a trio of nice properties known as the Macdonald “conjectures”. In recent work, the author has constructed elliptic analogues: a family of multivariate functions on an elliptic curve satisfying analogues of the Macdonald conjectures, and degenerating to Macdonald and Koornwinder polynomials under suitable limits. This article will discuss the two main constructions for these functions, focusing on the more algebraic/combinatorial of the two approaches.

Mathematics Subject Classification (2010). Primary 33D52, Secondary 14H52

Keywords. Macdonald polynomials, elliptic curves, special functions

1. Introduction

One of the most important advances in the theory of symmetric functions in recent decades was Macdonald’s introduction of his eponymous polynomials, which introduce two parameters into the Schur functions while retaining many of their properties. Recall (see, e.g. [15]) that when q and t are complex numbers with $|q|, |t| < 1$, the Macdonald polynomial $P_\lambda(x_1, \dots, x_n; q, t)$ for a partition λ is the unique polynomial which is invariant under permutations of its arguments, has leading (dominant) monomial $\prod_i x_i^{\lambda_i}$, and is orthogonal with respect to an appropriate density on the unit torus (here, and below, all integrals are with respect to the uniform measure on the unit torus $|x_i| = 1$; this can also

*Partially supported by the NSF, grant no. DMS-0401387.

Mathematics MC 253-37, California Institute of Technology, Pasadena, CA 91125 USA.
E-mail: rains@caltech.edu.

be viewed as a contour integral, replacing the measure by $\prod_i dx_i/2\pi\sqrt{-1x_i}$:

$$\int P_\lambda(x_1, \dots, x_n; q, t) P_\mu\left(\frac{1}{x_1}, \dots, \frac{1}{x_n}; q, t\right) \prod_{1 \leq i < j \leq n} \frac{(x_i/x_j, x_j/x_i; q)}{(tx_i/x_j, tx_j/x_i; q)} \propto \delta_{\lambda\mu}.$$

Here $(x; q)_k$ is the q -Pochhammer symbol

$$(x; q)_k = \prod_{0 \leq j < k} (1 - q^j x),$$

and we take the conventions of omitting k when $k = \infty$, and denoting a product of q -Pochhammer symbols via multiple arguments:

$$\begin{aligned} (r, s, \dots, z; q) &:= (r; q)(s; q) \cdots (z; q), \\ (z_i^{\pm 1} z_j^{\pm 1}; q) &:= (z_i z_j, z_i/z_j, z_j/z_i, 1/z_i z_j; q), \end{aligned}$$

and so forth. (We will take similar conventions for $\Gamma_{p,q}$ and θ_p below.)¹ It should be noted that since the natural ordering on monomials is only a partial ordering, it is by no means obvious that the Macdonald polynomials even exist. (For univariate polynomials, one can always simply apply Gram-Schmidt to the sequence $1, x, x^2, \dots$, and obtain an orthogonal basis of monic polynomials; but Gram-Schmidt with respect to a partial ordering only implies orthogonality when the corresponding elements are comparable in the partial order.) When $t = q$, the density becomes

$$\prod_{1 \leq i < j \leq n} |x_i - x_j|^2,$$

the unnormalized probability density of the eigenvalues of a random unitary matrix. It follows that Macdonald polynomials with $t = q$ are precisely the irreducible characters of the unitary group. (Characters of the unitary group extend to symmetric functions of infinitely many variables (the Schur functions), and the same applies to Macdonald polynomials (and Koornwinder polynomials, see [21]), but we will be concentrating on the finite case.)

The three main properties of the Macdonald polynomials (beyond mere existence) are the so-called Macdonald conjectures (later proved by Macdonald [15]): evaluation (an explicit product formula for the value at $x_i = t^{n-i}$), norm (an explicit product formula for the nonzero inner products), and perhaps most strikingly the symmetry property:

$$\frac{P_\lambda(\dots, q^{\mu_i} t^{n-i}, \dots; q, t)}{P_\lambda(\dots, t^{n-i}, \dots; q, t)} = \frac{P_\mu(\dots, q^{\lambda_i} t^{n-i}, \dots; q, t)}{P_\mu(\dots, t^{n-i}, \dots; q, t)}.$$

¹Note that when $t \in q^{\mathbb{N}}$, each ratio of Pochhammer symbols in the density becomes a finite Pochhammer symbol, so the density becomes polynomial and integration over the torus can be replaced by taking the constant term of the appropriate product of polynomials. It is common to formulate orthogonality of Macdonald polynomials in this purely algebraic way, but the (contour) integral formulation both allows more general parameters and extends more cleanly to the more general cases of interest.

Shortly after introducing the above (which we will refer to as “ordinary”) Macdonald polynomials, Macdonald came up with an analogous construction (see, e.g., [16]) for arbitrary root systems, with analogues of the three Macdonald conjectures, settled in nearly all cases by Cherednik using “double affine Hecke algebras” (see, e.g., [4]). Surprisingly, the last case of the conjectures to be settled (by Sahi [28], drawing on work of Noumi [17] and Van Diejen [7]) was the most general *classical* root system, largely because, as observed by Koornwinder [12], the existence of a nonreduced classical root system leads to an explosion of parameters; the Koornwinder polynomials have, in fact, four parameters in addition to q and t , with inner product (assuming all parameters lie in the unit disc)

$$\int K_\lambda(x_1, \dots, x_n; t_0, t_1, t_2, t_3; q, t) K_\mu(x_1, \dots, x_n; t_0, t_1, t_2, t_3; q, t) \prod_{1 \leq i \leq n} \frac{(x_i^{\pm 2}; q)}{(t_0 x_i^{\pm 1}, t_1 x_i^{\pm 1}, t_2 x_i^{\pm 1}, t_3 x_i^{\pm 1}; q)} \prod_{1 \leq i < j \leq n} \frac{(x_i^{\pm 1}; q)}{(t x_i^{\pm 1} x_j^{\pm 1}; q)} \propto \delta_{\lambda\mu}$$

These again satisfy analogues of the Macdonald conjectures, though the symmetry property becomes somewhat more complicated:

$$\frac{K_\lambda(\dots, q^{\mu_i} t^{n-i} t_0, \dots; t_0, t_1, t_2, t_3; q, t)}{K_\lambda(\dots, t^{n-i} t_0, \dots; t_0, t_1, t_2, t_3; q, t)} = \frac{K_\mu(\dots, q^{\lambda_i} t^{n-i} \hat{t}_0, \dots; \hat{t}_0, \hat{t}_1, \hat{t}_2, \hat{t}_3; q, t)}{K_\mu(\dots, t^{n-i} \hat{t}_0, \dots; \hat{t}_0, \hat{t}_1, \hat{t}_2, \hat{t}_3; q, t)},$$

where

$$\hat{t}_0^2 = t_0 t_1 t_2 t_3 / q, \quad \hat{t}_1 = t_0 t_1 / \hat{t}_0, \quad \hat{t}_2 = t_0 t_2 / \hat{t}_0, \quad \hat{t}_3 = t_0 t_3 / \hat{t}_0.$$

In [21], the author devised a new approach to the study of Koornwinder polynomials which in particular led to a new proof of the corresponding Macdonald conjectures. Though this new proof to date applies only to the Koornwinder case (and loses much of the elegance of Cherednik’s approach), the loss of the ability to treat nonclassical root systems is quite adequately compensated for by the ability, as we will see, to generalize in a quite different direction.

In [9] (partially anticipated by [6]), Frenkel and Turaev observed that, just as many hypergeometric identities (e.g., identities of sums of binomial coefficients) had natural q -analogues, one could extend the most general q -hypergeometric identities (Jackson’s summation and Bailey’s transformation; see [10] for an excellent survey of q -hypergeometric functions) to identities on an elliptic curve (i.e., replacing the additive and multiplicative groups by an arbitrary 1-dimensional algebraic group, so that the ratio of the k th and $k + 1$ st term is an elliptic function of k). Integral analogues of their identities were derived by Spiridonov, who in particular proved the *elliptic beta integral* [29]:

$$\frac{(p; p)(q; q)}{2} \int \frac{\prod_{0 \leq r < 6} \Gamma_{p,q}(u_r x^{\pm 1})}{\Gamma_{p,q}(x^{\pm 2})} = \prod_{0 \leq r < s < 6} \Gamma_{p,q}(u_r u_s),$$

where the parameters satisfy $|p|, |q|, |u_r| < 1$ as well as the *balancing condition*

$$u_0 u_1 u_2 u_3 u_4 u_5 = pq,$$

and $\Gamma_{p,q}$ denotes Ruijsenaars' *elliptic gamma function* [27]:

$$\Gamma_{p,q}(x) := \prod_{0 \leq j,k} \frac{1 - p^{j+1} q^{k+1} / x}{1 - p^j q^k x}.$$

This becomes a (reciprocal of a) q -Pochhammer symbol when $p = 0$ (and a p -Pochhammer symbol when $q = 0$, by symmetry), and satisfies the recurrence relations and reflection relation

$$\Gamma_{p,q}(qx) = \theta_p(x) \Gamma_{p,q}(x), \quad \Gamma_{p,q}(px) = \theta_q(x) \Gamma_{p,q}(x), \quad \Gamma_{p,q}(pq/x) = \Gamma_{p,q}(x)^{-1},$$

where

$$\theta_p(x) = \prod_{0 \leq k} (1 - p^k x)(1 - p^{k+1} / x)$$

is essentially Jacobi's theta function. The significance of the elliptic beta integral for our purposes is that it has as a limiting case (take $p \rightarrow 0$ with two parameters of order $p^{1/2}$) the identity [1]

$$\frac{(q; q)}{2} \int \frac{(z^{\pm 2}; q)}{(t_0 z^{\pm 1}, t_1 z^{\pm 1}, t_2 z^{\pm 1}, t_3 z^{\pm 1}; q)} = \frac{(t_0 t_1 t_2 t_3; q)}{(t_0 t_1, t_0 t_2, t_0 t_3, t_1 t_2, t_1 t_3, t_2 t_3; q)};$$

this is just the normalization for the orthogonality density for the Askey-Wilson polynomials, of which the Koornwinder polynomials are the multivariate analogues. Indeed, Spiridonov [30] also constructed analogues of the Askey-Wilson polynomials associated to the elliptic beta integral (themselves analogues of discrete elliptic biorthogonal functions [31], and generalizing known hypergeometric biorthogonal rational functions [19, 33]), and with Van Diejen [8] conjectured a multivariate analogue of the elliptic normalization (the *elliptic Selberg integral*):

$$\begin{aligned} & \frac{((p; p)(q; q) \Gamma_{p,q}(t))^n}{2^n n!} \int \prod_{1 \leq j < k \leq n} \frac{\Gamma_{p,q}(t x_j^{\pm 1} x_k^{\pm 1})}{\Gamma_{p,q}(x_j^{\pm 1} x_k^{\pm 1})} \prod_{1 \leq j \leq n} \frac{\prod_{0 \leq r < 6} \Gamma_{p,q}(u_r x_j^{\pm 1})}{\Gamma_{p,q}(x_j^{\pm 2})} \\ &= \prod_{0 \leq j < n} \left(\Gamma_{p,q}(t^{j+1}) \prod_{0 \leq r < s < 6} \Gamma_{p,q}(t^j u_r u_s) \right), \end{aligned} \tag{1}$$

with balancing condition $t^{2n-2} u_0 u_1 u_2 u_3 u_4 u_5 = pq$. In [24], the author gave two different proofs of this integral, and showed that one could combine ideas from the two proofs to give an explicit construction of a family of associated biorthogonal elliptic functions (by applying a sequence of difference and integral

operators starting with 1). This construction made two of the Macdonald conjectures quite straightforward, but left the symmetry conjecture open. This was established in [22] via a different construction for the biorthogonal functions. While the first construction is intrinsically complex analytic in nature (as the integral operators are defined via multivariate contour integrals), the second construction is much more combinatorial and algebraic, so (as the associated talk is in the Combinatorics section) we will focus on the second construction here, modified somewhat to make the arguments self-contained.

2. Interpolation Functions

The main tool in the author’s alternative proof of the Macdonald conjectures for Koornwinder polynomials was a family of polynomials constructed by Okounkov [18], his *(BC-type) interpolation polynomials*. The polynomial $P_\lambda^{*(n)}(; q, t, s)$ associated to a partition λ is the (generically) unique polynomial with the following properties:

- 1. $P_\lambda^{*(n)}$ is invariant under the Weyl group C_n (i.e., permutation and inversion of variables), with leading (dominant) monomial $\prod_i x_i^{\lambda_i}$.
- 2. For any partition μ ,

$$P_\lambda^{*(n)}(\dots, q^{\mu_i} t^{n-i} s, \dots; q, t, s) = 0$$

unless $\mu \supset \lambda$ (meaning that $\mu_i \geq \lambda_i$ for each i ; i.e., the Young diagram of μ contains that of λ).

In addition to constructing these polynomials (which do not trivially exist, since there are typically far more equations than variables), Okounkov observed that the expansion of Koornwinder polynomials in the basis of interpolation polynomials (the “binomial formula”) was particularly nice:

$$\begin{aligned} & \frac{K_\lambda(; t_0, t_1, t_2, t_3; q, t)}{K_\lambda(\dots, t^{n-i} t_0, \dots; t_0, t_1, t_2, t_3; q, t)} \\ &= \sum_{\mu} c_\mu P_\mu(\dots, q^{\lambda_i} t^{n-i} \hat{t}_0, \dots; q, t, \hat{t}_0) P_\mu(; q, t, t_0), \end{aligned}$$

where

$$c_\mu = \frac{1}{P_\mu(\dots, q^{\mu_i} t^{n-i} \hat{t}_0, \dots; q, t, \hat{t}_0) K_\mu(\dots, t^{n-i} t_0, \dots; t_0, t_1, t_2, t_3; q, t)}.$$

Together with an explicit product formula for the value of the interpolation polynomial, the evaluation Macdonald conjecture makes c_μ explicitly invariant under the involution $t_r \mapsto \hat{t}_r$, and thus implies the symmetry conjecture. The key observation of [21] was that even without the Macdonald conjectures, if one

used this (with the principal specialization of K_λ replaced by its conjectural product representation) as a definition of a family of symmetric polynomials K_λ , then those polynomials would automatically satisfy the evaluation and symmetry conjectures. Thus to show that the Koornwinder polynomials satisfied evaluation and symmetry, it would suffice to show that these new polynomials satisfied orthogonality with respect to the Koornwinder inner product.

The main tool for this was a certain difference operator:

$$(D_q^{(n)}(t_0, t_1; t)f)(z_1, \dots, z_n) := \prod_{1 \leq i \leq n} (1 + R(z_i)) \frac{(1 - t_0 z_i)(1 - t_1 z_i)}{1 - z_i^2} \prod_{1 \leq i < j \leq n} \frac{1 - t z_i z_j}{1 - z_i z_j} f(\dots, q^{1/2} z_i, \dots),$$

where f is C_n -invariant and $R(z_i)$ is the operator that takes $z_i \mapsto 1/z_i$. (Since f is invariant, evaluating it at $q^{1/2}/z_i$ is equivalent to evaluating it at $q^{-1/2}z_i$, so this is indeed a difference operator.) This takes C_n -invariant Laurent polynomials to C_n -invariant Laurent polynomials (proof: if one clears the denominators, the result is antisymmetric under C_n , so is a multiple of the cleared denominator), and is triangular with respect to the dominance ordering on monomials. Most importantly, though it is not self-adjoint with respect to the Koornwinder inner product, it does satisfy a useful adjointness property (where $\langle \cdot \rangle$ denotes the normalized inner product, so that $\langle 1, 1 \rangle = 1$):

$$\langle f, D_q^{(n)}(t_0, t_1; t)g \rangle_{t_0, t_1, q^{1/2}t_2, q^{1/2}t_3; q, t}^{(n)} \propto \langle g, D_q^{(n)}(t_2, t_3; t)f \rangle_{t_2, t_3, q^{1/2}t_0, q^{1/2}t_1; q, t}^{(n)}$$

(Sketch of proof: The densities are symmetric, so $R(z_i)f$ and f have the same integral; we can thus replace the applications of $1 + R(z_i)$ by multiplication by 2. But then the two integrals are related by $z_i \mapsto q^{-1/2}/z_i$.) This suffices to show that these operators act nicely on the Koornwinder (orthogonal) polynomials. Moreover, the combined operator

$$D_q^{(n)}(t_0, t_1; t)D_q^{(n)}(q^{-1/2}t_2, q^{-1/2}t_3; t)$$

is self-adjoint, so has Koornwinder polynomials as eigenfunctions, with generically distinct eigenvalues. Thus to show that our new polynomials are Koornwinder polynomials, it suffices to understand the action of these operators on interpolation polynomials. And, indeed, this action is also quite nice:

$$D_q^{(n)}(s, u/s; t)P_\lambda^{*(n)}(; q, t, q^{1/2}s) = \prod_{1 \leq j \leq n} q^{-\lambda_j/2} (1 - q^{\lambda_j} t^{n-i} u) P_\lambda(; q, t, s)$$

(Sketch: If we evaluate the left-hand side at a partition, the resulting sum over values of $P_\lambda^{*(n)}$ involves only partitions (the sum is over integer vectors μ with $\lambda_i - 1 \leq \mu_i \leq \lambda_i$; if $\mu_n < 0$, the factor $1 - s/z_n$ vanishes, while if $\mu_i < \mu_{i+1}$, the factor $1 - t z_{i+1}/z_i$ vanishes). It follows that the right-hand side vanishes precisely when it is supposed to; the “eigenvalue” follows by comparing leading

monomials.) This gives the required action of $D_q^{(n)}(t_0, t_1; t)$ on Koornwinder polynomials (defined via the binomial formula), and the action of $D_q^{(n)}(t_2, t_3; t)$ follows once one shows that K_λ is invariant under permutations of t_0, t_1, t_2, t_3 ; this in turn follows once one understands how the different bases of interpolation polynomials are related.

To generalize this argument to the elliptic level, we will thus need an elliptic analogue of interpolation polynomials. Here, though, we immediately encounter difficulties: the notion of “leading monomial” is essentially meaningless at the elliptic level. This is particularly problematical, since this notion also plays a significant role in the definition of Koornwinder polynomials.

We also encounter difficulties if we try to generalize the difference operator. Indeed, the fact that $\theta_0(x) = 1 - x$ suggests an obvious analogue of the difference operator:

$$\prod_{1 \leq i \leq n} (1 + R(z_i)) \frac{\theta_p(t_0 z_i) \theta_p(t_1 z_i)}{\theta_p(z_i^2)} \prod_{1 \leq i < j \leq n} \frac{\theta_p(t z_i z_j)}{\theta_p(z_i z_j)} f(\dots, q^{1/2} z_i, \dots).$$

Unfortunately, this operator does not act nicely on elliptic functions. Indeed, the problem is that though $1 - z^2$ has only two zeros, the theta function $\theta_p(z^2)$ has *four* zeros (modulo p), namely ± 1 and $\pm \sqrt{p}$. This is easily fixed, however: just add two factors to the numerator. We are thus led to consider the following difference operator:

$$(D_q^{(n)}(t_0, t_1, t_2, t_3; t; p)f)(z_1, \dots, z_n) := \prod_{1 \leq i \leq n} (1 + R(z_i)) \frac{\theta_p(t_0 z_i, t_1 z_i, t_2 z_i, t_3 z_i)}{\theta_p(z_i^2)} \prod_{1 \leq i < j \leq n} \frac{\theta_p(t z_i z_j)}{\theta_p(z_i z_j)} f(\dots, q^{1/2} z_i, \dots),$$

which does act nicely on suitable spaces of theta functions. Fix a complex number p with $0 < |p| < 1$, and define a BC_n -symmetric theta function (of degree m) to be a function $f(z_1, \dots, z_n)$ with the following properties:

- 1. f is a holomorphic function on $(\mathbb{C} \setminus 0)^n$, invariant under the action of C_n .
- 2. f is quasiperiodic in each variable:

$$f(pz_1, z_2, \dots, z_n) = (pz_1^2)^{-m} f(z_1, z_2, \dots, z_n).$$

These span an $\binom{n+m}{m}$ -dimensional space, which we note is the same as the number of partitions $\lambda \subset m^n$ (i.e., partitions with at most n parts, each at most m). We similarly define BC_n -symmetric elliptic functions by allowing meromorphic functions but requiring periodicity ($m = 0$).

Proposition 1. *If $t^{n-1}t_0t_1t_2t_3 = q^{-m}p$, then $D_q^{(n)}(t_0, t_1, t_2, t_3; t; p)$ preserves the space of BC_n -symmetric theta functions of degree m .*

Proof. The image is certainly hyperoctahedrally invariant, and each term has the same quasiperiodicity properties, so it remains only to show the result is holomorphic; but this follows by symmetry considerations as before. \square

Remark. Note that if one solves for t_3 , one can then take a limit $p \rightarrow 0$ and obtain a difference operator preserving the space of C_n -symmetric Laurent polynomials of degree at most m in each variable. The analogue of the proposition is somewhat complicated to prove directly, however, as the degree would naturally increase to $m + 1$.

When given an operator on a space of functions, one is naturally led to consider the eigenvalues and eigenfunctions. Unfortunately, these appear to be badly behaved in general (certainly, no explicit formulas are known for either). However, we may note that the difference equation for the interpolation polynomials is not itself an eigenvalue equation (indeed, Okounkov showed in an appendix to [18] that no such equation exists). Instead, we have in that case a family of operators satisfying a generalized eigenvalue problem: as u varies, the images

$$D_q^{(n)}(q^{-1/2}s, u; t)P_\lambda^{*(n)}(; q, t, s)$$

are all proportional to each other. Experimentation (with the $p \rightarrow 0$ limit) shows that if we consider two instances of the elliptic operators with two parameters in common, the resulting generalized eigenvalue problem has nice (i.e., having explicit product formulas) eigenvalues, and the corresponding eigenfunctions depend only on the common parameters. More precisely, one has the following.

Theorem 2. *For generic $a, b, q, t \in \mathbb{C} \setminus 0$, there is a basis $P_\lambda^{*(m,n)}(; a, b, q, t)$ (where λ ranges over partitions with $\lambda \subset m^n$) of the space of BC_n -symmetric theta functions satisfying the condition that*

$$\frac{D_q^{(n)}(q^{-1/2}a, q^{-1/2}b, q^{1/2}c, q^{1/2}p/q^m t^{n-1}abc; t; p)}{\prod_{1 \leq i \leq n} \theta_p(q^{\lambda_i} t^{n-i} ac, q^{m-\lambda_i} t^{i-1} bc)} P_\lambda^{*(m,n)}(; a, b, q, t; p)$$

is independent of c . One can moreover normalize the bases so that

$$\begin{aligned} & D_q^{(n)}(q^{-1/2}a, q^{-1/2}b, q^{1/2}c, \frac{q^{1/2}p}{q^m t^{n-1}abc}; t; p) P_\lambda^{*(m,n)}(; a, b, q, t; p) \\ &= \prod_{1 \leq i \leq n} \theta_p(q^{m-1} t^{n-i} ab, q^{\lambda_i} t^{n-i} ac, q^{m-\lambda_i} t^{i-1} bc) P_\lambda^{*(m,n)}(; q^{-1/2}a, q^{-1/2}b; q, t; p). \end{aligned}$$

Remark. We normalize the functions so that

$$\begin{aligned} & P_\lambda^{*(m,n)}(\dots, t^{n-i}v, \dots; a, b, q, t; p) \\ &= \prod_{1 \leq i \leq n} \prod_{1 \leq j \leq \lambda_i} \theta_p(t^{n-i} q^{j-1} av, t^{1-i} q^{j-1} a/v) \prod_{\lambda_i < j \leq m} \theta_p(t^{i-1} q^{m-j} bv, t^{i-n} q^{m-j} b/v) \end{aligned}$$

for all v ; one can show from the difference equation that if the function satisfies the normalization for one point v , it satisfies it for qv , and thus by analytic continuation for all v .

Remark. This implies (by comparing actions on generalized eigenfunctions) the following commutation relation for the difference operators (which can also be proved directly):

$$D_q^{(n)}(a, b, c, d; q, t; p)D_q^{(n)}(q^{1/2}a, q^{1/2}b, q^{-1/2}e, q^{-1/2}f; q, t; p) \\ = D_q^{(n)}(a, b, e, f; q, t; p)D_q^{(n)}(q^{1/2}a, q^{1/2}b, q^{-1/2}c, q^{-1/2}d; q, t; p),$$

whenever $cd = ef$. Unlike the case of commuting operators, however, this does not immediately imply the existence of joint generalized eigenfunctions.

Supposing for the moment that this theorem holds, we find that the resulting eigenfunctions (the *interpolation theta functions*) satisfy an analogue of the vanishing property of Okounkov’s interpolation polynomials.

Theorem 3. *The interpolation theta functions satisfy the vanishing condition*

$$P_\lambda^{*(m,n)}(\dots, q^{\mu_i}t^{n-i}a, \dots; a, b, q, t; p) = 0$$

for any partition μ with $\lambda \not\subseteq \mu$.

Remark. Note that though μ must have at most n parts for this to make sense, we do *not* constrain the parts to be at most m . Note thus that one can analytically continue in any unconstrained parts. In other words, any part before the last part with $\mu_i < \lambda_i$ can be replaced by a variable.

Since the operators (and the normalization) are symmetric between a and b , we also find that $P_\lambda^{*(m,n)}$ satisfies a symmetry:

$$P_\lambda^{*(m,n)}(; a, b, q, t; p) = P_{m^n - \lambda}^{*(m,n)}(; b, a, q, t; p),$$

where the complementary partition $m^n - \lambda$ is defined by

$$(m^n - \lambda)_i = m - \lambda_{n+1-i}.$$

(Take the complement of the Young diagram of λ inside the rectangle m^n , then rotate by 180 degrees.) The interpolation theta functions thus also satisfy a vanishing condition

$$P_\lambda^{*(m,n)}(\dots, q^{\mu_i}t^{n-i}b, \dots; a, b, q, t; p) = 0$$

for all partitions μ with $m^n - \lambda \not\subseteq \mu$. Note that the limit which takes the above difference equation to that satisfied by Okounkov’s polynomials involves taking $p \rightarrow 0$ with $b \sim p^{1/2}$, and thus the second set of vanishing conditions moves off to infinity, and essentially (in a way not yet fully understood) becomes the constraint on the dominant monomial.

There are two main strategies for proving these two theorems. One strategy is to mimic Okounkov’s construction by giving an explicit construction for the interpolation functions as $n(n + 1)/2$ -dimensional contour integrals, which

requires only understanding the commutation relations between the above difference operators and the (raising) integral operators of [24] to obtain the generalized eigenvalue equations. One then observes that the first theorem implies the second. Indeed, one again finds that evaluating the image of the difference operator at a partition resolves to a sum of terms, in each of which the interpolation function is evaluated at a partition (differing from the original by a vertical strip; i.e., subtracting at most 1 from each part). One thus obtains linear combinations (with explicit, nice coefficients) of the form

$$\sum_{\nu \subset \mu} f_{\lambda, \mu, \nu}(c) P_{\lambda}^{*(m, n)}(\dots, q^{\nu_i} t^{n-i} a; a, b, q, t; p)$$

which are guaranteed by the generalized eigenvalue equation to be independent of c . Moreover, if $\mu \neq \lambda$, one finds that the coefficient $f_{\lambda, \mu, \mu}$ indeed depends nontrivially on c . So subtracting two such expansions and dividing, we obtain an expansion of the form

$$P_{\lambda}^{*(m, n)}(\dots, q^{\mu_i} t^{n-i} a; a, b, q, t; p) = \sum_{\nu \subsetneq \mu} c_{\nu} P_{\lambda}^{*(m, n)}(\dots, q^{\nu_i} t^{n-i} a; a, b, q, t; p).$$

If we start with a partition not containing λ , then the terms on the right all involve strictly smaller partitions not containing λ , and thus the vanishing conditions follow by induction.

One may also, as in [22], take a partially opposite approach: choose a suitable subset of the vanishing conditions guaranteed to (generically) have a unique solution, and show that these conditions are preserved by the difference operators, so that the solutions satisfy the difference equations; the remaining vanishing conditions then follow as above.

There are several nice special cases of the interpolation theta functions. If $n = 1$, the vanishing conditions determine all of the zeros, and we thus obtain a product expression for $P^{*(m, 1)}$:

$$P_l^{*(m, 1)}(z; a, b, q, t; p) = \prod_{0 \leq j < l} \theta_p(q^j a z^{\pm 1}) \prod_{0 \leq j < m-l} \theta_p(q^j b z^{\pm 1}).$$

If $t = 1$, the difference equation acts independently on each variable, so $P_{\lambda}^{*(m, n)}$ is proportional to the symmetrization of $\prod_i P_{\lambda_i}^{*(m, 1)}(z_i)$. Similarly, if $t = q$, then the difference equation has a sufficiently nice determinantal structure, so that $P_{\lambda}^{*(m, n)}$ satisfies an analogue of the Weyl character formula:

$$P_{\lambda}^{*(m, n)}(z_1, \dots, z_n; a, b, q, q; p) \propto \frac{\det_{1 \leq i, j \leq n} P_{\lambda_i + n - i}^{*(m+n-1, 1)}(z_j; a, b, q, q; p)}{\prod_{1 \leq i < j \leq n} z_i^{-1} \theta_p(z_i z_j^{\pm 1})}.$$

If $q^m t^{n-1} a b = 1$, then the two vanishing conditions both involve evaluation at partitions based at a , and one thus finds that

$$P_{\lambda}^{*(m, n)}(\dots, q^{\mu_i} t^{n-i} a, \dots; a, b, q, t; p) \propto \delta_{\lambda \mu}.$$

There is also a somewhat unexpected special case: if $q^{m+1}t^n ab = pq$, then one finds that the function

$$\prod_{1 \leq i \leq n, 1 \leq j \leq m} \theta_p(at^{n-\lambda'_j} q^{j-1} z_i^{\pm 1})$$

satisfies all of the required vanishing conditions, and thus $P_\lambda^{*(m,n)}$ has this factored form. Indeed, for the product not to vanish at $z_i = q^{\mu_i} t^{n-i} a$, one must have either $\lambda'_j \neq i$ or $\mu_i \neq j - 1$ for each i, j , and thus $\mu_{\lambda'_j} \neq j - 1$ for each j . Incrementing j decreases $\mu_{\lambda'_j} + 1 - j$ by at most 1; since $\mu_{\lambda'_1} > 0$, and the sequence is not allowed to hit 0, it must remain positive, and thus $\mu_{\lambda'_j} \geq j$ for all j . In particular, every corner of the diagram of λ is contained in the diagram of μ , and thus $\lambda \subset \mu$. (Combinatorially, this argument is equivalent to that of a lemma in Okounkov’s proof of the Cauchy identity for interpolation polynomials, so we call this the Cauchy case; it also relates to a Cauchy identity for interpolation theta functions.) A final key special case is when $q^{m+1}t^{n-1}ab = 1$. Since

$$P_\lambda^{*(m,n)}(; a, b; q, t; p) \propto D^{(n)}(a, b, c, d; q, t; p) P_\lambda^{*(m,n)}(; q^{1/2}a, q^{1/2}b; q, t; p),$$

and the interpolation theta functions on the right are delta functions, we find that if we evaluate the function on the left at a partition, the resulting sum on the right collapses to a single term. We thus find that for a partition μ ,

$$P_\lambda^{*(m,n)}(\dots, q^{\mu_i} t^{n-i} a, \dots; a, 1/q^{m+1} t^{n-1} a; q, t; p)$$

vanishes unless $\lambda_i \leq \mu_i \leq \lambda_i + 1$ for each i , and the nonzero values are essentially just the coefficients in the difference equation.

Since the interpolation theta functions are all quasiperiodic with the same multiplier, any ratio of interpolation theta functions will be an elliptic function. It turns out that if we divide by $P_0^{*(m,n)}$, the remaining dependence on m is easy to remove (and the introduced denominator is nice), as follows from the identity

$$\begin{aligned} P_\lambda^{*(m+1,n)}(\dots, z_i, \dots; a, b; q, t; p) \\ = \prod_{1 \leq i \leq n} \theta_p(bz_i, b/z_i) P_\lambda^{*(m,n)}(\dots, z_i, \dots; a, qb; q, t; p). \end{aligned}$$

(Both sides easily satisfy the same difference equations and normalization.) We may thus define the *elliptic interpolation functions* by

$$R_\lambda^{*(n)}(; a, b; q, t; p) := \frac{P_\lambda^{*(m,n)}(; a, q^{-m}b; q, t; p)}{P_0^{*(m,n)}(; a, q^{-m}b; q, t; p)}$$

for sufficiently large m . These are elliptic functions of the variables, and quasiperiodic in a and b (since the difference operators are quasiperiodic). In

addition to the above symmetries, we note that

$$R_{\lambda+1^n}^{*(n)}(\dots, z_i, \dots; a, b; q, t; p) = \prod_{1 \leq i \leq n} \frac{\theta_p(az_i^{\pm 1})}{\theta_p(bz_i^{\pm 1}/q)} R_{\lambda}^{*(n)}(\dots, z_i, \dots; qa, b/q; q, t; p),$$

where $(\lambda + 1^n)_i := \lambda_i + 1$, and

$$R_{\lambda}^{*(n+k)}(\dots, z_i, \dots, a, ta, \dots, t^{k-1}a; a, b; q, t; p) \propto R_{\lambda}^{*(n)}(\dots, z_i, \dots; t^k a, b; q, t; p),$$

again since the transformations preserve the difference equations. The constant of proportionality in the latter equation follows from the case $v = t^{-n}a$ of the formula

$$R_{\lambda}^{*(n)}(\dots, t^{n-i}v, \dots; a, b; q, t; p) = \Delta_{\lambda}^0 \left(\frac{t^{n-1}a}{b} \mid t^{n-1}av, \frac{a}{v}; q, t; p \right),$$

where in general one defines

$$\begin{aligned} \Delta_{\lambda}^0(a \mid b_1, \dots, b_m; q, t; p) &= \prod_{1 \leq k \leq m} \prod_{1 \leq i} \prod_{1 \leq j \leq \lambda_i} \frac{\theta_p(q^{j-1}t^{1-i}b_k)}{\theta_p(q^{-j}t^{i-1}b_k/a)} \\ &= \Delta_{\lambda}^0(a \mid \frac{pqa}{b_1}, \dots, \frac{pqa}{b_m}; q, t; p)^{-1} \end{aligned}$$

It will be notationally convenient to define renormalized interpolation functions

$$R_{\lambda}^{*(n)}(; a, b(v); q, t; p) := \frac{R_{\lambda}^{*(n)}(; a, b; q, t; p)}{R_{\lambda}^{*(n)}(\dots, t^{n-i}v, \dots; a, b; q, t; p)}.$$

A key property of the interpolation functions is that the corresponding change of basis matrices are particularly nice.

Proposition 4. *There exist elliptic binomial coefficients $\binom{\lambda}{\mu}_{[a,b];q,t;p}$ such that*

$$R_{\lambda}^{*(n)}(; a, b(a'); q, t; p) = \sum_{\mu} \binom{\lambda}{\mu}_{[t^{n-1}a/b, a/a'];q,t;p} R_{\mu}^{*(n)}(; a', b(a); q, t; p)$$

That is, the given connection (change of basis) coefficients are invariant under simultaneous rescaling of a, a', b .

Proof. (Sketch) The difference operator $D_q^{(n)}(a, a', b, p/t^{n-1}aa'b; t; p)$ acts on both sides, and has the effect of rescaling a, a', b by $q^{1/2}$ without changing the connection coefficients. The claim follows by analytic continuation. □

Remark. The corresponding coefficients for interpolation polynomials are essentially just principal specializations of skew Macdonald polynomials.

Connection coefficients for interpolation theta functions follow immediately, and by observing that the b -based vanishing conditions are the same, we may conclude that $\binom{\lambda}{\mu} = 0$ unless $\mu \subset \lambda$. More significantly, when $t^{n-1}a'b = 1$, the resulting interpolation theta functions are delta functions, so we can compute the change of basis coefficients in that case as values of interpolation functions. But since the connection coefficients are invariant under rescaling, we can *always* arrange to be in this case. We thus find (after mild reparametrization) that

$$\binom{\lambda}{\mu}_{[t^{n-1}a/q^m b, t^{n-1}a q^m b]; q, t; p} = f_\lambda f'_\mu P_{m^n - \lambda}^{*(m, n)}(\dots, q^{(m^n - \mu)_i} t^{n-i} b, \dots; b, a, q, t; p)$$

for suitable factors f_λ, f'_μ (depending on the parameters as well as the specified partition, and having explicit product formulas).

Note that we have the inversion identity

$$\sum_{\mu \subset \kappa \subset \lambda} \binom{\lambda}{\kappa}_{[a, b]; q, t; p} \binom{\kappa}{\mu}_{[a/b, 1/b]; q, t; p} = \delta_{\lambda \mu},$$

since the inverse connection coefficients have the same form as the original connection coefficients. Less trivially, if we change basis from $R^{*(n)}(; a, b)$ to $R^{*(n)}(; a'', b)$ via $R^{*(n)}(; a', b)$, and compare to the direct change of basis coefficients, we obtain the following identity (with $c = a'/a, c' = a''/a$, and other reparametrizations):

$$\begin{aligned} & \frac{\Delta_\lambda^0(b|dc, c'; q, t; p)}{\Delta_\mu^0(b/c'|dc, 1/c'; q, t; p)} \binom{\lambda}{\mu}_{[b, c']; q, t; p} \\ &= \frac{\Delta_\lambda^0(b|dc', c; q, t; p)}{\Delta_\mu^0(b/c'|d, c/c'; q, t; p)} \\ & \sum_{\mu \subset \kappa \subset \lambda} \frac{\Delta_\kappa^0(b/c|d, c'/c; q, t; p)}{\Delta_\kappa^0(b/c|dc', 1/c; q, t; p)} \binom{\lambda}{\kappa}_{[b, c]; q, t; p} \binom{\kappa}{\mu}_{[b/c, c'/c]; q, t; p} \end{aligned}$$

(In particular, the left-hand side becomes a delta function in the limit $c' \rightarrow 1$, i.e., $a'' \rightarrow a$.) This has a hidden symmetry (actually two, one reflected in the fact that n has disappeared from the equation), in that replacing d by $pqb/cc'd$ leaves the sum invariant.

Remark. Note that when λ has only one part, $\binom{\lambda}{\mu}$ can be computed using only univariate interpolation functions, and in particular factors nicely. The sum in this case becomes (elliptic) hypergeometric, and is in fact precisely Frenkel and Turaev’s summation identity.

Lemma 5. *The binomial coefficients satisfy a complementation identity of the form*

$$\binom{\lambda}{\mu}_{[a, b]; q, t; p} = c_\lambda c'_\mu \binom{m^n - \mu}{m^n - \lambda}_{[q^{-2m} t^{2n-2} b/a, b]; q, t; p},$$

where c, c' are explicit factors depending only on the parameters and the specified partitions.

Proof. (Sketch) The above summation identity is consistent with this complementation (with the appropriate explicit factors), so if complementation holds with second parameter b and second parameter b' , it holds when the second parameter is bb' . So it will suffice to prove this when $b = 1/q$. But in that case, the expression for the binomial coefficient as a value of an interpolation function can be further reinterpreted as a coefficient of the difference equation. But the form of the difference equation ensures that its coefficients have a symmetry precisely of the above form. □

Remark. There is a slight lie involved here, since the binomial coefficient as defined is actually singular when $b = 1/q$; but eliminating the singularity is straightforward.

Remark. It follows from the definition and the vanishing conditions that $\binom{\lambda}{0} = 1$; complementation symmetry then gives a nice formula for $\binom{m^n}{\lambda}$. If we take $\lambda = m^n, \kappa = 0$ in the iterated connection coefficient identity, the result is a hypergeometric-type sum (in the sense that the ratios of terms corresponding to adjacent partitions are elliptic functions of q_i^λ), originally conjectured by Warnaar [32] and proved by Rosengren [26]. One of the first indications that the approach of [21] had an elliptic analogue was the fact that it gave a proof of the $p \rightarrow 0$ limit of this identity.

It follows that one has an expression of the form

$$R_\mu^{*(n)}(\dots, q^{\lambda_i} t^{n-i} a, \dots; a, b, q, t; p) = \frac{\binom{\lambda}{\mu}_{[t^{2n-2}a^2, t^{n-1}ab]; q, t; p}}{\Delta_\mu(t^{n-1}a/b|t^n, 1/t^{n-1}ab; q, t; p)},$$

where Δ_μ denotes a suitable product of theta functions; by convention, we extend the notation so that

$$\Delta_\mu(a|b_1, \dots, b_m; q, t; p) = \Delta_\mu(a|; q, t; p)\Delta_\mu^0(a|b_1, \dots, b_m; q, t; p);$$

for the explicit form of Δ_μ (which, roughly speaking, is the residue of the elliptic Selberg integral at a partition), see the introduction to either [24] or [22].

If we use this to rewrite complementation symmetry in terms of interpolation functions, we obtain a (dense) special case of the following familiar-looking identity:

$$\frac{R_\lambda^{*(n)}(\dots, q^{\mu_i} t^{n-i} c, \dots; a, b, q, t; p)}{R_\lambda^{*(n)}(\dots, t^{n-i} c, \dots; a, b, q, t; p)} = \frac{R_\mu^{*(n)}(\dots, q^{\lambda_i} t^{n-i} \hat{c}, \dots; \hat{a}, \hat{b}; q, t; p)}{R_\mu^{*(n)}(\dots, t^{n-i} \hat{c}, \dots; \hat{a}, \hat{b}; q, t; p)},$$

where

$$\hat{a} = c\sqrt{t^{n-1}ab}, \quad \hat{b} = ab/\hat{a}, \quad \hat{c} = ac/\hat{a}.$$

In other words, the interpolation functions satisfy an analogue of Macdonald’s symmetry conjecture. (In contrast, Okounkov’s polynomials do *not* satisfy such a symmetry; more precisely, if one takes a limit so that the left-hand side involves Okounkov’s polynomials, the functions on the right become a new family of rational functions (in which the hyperoctahedral symmetry is broken).)

Other identities involve a Pieri-type identity of the form

$$\prod_{1 \leq i \leq n} \frac{\theta_p(vz_i^{\pm 1})}{\theta_p((pq/b)z_i^{\pm 1})} R_\lambda^{*(n)}(\dots, z_i, \dots; a, b/q; q, t; p) = \sum_{\kappa} f_\kappa f'_\lambda \binom{\lambda + 1^n}{\kappa} R_\lambda^{*(n)}(\dots, z_i, \dots; a, b; q, t; p)_{[t^{n-1}a/b, 1/q]}$$

(Use connection coefficients to reduce to the case $v = a$), and a branching rule of the form

$$R_\lambda^{*(n+1)}(\dots, z_i, \dots, v; a, b; q, t; p) = \sum_{\kappa} f_\kappa f'_\lambda \binom{\lambda}{\kappa} R_\kappa^{*(n)}(\dots, z_i, \dots; a, b; q, t; p)_{[t^n a/b, t]; q, t; p}$$

The coefficients of the Pieri identity factor (as they are coefficients of the difference equation), while the coefficients of the branching rule *also* factor, by virtue of the following additional symmetry:

$$\binom{\lambda}{\mu}_{[a, b]; q, t; p} = \binom{\lambda'}{\mu'}_{[qta, b]; 1/t, 1/q; p};$$

this follows from the case $b = pq/t$, when the interpolation functions factor as mentioned above. (It should be noted that the elliptic interpolation functions were independently constructed by Coskun and Gustafson [5] using the branching rule in the explicit factored form.) One also has a Cauchy-type identity, obtained by expanding the Cauchy case of the interpolation functions via the connection coefficients, then analytically continuing. There is also a transformation version of the main summation identity above, an analogue of a hypergeometric transformation conjectured by Warnaar, generalizing the Frenkel-Turaev transformation.

As an aside, we note that one can use the branching rule to express the interpolation function as a (complicated) sum over semistandard Young tableaux. In the special case

$$R_\lambda^{*(n)}(vt^{n-1}, \dots, v; a, b(v); q, t; p) = 1,$$

this can be interpreted, at least formally, as giving a family of elliptic probability distributions on the set of semistandard Young tableaux of shape λ . When λ is a rectangle, a standard bijection allows us to interpret this as a (formal) probability distribution on plane partitions in a rectangular box. Finally,

when $q = t$, the sum greatly simplifies, and one obtains the following elliptic MacMahon identity [2]:

$$\sum_{\Pi \subset a \times b \times c} \prod_{(i,j,k) \in \Pi} \frac{q^3 \theta_p(q^{j+k-2i-1}u_1, q^{i+k-2j-1}u_2, q^{i+j-2k-1}u_3)}{\theta_p(q^{j+k-2i+1}u_1, q^{i+k-2j+1}u_2, q^{i+j-2k+1}u_3)}$$

$$= \prod_{1 \leq i \leq a, 1 \leq j \leq b, 1 \leq k \leq c} \frac{q \theta_p(q^{i+j+k-1}, q^{j+k-i-1}u_1, q^{i+k-j-1}u_2, q^{i+j-k-1}u_3)}{\theta_p(q^{i+j+k-2}, q^{j+k-i}u_1, q^{i+k-j}u_2, q^{i+j-k}u_3)},$$

where $u_1 u_2 u_3 = 1$, and the sum is over plane partitions contained in an $a \times b \times c$ box.

3. Biorthogonal Functions

The first key observation in the construction of the multivariate biorthogonal functions is that the difference operators satisfy a suitable adjointness property with respect to the elliptic Selberg integral (1) (where we recall the balancing condition $t^{2n-2}t_0t_1t_2t_3t_4t_5 = pq$):

$$\langle f, D_q^{(n)}(t_0, t_1, t_2, p/t^{n-1}t_0t_1t_2)g \rangle_{t_0, t_1, t_2, q^{1/2}t_3, q^{1/2}t_4, q^{1/2}t_5; t; p, q}^{(n)}$$

$$\propto \langle D_q^{(n)}(t_3, t_4, t_5, p/t^{n-1}t_4t_4t_5)f, g \rangle_{q^{1/2}t_0, q^{1/2}t_1, q^{1/2}t_2, t_3, t_4, t_5; t; p, q}^{(n)}$$

Here the constant of proportionality can be obtained from the fact that

$$D_q^{(n)}(t_0, t_1, t_2, p/t^{n-1}t_0t_1t_2)1 = \prod_{1 \leq i \leq n} \theta_p(t^{n-i}t_0t_1, t^{n-i}t_0t_2, t^{n-i}t_1t_2);$$

the left-hand side is constant by Proposition 1, and becomes the right-hand side when $z_i = t^{n-i}t_0$. In particular, taking $f = g = 1$ gives a recurrence for the normalization under certain q -shifts of the parameters; the symmetry of the density with respect to swapping p and q gives an analogous recurrence involving p -shifts, thus evaluating the elliptic Selberg integral up to a constant, which can be determined via a suitable limit.

Since the difference operators act nicely on interpolation functions, we can generalize the above computation to give the following. (The original proof can be found in [24].)

Lemma 6. *The inner product of interpolation functions with respect to the elliptic Selberg integral is given by the following formula:*

$$\langle R_\lambda^{*(n)}(; t_0, u_0; q, t; p), R_\mu^{*(n)}(; t_1, u_1; q, t; p) \rangle_{t_0, t_1, t_2, t_3, u_0, u_1; q, t; p}^{(n)}$$

$$= \Delta_\lambda^0(t^{n-1}t_0/u_0 | t^{n-1}t_0t_2, t^{n-1}t_0t_3, t^{n-1}t_0t_1, t^{n-1}t_0u_1)$$

$$\Delta_\mu^0(t^{n-1}t_1/u_1 | t^{n-1}t_1t_2, t^{n-1}t_1t_3, t^{n-1}t_1t_0, t^{n-1}t_1u_0)$$

$$R_\lambda^{*(n)}(\dots, q^{\mu_i}t^{n-i}t_1/v, \dots; t_0v, u_0v(t_1/v); q, t; p),$$

where $v^2 = t^{n-1}t_1u_1$.

Proof. (Sketch) Adjointness shows that both sides transform the same way under multiplying t_0, u_0, t_2 by $q^{1/2}$ and dividing t_1, u_1, t_3 by the same, and thus, composing two such operations, under the shift $(t_2, t_3) \mapsto (qt_2, t_3/q)$. Using q -theta p -difference operators, we find similarly that $(t_2, t_3) \mapsto (pt_2, t_3/p)$ preserves the relation between the two sides. It follows that the ratio of the two sides depends on t_2 and t_3 only via their product. Moreover, it follows via the connection coefficient formula that if the identity holds when $t_2t_3 = x$, it also holds whenever $t_0t_2 = x$. It therefore suffices to prove the identity for one value of x . Letting t_2 and t_3 approach the unit circle in such a way that $t_2t_3 \rightarrow 1$, we find that the integral becomes singular, but the normalized integral can be evaluated via residue calculus: the result is an $n - 1$ -dimensional elliptic Selberg integral (with t_2, t_3 both multiplied by t) in which the last variable of each interpolation function has been specialized to t_2 . But this can be evaluated using the branching rule and induction. \square

With this in mind, we can readily construct biorthogonal functions. The point is that the above matrix of inner products can be written using connection coefficients as a product of triangular matrices with explicit inverses.

Theorem 7. *Define the (multivariate) elliptic biorthogonal functions by the binomial formula*

$$\begin{aligned} &\tilde{R}_\lambda^{(n)}(; t_0:t_1, t_2, t_3; u_0, u_1; q, t; p) \\ &= \sum_{\mu \subset \lambda} \frac{\binom{\lambda}{\mu}_{[1/u_0u_1, 1/t^{n-1}t_0u_1]; q, t; p} R_\mu^{*(n)}(; t_0, u_0; q, t; p)}{\Delta_\mu^0(t^{n-1}t_0/u_0 | t^{n-1}t_0t_1, t^{n-1}t_0t_2, t^{n-1}t_0t_3, t^{n-1}t_0u_1; q, t; p)}. \end{aligned}$$

These are normalized so that

$$\tilde{R}_\lambda^{(n)}(\dots, t^{n-i}t_0, \dots; t_0:t_1, t_2, t_3; u_0, u_1; q, t; p) = 1,$$

and are biorthogonal with respect to the elliptic Selberg integral, in that

$$\langle \tilde{R}_\lambda^{(n)}(; t_0:t_1, t_2, t_3; u_0, u_1; q, t; p), \tilde{R}_\mu^{(n)}(; t_1:t_0, t_2, t_3; u_1, u_0; q, t; p) \rangle_{t_0, t_1, t_2, t_3, u_0, u_1; q, t; p}^{(n)}$$

vanishes unless $\lambda = \mu$, when the integral is

$$\frac{1}{\Delta_\lambda(1/u_0u_1 | t^n, t^{n-1}t_1t_0, 1/t^{n-1}t_0u_1, 1/t^{n-1}t_1u_0; q, t; p)}.$$

Proof. (Sketch) We can use the lemma to compute the inner product of the biorthogonal function with $R_\mu^{*(n)}(; t_1, u_1; q, t; p)$ by expanding then integrating term by term. The resulting sum can be simplified using the connection coefficient formula to obtain

$$R_\lambda^{*(n)}(\dots, q^{\mu_i}t^{n-i}t_1/v, \dots; t_1/v, u_0v(t_0v); q, t; p),$$

up to a factor involving only μ . But this value of the interpolation function is

$$\frac{\binom{\mu}{\lambda}_{[t^{n-1}t_1/u_1, t^{n-1}t_1u_0]; q, t; p}}{\Delta_\lambda(1/u_0u_1|t^n, t^{n-1}t_1t_0, 1/t^{n-1}t_0u_1, 1/t^{n-1}t_1u_0; q, t; p)},$$

and these binomial coefficients are inverse to those in the expansion of the dual biorthogonal functions. □

In particular, the norms can be given as an explicit product. As mentioned above, the other two Macdonald conjectures are essentially trivial from this expansion.

Theorem 8. *The elliptic biorthogonal functions satisfy the evaluation identity*

$$\begin{aligned} &\tilde{R}_\lambda^{(n)}(\dots, t^{n-i}t_1, \dots; t_0:t_1, t_2, t_3; u_0, u_1; q, t; p) \\ &= \Delta_\lambda^0(1/u_0u_1|t^{n-1}t_1t_2, t^{n-1}t_1t_3, 1/t^{n-1}t_1u_1, pqt^{n-1}t_0/u_0; q, t; p), \end{aligned}$$

and the symmetry property

$$\begin{aligned} &\tilde{R}_\lambda^{(n)}(\dots, q^{\mu_i}t^{n-i}t_0, \dots; t_0:t_1t_2, t_3; u_0, u_1; q, t; p) \\ &= \tilde{R}_\mu^{(n)}(\dots, q^{\lambda_i}t^{n-i}\hat{t}_0, \dots; \hat{t}_0:\hat{t}_1\hat{t}_2, \hat{t}_3; \hat{u}_0, \hat{u}_1; q, t; p), \end{aligned}$$

where

$$\begin{aligned} \hat{t}_0 &= \sqrt{\frac{t_0t_1t_2t_3}{pq}}, & \hat{t}_0\hat{t}_1 &= t_0t_1, & \hat{t}_0\hat{t}_2 &= t_0t_2, & \hat{t}_0\hat{t}_3 &= t_0t_3, \\ & & \frac{\hat{u}_0}{\hat{t}_0} &= \frac{u_0}{t_0}, & \frac{\hat{u}_1}{\hat{t}_0} &= \frac{u_1}{t_0} \end{aligned}$$

Moreover, the nonzero values of the inner products of biorthogonal functions can be expressed as explicit products of theta functions.

Proof. Symmetry is immediate: plug the appropriate point into the expansion for the biorthogonal function, rewrite the interpolation function as a binomial coefficient, then swap the two binomial coefficients. Evaluation is only slightly more complicated: this special case of the sum is just the case $\mu = 0$ of our main binomial coefficient summation. □

Other properties of these functions are also straightforward. For instance, they are clearly symmetrical in t_1, t_2, t_3 , but are also nearly symmetrical under swapping t_0 and t_1 :

$$\tilde{R}_\lambda^{*(n)}(; t_1:t_0, t_2, t_3; u_0, u_1; q, t; p) = \frac{\tilde{R}_\lambda^{*(n)}(; t_0:t_1, t_2, t_3; u_0, u_1; q, t; p)}{\tilde{R}_\lambda^{*(n)}(\dots, t^{n-i}t_1, \dots; t_0:t_1, t_2, t_3; u_0, u_1; q, t; p)},$$

as follows from an application of the connection coefficient formula and the main binomial coefficient summation. In addition, since we can invert the binomial

coefficients, we can also expand the interpolation functions in terms of biorthogonal functions. Since the interpolation functions are independent of t_1, t_2, t_3 , we can thus obtain a formula for connection coefficients of biorthogonal functions in which one changes those three parameters arbitrarily; these are expressed as sums over partitions involving two binomial coefficients. (An analogous formula for Askey-Wilson polynomials was given in [1]; the corresponding formula for Koornwinder polynomials was new.) There are two important special cases. If one changes only one parameter, the resulting sum can be simplified via the main identity to obtain a single binomial coefficient (up to some explicit Δ^0 factors). Changing two parameters can then be done in multiple ways (one parameter at a time, or via interpolation functions), and comparing the different formulas gives the main binomial coefficient transformations.

The biorthogonal functions also satisfy difference equations, of which the principal example is

$$D^{(n)}(u_0, t_0, t_1, p/t^{n-1}t_0t_1u_0; q, t; p) \tilde{R}_\lambda^{(n)}(; q^{1/2}t_0; q^{1/2}t_1, q^{-1/2}t_2, q^{-1/2}t_3; q^{1/2}u_0, q^{-1/2}u_1; q, t; p) \propto \tilde{R}_\lambda^{(n)}(; t_0:t_1, t_2, t_3; u_0, u_1; q, t; p),$$

where the constant is independent of λ . This gives rise to several generalized eigenvalue equations relating the six operators of the form

$$D^{(n)}(u_0, t_0, t_1, \frac{p}{t^{n-1}t_0t_1u_0}; q, t; p)D^{(n)}(q^{1/2}u_0, q^{-1/2}t_2, q^{-1/2}t_3, \frac{pq^{1/2}}{t^{n-1}t_2t_3u_0}; q, t; p).$$

Note, however, that these operators all act nontrivially on u_0 and u_1 , so do not produce eigenvalue equations. One can also define lowering and raising operators. For lowering operators, if one takes $c = p/a$ in the difference equation for interpolation functions, the result vanishes unless $\lambda_n > 0$, when one can extract a factor from the image interpolation function. Simplifying, we find that (with $\lambda - 1^n$ denoting the partition obtained by subtracting 1 from each part of λ)

$$D^{(n)}(b, p/b, bq, 1/t^{n-1}bq; q, t; p)R_\lambda^{*(n)}(; q^{-1/2}a, q^{3/2}b; q, t; p) \propto R_{\lambda-1^n}^{*(n)}(; a, b; q, t; p),$$

or 0 if $\lambda_n = 0$. Plugging this in to the binomial formula, one similarly finds that

$$D^{(n)}(u_0, p/u_0, u_0q, 1/t^{n-1}u_0q; q, t; p) \tilde{R}_\lambda^{(n)}(; q^{-1/2}t_0; q^{-1/2}t_1, q^{-1/2}t_2, q^{-1/2}t_3; q^{3/2}u_0, q^{1/2}u_1; q, t; p) \propto \tilde{R}_{\lambda-1^n}^{(n)}(; t_0:t_1, t_2, t_3; u_0, u_1; q, t; p).$$

This is adjoint to an operator of the form

$$(D^{+(n)}(u_0:t_0, t_1, t_2, t_3, t_4)f)(\dots, z_i, \dots) := \prod_{1 \leq i \leq n} (1 + R(z_i)) \prod_{1 \leq i \leq n} \frac{\prod_{0 \leq r \leq 4} \theta_p(t_r z_i)}{\theta_p(z_i^2) \theta_p(pqz_i/u_0)} \prod_{1 \leq i < j \leq n} \frac{\theta_p(tz_i z_j)}{\theta_p(z_i z_j)} f(\dots, q^{1/2}z_i, \dots),$$

and thus

$$\begin{aligned}
 & D^{+(n)}(u_0:t_0, t_1, t_2, t_3, p^2q/t^{n-1}u_0t_0t_1t_2t_3) \\
 & \tilde{R}_\lambda^{(n)}(; q^{1/2}t_0:q^{1/2}t_1, q^{1/2}t_2, q^{1/2}t_3; q^{-1/2}u_0, q^{-3/2}u_1; q, t; p) \\
 & \qquad \qquad \qquad \propto \tilde{R}_{\lambda+1^n}^{(n)}(; t_0:t_1, t_2, t_3; u_0, u_1; q, t; p).
 \end{aligned}$$

In addition, there is a collection of eight integral operators (defined in [24]), of which one increases the number of variables and one decreases the number of variables, and each shifts the parameters by half-integer powers of t . The approach of [24], via an analogue of Okounkov’s integral representation for interpolation polynomials, is based on the observation that $\tilde{R}_\lambda^{(n)}$ is either an image under an integral operator of an interpolation function in $n - 1$ variables or an image under the raising difference operator of an interpolation function with a smaller index partition. Note that if one evaluates the image of an integral operator at a partition, the result is a sum over partitions differing by a horizontal strip, and the coefficients are of the form $\binom{\lambda}{\mu}_{[a,t];q,t;p}$; thus the integral operators are in this sense dual to the difference operators. Also, the raising integral operator takes interpolation functions to interpolation functions, so provides a direct analogue of Okounkov’s integral representation (the $n(n+1)/2$ -dimensional integral alluded to above).

In the limit $t_0t_1 \rightarrow q^{-m}t^{1-n}$, the continuous inner product becomes a discrete inner product concentrated at points of the form $q^\lambda t^{n-i}t_0$ with $\lambda \subset m^n$, with mass

$$\Delta_\lambda(t^{2n-2}t_0^2|t^n, t^{n-1}t_0t_1, t^{n-1}t_0t_2, t^{n-1}t_0t_3, t^{n-1}t_0u_0, t^{n-1}t_0u_1; q, t; p).$$

This can either be seen by a residue calculus argument, or by observing that the analogue of Lemma 6 is just an instance of the main summation identity. In particular, the normalization of the discrete measure is just Warnaar’s conjecture alluded to above. The transformation analogue of this sum is also the discrete form of an integral, obtained by adding two parameters to the original elliptic Selberg integral (and changing the balancing condition to $t^{2n-2}u_0u_1u_2u_3u_4u_5u_6u_7 = p^2q^2$). The resulting integral is manifestly invariant under S_8 , and satisfies an analogue of the discrete transformation; together, these generate the Weyl group of type E_7 .

Some additional observations should be made at this point. First is that it is largely impossible to avoid using the interpolation functions in defining the biorthogonal functions. Indeed, one could define the biorthogonal functions via biorthogonality and triangularity with respect to the filtration by interpolation functions, but it is unclear how to define that filtration except via evaluation at partitions. More significantly, if $t^{n-1}t_1u_1 = 1$, then the binomial formula actually turns into the connection coefficient formula for interpolation functions; in other words,

$$\tilde{R}^{(n)}(; t_0:t_1, t_2, t_3; u_0, 1/t^{n-1}t_1; q, t; p) = R^{*(n)}(; t_1, u_0(t_0); q, t; p).$$

This, of course, explains why the interpolation functions satisfy an analogue of the symmetry property: this is simply a special case of the symmetry property for biorthogonal functions. In any event, this makes it clear that the interpolation functions are an intrinsic feature of the theory of elliptic biorthogonal functions: not only are they a special case, but a natural one at that (they form a 7-dimensional subfamily of the biorthogonal functions, but effectively have only 5 degrees of freedom (a, b, q, t, p)).

In addition, as mentioned above, these functions are true generalizations of the Koornwinder polynomials. Indeed, if we take $u_0, u_1 \sim \sqrt{p}$, then let $p \rightarrow 0$, the interpolation functions become Okounkov's interpolation polynomials (e.g., by taking a limit in the branching rule), and the binomial formula for biorthogonal functions becomes Okounkov's binomial formula for Koornwinder polynomials. The above proof does not descend well, however, as the interpolation function in the inner product formula remains a rational function in the limit. Another limit of the interpolation functions is the Macdonald polynomials

$$P_\lambda(\dots, z_i, \dots; q, t) \propto \lim_{p \rightarrow 0} p^{|\lambda|/4} R_\lambda^{*(n)}(\dots, p^{1/4} z_i, \dots; a, p^{1/2} b; q, t; p).$$

Since the interpolation functions are special cases of the biorthogonal functions, this gives an interpretation of Macdonald polynomials as limits of biorthogonal functions. In fact, by taking advantage of quasiperiodicity, one can arrange to take this limit in such a way that the inner product becomes the usual inner product, and the Macdonald “conjectures” arise in the limit. (Note that the inner product for ordinary Macdonald polynomials involves the product of an ordinary polynomial with a polynomial in $1/z_i$; thus the orthogonality of Macdonald polynomials should properly be considered *b*orthogonality.) For details of this limit (among others), see [23]; for the trigonometric limit (i.e., the q -hypergeometric case), Van de Bult and the author have obtained (manuscript in preparation) a complete classification of limits of biorthogonal functions, thus generalizing the usual q -Askey scheme of q -hypergeometric orthogonal polynomials. Similar ideas give a classification of q -hypergeometric limits of higher-order elliptic Selberg integrals; for the univariate case, see [3]. In particular, one finds that a large class of q -hypergeometric sums and integrals are classified by faces of the Hesse polytope (Gosset type 3_{21}), in such a way that the action of $W(E_7)$ on this polytope induces transformations; the analysis of biorthogonal functions involves the $W(E_6)$ -invariant Schläfli polytope (2_{21}) in a similar way.

One major open problem is to extend the double affine Hecke algebra approach (à la Cherednik and Sahi) to the elliptic level. One key ingredient of this approach for Macdonald and Koornwinder polynomials is the existence of certain nonsymmetric analogues (indexed by arbitrary, as opposed to dominant, weights of the appropriate group). This is somewhat problematical at the elliptic level, as in the absence of symmetry, spaces of multivariate theta functions tend to be smaller than the corresponding spaces of Laurent polynomials. For instance, if one considers the space of all (holomorphic) theta functions with the same quasiperiodicity as BC_n -symmetric theta functions of degree m , the

resulting space has dimension $(2m)^n$; in contrast, Laurent polynomials with degrees between $-m$ and m in each variable form a space of dimension $(2m + 1)^n$. It thus seems likely that one must break not just the C_n symmetry, but the quasiperiodicity as well (i.e., the action of the *affine* Weyl group), though by producing infinite-dimensional spaces, this would introduce its own set of difficulties. Some partial progress towards nonsymmetric analogues was obtained in [14], which established a nonsymmetric analogue of a limiting case of the elliptic binomial coefficient (the values of which are principal specializations of skew Macdonald polynomials). These nonsymmetric binomial coefficients (which take compositions as upper and lower arguments) can in turn be interpreted as values of certain nonsymmetric rational functions evaluated at compositions (work in progress of Lascoux, Warnaar, and the author).

A related problem is to understand more fully the difference equations satisfied by the biorthogonal functions. By composing our difference operators in suitable pairs, one obtains difference equations of the form

$$D(v)\tilde{R}_\lambda^{(n)}(;t_0:t_1, t_2, t_3; qu_0, u_1/q; t; p, q) = \prod_{1 \leq i \leq n} \frac{\theta_p(v\hat{t}_0q^{\lambda_i}t^{n-i}, vq^{-\lambda_i}t^{i-n}/\hat{t}_0)}{\theta_p(\hat{u}_0\hat{t}_0q^{\lambda_i}t^{n-i}, \hat{u}_0q^{-\lambda_i}t^{i-n}/\hat{t}_0)} \tilde{R}_\lambda^{(n)}(;t_0:t_1, t_2, t_3; u_0, u_1; t; p, q),$$

for $v \in \{\hat{t}_r, \hat{t}_r/q\}$, where $D(v)$ is a difference operator of the form

$$(D(v)f)(\dots, z_i, \dots) = \sum_{\sigma \in \{-1, 0, 1\}^n} c_\sigma(z) f(\dots, q^{\sigma_i} z_i, \dots)$$

with elliptic coefficients c_σ . One can show in general that such an operator exists for all v , but no explicit formula is known (not even for the case $v = \hat{u}_0$, when the “eigenvalues” are trivial). Dually, one has a reasonable formula for a branching rule for biorthogonal functions only in the case that a variable is being specialized to t_r or t_r/t .

A final collection of open problems is a family of conjectured quadratic transformations which were stated (together with proofs of a number of special cases and internal consistency checks) in [20]. The typical such conjecture involves integrating a biorthogonal function against an elliptic Selberg density with different values of p, q , and t . For instance, one conjectures that the integral of

$$\tilde{R}_\lambda^{(2n)}(\dots, p^{1/4}z_i^{\pm 1}, \dots; p^{1/4}t_0:p^{-1/4}t_0, p^{1/4}t_1, p^{-1/4}t_1:p^{1/4}u_0, p^{-1/4}tu_0; q, t; p)$$

with respect to the elliptic Selberg density with parameters

$$(t_0, t_1, u_0, \pm t^{1/2}, -q^{1/2}t^{1/2}; q, t; p^{1/2})$$

vanishes unless each part of λ has even multiplicity (when the integral is an explicit ratio of Δ symbols). In a suitable limit $p \rightarrow 0$, the integral becomes

$$\int P_\lambda(\dots, x_i^{\pm 1}, \dots; q, t) \prod_{1 \leq i < j \leq n} \frac{(x_i^{\pm 1}x_j^{\pm 1}; q)}{(tx_i^{\pm 1}x_j^{\pm 1}; q)} \prod_{1 \leq i \leq n} \frac{(x_i^{\pm 2}; q)}{(tx_i^{\pm 1}; q)},$$

the requisite vanishing of which was established in [25]. When $t = q$, this becomes the fact that the integral of a Schur function over the symplectic group vanishes similarly (per the theory of symmetric spaces). These conjectures include analogues of other such representation-theoretical vanishing identities, and are equivalent in the limit $n \rightarrow \infty$ to Littlewood-type identities, e.g., the fact established by Macdonald [15] that Macdonald polynomials have a generating function of the form

$$\sum_{\lambda} c_{\lambda} P_{\lambda}(x_1, \dots, x_n; q, t) = \prod_{1 \leq i < j \leq n} \frac{(tx_i x_j; q)}{(x_i x_j; q)}$$

where c_{λ} vanishes unless every part of λ has even multiplicity (and has an explicit product formula when nonzero). Another example is a generating function for Macdonald polynomials conjectured by Kawanaka [11] and recently established (using elliptic special functions) by Langer, Schlosser, and Warnaar [13]: an expansion of the form

$$\sum_{\lambda} c_{\lambda} P_{\lambda}(x; q^2, t^2) = \prod_{1 \leq i} \frac{(tx_i; q)}{(x_i; q)} \prod_{1 \leq i < j \leq n} \frac{(t^2 x_i x_j; q^2)}{(x_i x_j; q^2)}$$

with coefficients c_{λ} given by an explicit product over the Young diagram of λ . Enough special cases of the conjectures of [20] have been established to provide a different, also elliptic, proof of Kawanaka's conjecture. Though either proof could in principle have been couched in purely q -hypergeometric terms, it seems unlikely that they would have been discovered except via the elliptic theory.

References

- [1] R. Askey and J. Wilson. *Some basic hypergeometric orthogonal polynomials that generalize Jacobi polynomials*. Number 319 in *Memoirs of the AMS*. Amer. Math. Soc., Providence, RI, 1985.
- [2] A. Borodin, V. Gorin, and E. M. Rains. q -Distributions on boxed plane partitions. arXiv:0905.0679.
- [3] F. J. van de Bult and E. M. Rains. Basic hypergeometric functions as limits of elliptic hypergeometric functions. *Symmetry Integrability Geom. Methods Appl.*, 5:Paper 059, 31, 2009.
- [4] I. Cherednik. *Double affine Hecke algebras*, volume 319 of *London Mathematical Society Lecture Note Series*. Cambridge University Press, Cambridge, 2005.
- [5] H. Coskun and R. A. Gustafson. Well-poised Macdonald functions W_{λ} and Jackson coefficients ω_{λ} on BC_n . In *Jack, Hall-Littlewood and Macdonald polynomials*, volume 417 of *Contemp. Math.*, pages 127–155. Amer. Math. Soc., Providence, RI, 2006.
- [6] E. Date, M. Jimbo, A. Kuniba, T. Miwa, and M. Okado. Exactly solvable SOS models. II. Proof of the star-triangle relation and combinatorial identities. In

- Conformal field theory and solvable lattice models (Kyoto, 1986)*, volume 16 of *Adv. Stud. Pure Math.*, pages 17–122. Academic Press, Boston, MA, 1988.
- [7] J. F. van Diejen. Self-dual Koornwinder-Macdonald polynomials. *Invent. Math.*, 126(2):319–339, 1996.
 - [8] J. F. van Diejen and V. P. Spiridonov. An elliptic Macdonald-Morris conjecture and multiple modular hypergeometric sums. *Math. Res. Lett.*, 7(5-6):729–746, 2000.
 - [9] I. B. Frenkel and V. G. Turaev. Elliptic solutions of the Yang-Baxter equation and modular hypergeometric functions. In *The Arnold-Gelfand mathematical seminars*, pages 171–204. Birkhäuser Boston, Boston, MA, 1997.
 - [10] G. Gasper and M. Rahman. *Basic Hypergeometric Series*, volume 35 of *Encyclopedia of Mathematics and its Applications*. Cambridge University Press, 1990.
 - [11] N. Kawanaka. A q -series identity involving Schur functions and related topics. *Osaka J. Math.*, 36(1):157–176, 1999.
 - [12] T. H. Koornwinder. Askey-Wilson polynomials for root systems of type BC . In Donald St. P. Richards, editor, *Hypergeometric functions on domains of positivity, Jack polynomials, and applications (Tampa, FL, 1991)*, Contemp. Math. 138, pages 189–204. Amer. Math. Soc., Providence, RI, 1992.
 - [13] R. Langer, M. J. Schlosser, and S. O. Warnaar. Theta functions, elliptic hypergeometric series, and Kawanaka’s Macdonald polynomial conjecture. *SIGMA Symmetry Integrability Geom. Methods Appl.*, 5:Paper 055, 20, 2009.
 - [14] A. Lascoux, E. M. Rains, and S. O. Warnaar. Nonsymmetric interpolation Macdonald polynomials and \mathfrak{gl}_n basic hypergeometric series. *Transform. Groups*, 14(3):613–647, 2009.
 - [15] I. G. Macdonald. *Symmetric Functions and Hall Polynomials*. Oxford Univ. Press, Oxford, England, second edition, 1995.
 - [16] I. G. Macdonald. *Affine Hecke algebras and orthogonal polynomials*, volume 157 of *Cambridge Tracts in Mathematics*. Cambridge University Press, Cambridge, 2003.
 - [17] M. Noumi. Macdonald-Koornwinder polynomials and affine Hecke rings. *Sūrikaiseikikenkyūsho Kōkyūroku*, (919):44–55, 1995. Various aspects of hypergeometric functions (Japanese) (Kyoto, 1994).
 - [18] A. Okounkov. BC -type interpolation Macdonald polynomials and binomial formula for Koornwinder polynomials. *Transform. Groups*, 3(2):181–207, 1998.
 - [19] M. Rahman. An integral representation of a ${}_{10}\phi_9$ and continuous bi-orthogonal ${}_{10}\phi_9$ rational functions. *Can. J. Math.*, 38:605–618, 1986.
 - [20] E. M. Rains. Elliptic Littlewood identities. arXiv:0806:0871.
 - [21] E. M. Rains. BC_n -symmetric polynomials. *Transform. Groups*, 10(1):63–132, 2005.
 - [22] E. M. Rains. BC_n -symmetric abelian functions. *Duke Math. J.*, 135(1):99–180, 2006.
 - [23] E. M. Rains. Limits of elliptic hypergeometric integrals. *Ramanujan J.*, 18(3):257–306, 2009.

-
- [24] E. M. Rains. Transformations of elliptic hypergeometric integrals. *Ann. Math.*, 171(1):169–243, 2010.
- [25] E. M. Rains and M. Vazirani. Vanishing integrals of Macdonald and Koornwinder polynomials. *Transform. Groups*, 12(4):725–759, 2007.
- [26] H. Rosengren. A proof of a multivariable elliptic summation formula conjectured by Warnaar. In *q-series with applications to combinatorics, number theory, and physics (Urbana, IL, 2000)*, volume 291 of *Contemp. Math.*, pages 193–202. Amer. Math. Soc., Providence, RI, 2001.
- [27] S. N. M. Ruijsenaars. A generalized hypergeometric function satisfying four analytic difference equations of Askey-Wilson type. *Comm. Math. Phys.*, 206(3):639–690, 1999.
- [28] S. Sahi. Nonsymmetric Koornwinder polynomials and duality. *Ann. of Math. (2)*, 150(1):267–282, 1999.
- [29] V. P. Spiridonov. On the elliptic beta function. *Uspekhi Mat. Nauk*, 56(1(337)):181–182, 2001.
- [30] V. P. Spiridonov. Theta hypergeometric integrals. *Algebra i Analiz (St. Petersburg Math. J.)*, 15:161–215, 2003.
- [31] V. P. Spiridonov and A. S. Zhedanov. Spectral transformation chains and some new biorthogonal rational functions. *Commun. Math. Phys.*, 210:49–83, 2000.
- [32] S. O. Warnaar. Summation and transformation formulas for elliptic hypergeometric series. *Constr. Approx.*, 18(4):479–502, 2002.
- [33] J. A. Wilson. Orthogonal functions from Gram determinants. *SIAM J. Math. Anal.*, 22(4):1147–1155, 1991.

Percolation on Sequences of Graphs

Oliver Riordan*

Abstract

Recently many new random graph models have been introduced, motivated originally by attempts to model disordered large-scale networks in the real world, but now also by the desire to understand mathematically the space of (sequences of) graphs. This article will focus on two topics. Firstly, we discuss the percolation phase transition in these new models, and in general sequences of dense graphs. Secondly, we consider the question ‘when are two graphs close?’ This is important for deciding whether a graph model fits some real-world example, as well as for exploring what models are possible. Here the situation is well understood for dense graphs, but wide open for sparse graphs.

The material discussed here is from a variety of sources, primarily work of Bollobás, Janson and Riordan and of Borgs, Chayes, Lovász, Sós, Szegedy and Vesztergombi. The viewpoint taken here is based on recent papers of Bollobás and the author.

Mathematics Subject Classification (2010). Primary 05C80; Secondary 60C05.

Keywords. Inhomogeneous random graphs, phase transition, metrics on graphs

1. Introduction

In the last decade or so, a great many new random graph models have been introduced. The original motivation, which is still important, was the desire to have models that better fit many real-world networks. These are often disordered, or determined by underlying mechanisms that cannot possibly be precisely modelled, and so are naturally modelled by some type of random graph. Typically, the networks are *sparse*: the number of edges is much smaller than the maximum possible, and in fact usually roughly linear in the number of vertices. Unlike the classical random graphs, real-world networks are often highly ‘inhomogeneous’, for example with power-law distribution of parameters such

*Mathematical Institute, University of Oxford, 24–29 St Giles’, Oxford OX1 3LB, UK.
E-mail: riordan@maths.ox.ac.uk.

as vertex degrees. Hence the mathematical models should also be inhomogeneous. Many of the models that have been proposed are rather *ad hoc*, but nonetheless a unified mathematical theory seems to be emerging that covers a large subclass of these sparse, inhomogeneous models. One of the aims of this survey is to outline this theory. In doing so, there are many mathematical questions one could consider about these random graphs; we shall focus on the percolation phase transition, i.e., the emergence of a giant component as some parameter is varied.

The second aim is to consider the rather open-ended questions of what models are in principle possible, and how to tell whether a certain model fits a certain real-world network; these turn out to be closely related. It might seem that such a question is unlikely to have a clear-cut answer; however, for *dense* graphs, there is a very satisfactory answer. The situation for sparse graphs is rather different, as we shall see; here there are some results, but many more interesting open questions.

2. The Classical Models

Any survey of random graphs must start with the classical models $G(n, p)$ and $G(n, m)$ of Gilbert [43] and Erdős and Rényi [35]. Throughout this paper, by a *graph* we mean a pair (V, E) where V is a (usually finite) set, the set of *vertices*, and E , the set of *edges*, is a set of unordered pairs from V . Thus multiple edges and loops are not allowed: each pair ij either is or is not an edge. Almost always, n will denote the number of vertices; usually the vertex set is taken to be $[n] = \{1, 2, \dots, n\}$ for simplicity.

Although there are some earlier appearances of random graphs (for example in the paper of Solomonoff and Rapoport [57] on a not quite precisely defined model of a neural network), the study of random graphs really started in 1959 with the first of a series of papers by Erdős and Rényi [35, 36, 37, 38, 39, 40]. Given parameters n and m specifying the number of vertices and edges, they considered $G(n, m)$, a random graph chosen uniformly from among all $\binom{N}{m}$ graphs on $[n]$ with m edges, where $N = \binom{n}{2}$. In other words, they studied the probability space $\mathcal{G}(n, m)$ whose elements are all graphs on $[n]$ with m edges, where the probability measure is the normalized counting measure.

At the same time, Gilbert [43] considered a closely related model: given parameters n and p , let $G(n, p)$ denote the graph on $[n]$ obtained by including each possible edge ij with probability p , independently of all other edges. (In practice there is no danger of confusion with the notation: $G(n, x)$ denotes the former model if $x \geq 1$ is an integer, and the latter if $0 < x < 1$.) Note that the number of edges of $G(n, p)$ has a binomial distribution $\text{Bi}(N, p)$, which is tightly concentrated around its mean $Np = \binom{n}{2}p \sim n^2p/2$. Although for some questions the difference is important (see, for example, [53]), for most questions the two models are essentially equivalent, provided we choose corresponding

parameters, i.e., $p = m/\binom{n}{2}$. Although Erdős and Rényi proved their results for $G(n, m)$, we shall phrase them in terms of the more convenient model $G(n, p)$.

The models $G(n, p)$ and $G(n, m)$ are clearly fundamental combinatorial objects, and it is no surprise that thousands of papers have been written about them and related models. Still, even now, one of the most striking properties of $G(n, p)$ is captured by one of the original results of Erdős and Rényi, concerning the *phase transition*: in the natural parametrization $p = c/n$, c constant, there is a sudden change in the global structure as c crosses 1.

Listing the connected components of a graph G in decreasing order of their size (measured by number of vertices), let $C_i(G)$ denote the number of vertices in the i th largest component. We say that an event (formally a sequence of events) holds *with high probability*, or *whp*, if its probability tends to 1 as $n \rightarrow \infty$. Given a sequence (X_n) of real-valued random variables, and a real number a , we write $X_n \xrightarrow{\mathbb{P}} a$ if X_n converges to a in probability, i.e., $\mathbb{P}(|X_n - a| \geq \varepsilon) \rightarrow 0$ for any fixed $\varepsilon > 0$. With these definitions, the phase-transition result of Erdős and Rényi [36] may be stated as follows.

Theorem 2.1. *Let c be a constant, and let $G = G(n, c/n)$.*

If $c < 1$ then there is a constant $C > 0$ such that $C_1(G) \leq C \log n$ whp.

If $c > 1$ then there is a constant ρ such that $C_1(G)/n \xrightarrow{\mathbb{P}} \rho$.

If $c = 1$ then $C_1(G)$ is of order $n^{2/3}$.

We have deliberately left the third statement in a rather vague form; more precisely, the random quantities $C_1(G)/n^{2/3}$ and $n^{2/3}/C_1(G)$ are both bounded in probability.

Theorem 2.1 is an extremely important result about random graphs, and at first sight rather surprising. Leaving aside the third statement for the moment, it shows that a very small change in the parameters can precipitate a very large change in the component structure of the graph $G(n, p)$, from all components ‘small’ to small components together with a *giant component*, i.e., a component containing order n vertices. In other words, the property of having a giant component has a *sharp threshold* at $p^*(n) = 1/n$: if $\liminf p(n)/p^*(n) > 1$ then whp $G(n, p)$ has a giant component, while if $\limsup p(n)/p^*(n) < 1$ then whp it does not.

This result is closely analogous to the phase transition seen in (for example) percolation, although the scaling is different. In percolation, one considers a random subgraph of some infinite graph, for example the square lattice, obtained by selecting each edge (say) independently with probability p . It turns out that there is a critical probability p_c such that for $p < p_c$ all components are finite (with probability 1), and their size distribution has an exponential tail, while for $p > p_c$ there is an infinite component containing a constant fraction of the vertices. It is interesting to note that the theories of random graphs and of percolation started at essentially the same time, the latter with the 1957 paper of Broadbent and Hammersley [27].

At first sight, Theorem 2.1 might seem to be the end of the story, but this is very far from the case. Firstly, one could consider an arbitrary function $p = p(n)$, rather than $p = c/n$. Passing to subsequences, one can assume without loss of generality that $np \rightarrow c$ for some constant c . For $c \neq 1$, the result for $p = c/n$ carries over essentially unchanged to the case $np \rightarrow c$. For $c = 1$, contrary to the claim in [36], this is not the case: indeed, it cannot be the case, since some ‘smooth transition’ must occur between the cases $c < 1$, $c = 1$ and $c > 1$. This error was first noticed by Bollobás [12] in 1984, who proved detailed results about the behaviour around the phase transition. Even more precise results of this type have now been proved, by Łuczak [51] in 1990, and by many people in more recent papers, including the seminal paper of Janson, Knuth, Łuczak and Pittel [46].

Here we shall focus on extensions of Theorem 2.1 in an orthogonal direction, considering analogous questions and results for a wide range of models, rather than looking for more and more precise results for a single (fundamental!) model.

Before turning to this, note that the constant $\rho = \rho(c)$ in Theorem 2.1, given in [36] by an explicit formula in terms of an infinite sum, has a very simple interpretation. Consider the Galton–Watson branching process \mathfrak{X}_c defined as follows: start with a single individual in generation 0. Each individual in generation t has a random number of children with a Poisson distribution with mean c ; together these (disjoint sets of) children make up generation $t + 1$. The numbers of children are independent for different individuals in the same generation, and independent of the history. Heuristically, if one ‘explores’ the component of a random vertex v of $G(n, c/n)$, then, at least at first, the (breadth first) search tree constructed is very close in distribution to \mathfrak{X}_c , so it is no surprise that $\rho(c)$ is equal to the probability that \mathfrak{X}_c survives forever. Of course, this is very far from a proof of Theorem 2.1. Let us note for later that the branching process survival probability $\rho(c)$ satisfies the equation

$$\rho(c) = 1 - e^{-c\rho(c)}. \quad (1)$$

To see this, let S_1 denote the set of individuals in generation 1 of the process, i.e., the children of the initial individual, so $|S_1|$ has a Poisson distribution $\text{Po}(c)$ by definition. Given S_1 , the descendants of each individual $x \in S_1$ form an independent copy of the process \mathfrak{X}_c . Let S_1^+ denote the subset of S_1 containing those x whose descendants survive forever. Then each $x \in S_1$ is included in S_1^+ independently with probability $\rho(c)$, and it follows that $|S_1^+|$ has the distribution $\text{Po}(c\rho(c))$. Since the process survives if and only if $|S_1^+| > 0$, this establishes (1). Note that this equation on its own does not determine $\rho(c)$ (since 0 is always a solution), but standard arguments show that $\rho(c)$ is the maximum solution to this equation, which is positive if and only if $c > 1$.

Although our focus in the rest of this paper will be on results with $o(n)$ error terms such as Theorem 2.1, let us mention that comparison with branching

processes can also be used to prove precise results about the window of the phase transition; see Bollobás and Riordan [20].

Of course, $G(n, p)$ and $G(n, m)$ are not the only classical models; perhaps the next most studied is the random r -regular graph. The definition is very simple: select a graph uniformly at random from among all r -regular graphs on $[n]$. However, this definition is not so easy to work with, in part because there is no simple formula for the number of such graphs. In 1978, Bender and Canfield [9] gave an asymptotic formula for this number. In 1980, Bollobás [11] reproved this formula in a probabilistic way, and gave a practical method for constructing a random r -regular graph, the *configuration model*. The key point is that this model makes random r -regular graphs accessible to study by probabilistic methods. It also works just as well for constructing and analyzing a random graph with any given degree sequence.

In the next section we discuss a variety of new random graph models, concentrating on the sparse case, where the number of edges is linear in the number of vertices. This is important for modelling real-world networks, and is also mathematically interesting, since it is in this range that the percolation phase transition occurs.

3. Models of Real-world Networks

Although it is impossible to pin down a precise starting point, one of the catalysts for the recent surge of interest in new random graph models was the 1998 paper of Watts and Strogatz [61] concerning the ‘small world phenomenon’ of small (logarithmic) diameter in many real-world networks. Mathematically, this is *not* surprising, and essentially the Watts–Strogatz model had been analyzed by Bollobás and Chung [14] a decade earlier. In the next few years, attention shifted away from the small-world phenomenon, and towards the inhomogeneity of many real-world networks.

Although an individual realization of the symmetric model $G(n, p)$ will not be symmetric, it still turns out to be fairly ‘homogeneous’. For example, in the sparse case, the distribution of vertex degrees is asymptotically Poisson, and so is tightly concentrated: degrees much larger than the mean are exponentially unlikely. In contrast, many networks in the real world are highly inhomogeneous: their degree distributions are typically far from concentrated, often having a power-law tail. (Inhomogeneity is usually present in many other features of the network; the degree sequence tends to be singled out because it is the easiest to measure.) Observations of this kind were made by Barabási and Albert [7], Faloutsos, Faloutsos and Faloutsos [41] and Kleinberg, Kumar, Raghavan, Rajagopalan and Tomkins [47] around 1999. Such observations triggered the introduction of a vast number of new, inhomogeneous models, of which we shall briefly mention only a few; for early surveys of this area see Albert and Barabási [2] or Dorogovtsev and Mendes [32]; there are now several books about such models, for example the recent book by Dorogovtsev [31].

Perhaps the earliest and certainly the best known of the new inhomogeneous models is the ‘growth with preferential attachment’ model of Barabási and Albert [7], generally known as the BA model. Fix an integer $m \geq 2$; the BA graph with parameter m grows by adding vertices one at a time. When each new vertex is added, so are m edges from this new vertex to m distinct old vertices, chosen randomly with probabilities proportional to their degrees. This model is not in fact mathematically precise; a precise interpretation is the Linearized Chord Diagram model of Bollobás and Riordan [19], which also has the great advantage of giving a static description of the n -vertex graph. Indeed, one simply takes a uniformly random pairing of an ordered set of $2mn$ objects. Then, thinking of each pair as having a ‘left’ and ‘right’ endpoint (determined by the order), starting from the left, one identifies all objects up to the first right endpoint to form vertex 1, then all objects beyond this up to the next right endpoint to form vertex 2, and so on (for the details, see [19]). This gives a graph with a small number of loops and multiple edges, whose evolution with n satisfies a precise form of the Barabási–Albert rule. As shown heuristically by Barabási and Albert [7] and rigorously by Bollobás, Riordan, Spencer and Tusnády [23], these models do indeed generate graphs with power-law degree distributions.

An alternative approach, simply building in the power law rather than trying to explain it, was taken by Aiello, Chung and Lu [1]: generate a power-law distribution, and then generate a random graph with this degree sequence, using the configuration model of Bollobás [11]. One then studies other features of the resulting graphs, to see what properties one can expect in ‘typical’ graphs with the given degree distribution.

The nature of models such as the BA model makes them difficult to analyze rigorously: since the random graph is defined recursively by a growth rule, the description of the distribution of the n -vertex graph is somewhat involved. Nevertheless, using the static LCD description, properties such as the diameter [19] and the critical point of the phase transition [18] were found by Bollobás and Riordan. (In the latter case the answer is that there is no critical point: surprisingly, for any constant edge density there is whp a giant component.) Often, however, instead of analyzing the model directly, one analyzes a simplified variant with (essentially) the same edge probabilities, but independence between edges; this is sometimes called a ‘mean-field’ approximation. For example, very precise results about the size of the giant component in this case are given in [54].

4. Inhomogeneous Graphs and Branching Process

The defining properties of $G(n, p)$ are that edges are present independently of each other, and that each edge has the same probability of being present.

Can we still say something if these assumptions are weakened or dropped? One reason to consider this question is that, from a mathematical point of view, the vast array of new models hinted at in the previous section is somewhat unsatisfactory. Even when the individual models are tractable, with so many models one needs many separate arguments for different cases; for a time there seemed to be no general theory. In response to this situation, in 2007, Bollobás, Janson and Riordan [15] introduced a very general model of inhomogeneous sparse random graphs that includes many of the previous models as special cases, as well as having special cases that approximate many others. We shall not give the full details of the definition, which are rather involved, but only an outline.

The key idea is that each vertex has a ‘type’, and the probability of an edge ij depends on the types of the vertices i and j . One could consider two types of vertex, or more generally any finite number. Indeed, this extension of $G(n, p)$ is essentially ‘folklore’ within the random graphs community; its phase transition was studied explicitly by Söderberg [56]. However, the BJR model is much more general: the set \mathcal{S} of types is typically infinite. Without losing much generality, it turns out that one can take $\mathcal{S} = [0, 1]$. Vertices are assigned types from this set according to some rule: the simplest is that the types x_1, \dots, x_n are independent and uniformly distributed on $[0, 1]$, but other rules are possible. Given the types, each edge ij is present independently with probability $\kappa(x_i, x_j)/n$, where $\kappa : [0, 1]^2 \rightarrow [0, \infty)$ is a *kernel*, i.e., a symmetric, measurable function. (More precisely, one takes the smaller of $\kappa(x_i, x_j)/n$ and 1 as the edge probability.) Depending on how the vertex types are chosen, some additional assumptions on κ may be needed, but in the i.i.d. case, they are not (see [16]).

The basic motivation of the BJR model is to allow inhomogeneity by allowing different vertex types, and then taking the edge probabilities as an arbitrary function of the types. The normalization, dividing by n , is natural for the main questions considered in [15], such as finding the critical point of the phase transition. It is not hard to check that the degree distribution is asymptotically mixed Poisson $\text{Po}(\Lambda)$, where the distribution of Λ is simply the marginal distribution of κ ; this can easily be chosen to have a power-law tail with any desired exponent by a suitable choice of κ .

The construction of the BJR model in terms of a continuum object (the kernel) makes it amenable to analysis: one can hope to relate properties of the graph to those of the kernel. Again, a branching process plays a crucial role; this time it is a multi-type process where each individual has a type, and the distribution of the number and types of the children depends on the type of the parent. Specifically, the types of the children of an individual of type x form a Poisson process with intensity $\kappa(x, y) dy$ on $[0, 1]$. As shown in [15], ignoring the technical complications, in this very general setting the existence and size of any giant component are related to the survival probability of the branching process. This may be found by solving a certain functional equation

analogous to (1): the survival probability is a function ρ of the type of the initial individual, and this function ρ satisfies

$$\rho = 1 - e^{-T_\kappa \rho}, \quad (2)$$

where T_κ is the integral operator with kernel κ , defined by

$$(T_\kappa(f))(x) = \int_0^1 \kappa(x, y) f(y) \, dy.$$

As shown in [15], this gives a (relatively) simple criterion for the critical point of the phase transition: it occurs when $\|T_\kappa\| = 1$, where $\|\cdot\|$ denotes the L^2 norm. Furthermore, the normalized size of the giant component is asymptotically equal to the survival probability $\rho(\kappa) = \int_0^1 \rho(x) \, dx$ of the branching process, under the natural additional assumption of *irreducibility*. Roughly speaking, irreducibility says that the set \mathcal{S} of types cannot be split into two non-trivial parts so that individuals with types in one part only have children with types in the same part. Given a sequence (X_n) of random variables and a deterministic function $f(n)$, we say that $X_n = o_p(f(n))$ if $X_n/f(n) \xrightarrow{P} 0$. We write ‘ $X_n = \Theta(n)$ whp’ if there are constants $0 < c < C$ such that $c \leq X_n/f(n) \leq C$ holds whp as $n \rightarrow \infty$.

Theorem 4.1. *Let $\kappa : [0, 1]^2 \rightarrow [0, \infty)$ be a kernel, and let $G_n = G(n, \kappa)$ denote the random graph constructed by the BJR model.*

- (i) *If $\|T_\kappa\| \leq 1$, then $C_1(G_n) = o_p(n)$.*
- (ii) *If $\|T_\kappa\| > 1$, then $C_1(G_n) = \Theta(n)$ whp. Furthermore, if κ is irreducible, then $C_1(G_n)/n \xrightarrow{P} \rho(\kappa)$ and $C_2(G_n) = o_p(n)$.*

Note that for $p < 1$ the random graph $G(n, p\kappa)$ can be seen as the random subgraph of $G(n, \kappa)$ obtained by selecting each edge independently with probability p , so the result above can be seen as establishing the percolation threshold $p_c = \|T_\kappa\|^{-1}$ for percolation on the *random* graph $G(n, \kappa)$. In a similar way, most of the percolation threshold results discussed here can be seen either as results about a model with a variable parameter, or as results about the threshold for percolation on a random subgraph of a given (random) graph, when the probability p that each edge is included in the subgraph is varied.

Theorem 4.1 includes previous results about many specific models as special cases; not only (variants of) many of the ‘scale-free’ models mentioned in the previous section, but also other inhomogeneous models such as Dubins’ model – see the discussion in Section 16 of [15]. The BJR model also has applications in other areas, for example to a random graph related to quantum theory; see Janson [45].

There is a natural *dense* analogue of the model $G(n, \kappa)$, introduced independently by Lovász and Szegedy [49], which they called a *W-random graph* (W being their symbol for a kernel). By a *standard kernel* we mean a kernel taking values in $[0, 1]$; these are often called *graphons*. Given a standard kernel κ , the random graph $G_1(n, \kappa)$ is defined by choosing x_1, \dots, x_n independently and uniformly from $[0, 1]$ and then, given this sequence, including each possible edge ij with probability $\kappa(x_i, x_j)$, independently of the other edges. This differs from the BJR model in two ways: the main one is simply the normalization; the second is that vertex types are independent, rather than distributed according to some more general rule. Of course, one can interpolate between the models: given a ‘normalizing density’ $p = p(n)$, one can take the edge probabilities to be $p\kappa(x_i, x_j)$. The model of Lovász and Szegedy illustrates a recurring theme: in the study of dense graphs, any kernels that arise tend to be standard kernels (or, slightly more generally, bounded kernels). In the study of sparse graphs, unbounded kernels arise naturally.

In section 7 we shall return to the discussion of sparse inhomogeneous random graphs. But first we shall take what seems to be a detour, considering *non-random* dense graphs.

5. Metrics on Dense Graphs

How does one measure the similarity between two graphs? This is an important question whenever any kind of graph is used as a model for a real-world network: the model graph will presumably not be exactly the same as the real-world example, so if it is claimed to be ‘close’, then in what sense? Here the sparse case is most interesting, but we shall first discuss the dense case, where we think of our graphs as having n vertices and $\Theta(n^2)$ edges. Initially, we consider how to define the distance between two graphs with the same number of vertices. Usually, though not always, the labelling is irrelevant, i.e., one considers isomorphic graphs to be the same.

The smallest change one can make to an n -vertex graph is to add or delete an edge. This naturally suggests a metric d_{edit} :

$$d_{\text{edit}}(G, H) = \min\{|E(G) \Delta E(H')| : H' \cong H\}, \quad (3)$$

i.e., the smallest number of edge additions or deletions needed to turn G into a graph isomorphic to H . This metric seems to have been first explicitly defined by Axenovich, Kézdy and Martin [6], although it had been used implicitly much earlier, e.g., by Erdős [34] and Simonovits [55] in 1966, and in many subsequent papers.

There is another way of viewing the edit distance that makes it possible to compare graphs with different numbers of vertices. Given a graph G on the vertex set $[n]$, there is a natural way to represent G by a function $\kappa_G :$

$[0, 1]^2 \rightarrow \{0, 1\}$: split $[0, 1]$ into n equal intervals I_i , and let κ_G be constant on each set $I_i \times I_j$, taking the value 1 if ij is an edge of G and 0 otherwise. In other words, represent the adjacency matrix of G as a function in the obvious way, obtaining a (special type of) standard kernel. Up to a normalizing factor of $n^2/2$, the distance $d_{\text{edit}}(G, H)$ coincides with $\min\{\|\kappa_G - \kappa_{H'}\|_1 : H' \cong H\}$, where $\|\cdot\|_1$ denotes the L^1 norm on functions on $[0, 1]^2$. This definition makes sense even when G and H have different numbers of vertices, although it is not very natural in this case.

Once we pass to kernels instead of graphs, it is not clear what the appropriate sense of relabelling should be. Given a kernel κ and a measure-preserving bijection $\tau : [0, 1] \rightarrow [0, 1]$, let $\kappa^{(\tau)}$ be the kernel defined by

$$\kappa^{(\tau)}(x, y) = \kappa(\tau(x), \tau(y));$$

we call $\kappa^{(\tau)}$ a *rearrangement* of κ , and write $\kappa \approx \kappa'$ if κ' is a rearrangement of κ . Then one can define an L^1 metric with rearrangement on $L^1([0, 1]^2)$ by $d_1(\kappa_1, \kappa_2) = \inf\{\|\kappa_1 - \kappa'_2\|_1 : \kappa'_2 \approx \kappa_2\}$, and this gives a metric on graphs by mapping G to κ_G as before. In fact, d_1 is a pseudometric, as some pairs of graphs map to the same kernel; one way to obtain a genuine metric on the space of isomorphism classes of graphs, together with kernels modulo the appropriate notion of equivalence, is simply to code in the number of vertices. For example, writing $|G|$ for the number of vertices of G , map a graph G to the pair $(\kappa_G, 1/|G|)$, and a kernel κ to $(\kappa, 0)$, and define a modified metric d'_1 by $d'_1((\kappa_1, x_1), (\kappa_2, x_2)) = d_1(\kappa_1, \kappa_2) + |x_1 - x_2|$, say. (Such a modification was used in a related context by Diaconis and Janson [30].)

The edit distance and its variants are natural in many contexts, but they have many drawbacks. For example, there is a strong intuitive sense in which two instances of the dense random graph $G(n, 1/2)$ are with high probability very similar. This observation goes back to Erdős and Rényi; it is hard to make precise but, for example, for ‘most’ properties \mathcal{P} it turns out that the probability that $G(n, 1/2)$ has property \mathcal{P} tends either to 0 or to 1. In other words, for large n , typical instances of $G(n, 1/2)$ either have the property or do not. Also, numerical quantities such as the number of edges or triangles, or the size of the maximum cut, are tightly concentrated, so two typical instances of $G(n, 1/2)$ are close in many senses. However, it is easy to see that they will be very far apart in the edit distance – essentially as far apart as possible given their numbers of edges.

One concrete way of quantifying the similarity between typical instances of $G(n, p)$ is to consider small subgraph counts. Given a ‘fixed’ graph F , let $X_F(G)$ denote the number of subgraphs of a graph G that are isomorphic to F . (Here we consider all subgraphs, not just induced ones, though it makes very little difference.) It turns out to be more convenient to normalize differently, considering the number of *embeddings* or injective homomorphisms of F into G : $\text{emb}(F, G) = \text{aut}(F)X_F(G)$. If F has k vertices and G has n vertices, then it is natural to normalize by dividing by the number $\text{emb}(F, K_n) = n(n-1) \cdots (n-k)$

$k + 1) \sim n^k$ of embeddings of F into the complete graph on n vertices. This gives the ‘normalized subgraph count’

$$s(F, G) = \frac{\text{emb}(F, G)}{n(n - 1) \cdots (n - k + 1)} = \frac{X_F(G)}{X_F(K_n)} \in [0, 1]. \tag{4}$$

Then one can say that two graphs G and H are ‘similar’ if they have similar normalized counts of edges, triangles, etc. (One can also use homomorphism counts $t(F, G)$ as in [25]; this makes no difference in the dense case and is actually somewhat cleaner. However, in the sparse setting there are advantages to embedding counts; see the discussion in [21].)

Explicitly, one can define a metric on isomorphism classes of graphs by, for example,

$$d_{\text{sub}}(G, H) = \sum_F 2^{-|F|^2} |s(F, G) - s(F, H)|,$$

where the sum runs over one representative F of each isomorphism class of finite graphs. Of course, the normalizing factor $2^{-|F|^2}$ is somewhat arbitrary (as long as it is chosen so that the sum always converges) – we are really interested not in the metric itself, but rather in the induced topology on the completion of the set of graphs, i.e., in which sequences of graphs are Cauchy with respect to the metric. In this case (G_n) is Cauchy if and only if $s(F, G_n)$ converges for each F .

It is immediate that the above metric leads to a compact completion: every sequence (G_n) has a subsequence with the property that, for every F , $s(F, G_n)$ converges to some number $s_F \in [0, 1]$. At first sight it is not clear whether the completion has a nice description. However, Lovász and Szegedy [49] showed that it does: every limit point (s_F) corresponding to a sequence (G_n) with $|G_n| \rightarrow \infty$ may be obtained from a standard kernel $\kappa : [0, 1]^2 \rightarrow [0, 1]$ in a natural way. More explicitly, the limit point corresponds to ‘subgraph counts in the kernel’, defined by

$$s(F, \kappa) = \int_{[0, 1]^k} \prod_{ij \in E(F)} \kappa(x_i, x_j) \prod_{i=1}^k dx_i$$

for each graph F on $\{1, 2, \dots, k\}$.

The metric just defined captures ‘local’ information well, but appears to ignore ‘global structure’, so one would like to consider alternative metrics that take this into account. The edit distance is one such, but as noted above appears to be too fine in many contexts. (Certainly it gives a non-compact completion, for example.) A natural alternative is to say that two graphs on the same vertex set are close if for each (large) subset of the vertices, the graphs contain roughly the same number of edges within this set. Equivalently, if for each pair (U, V) of sets of vertices, the number $e_G(U, V)$ of edges of G from U to V is ‘close’ to $e_H(U, V)$, meaning the difference is small compared to n^2 .

This ‘global’ measure of similarity between graphs is extremely important, for example whenever Szemerédi’s Lemma [58] is used.

Borgs, Chayes, Lovász, Sós and Vesztergombi [25] made the above notion precise, in the following way. Given an integrable function $W : [0, 1]^2 \rightarrow \mathbb{R}$, its *cut norm* is

$$\|W\|_{\square} = \sup_{A, B \subseteq [0, 1]} \left| \int_{A \times B} W(x, y) \, dx \, dy \right|,$$

where the supremum is taken over all measurable sets. This norm agrees up to a constant factor with the operator norm of the corresponding operator $T_W : L^{\infty} \rightarrow L^1$, or, equivalently, the injective tensor product norm on $L^1 \hat{\otimes} L^1$. In combinatorics, it was introduced by Frieze and Kannan [42], who used it to define a ‘weak’ type of Szemerédi partition and showed that such partitions always exist with many fewer classes than may be needed for the usual ‘strong’ partition.

Allowing rearrangement as above for the L^1 metric, one arrives at the *cut metric* of Borgs, Chayes, Lovász, Sós and Vesztergombi [25], defined by

$$\delta_{\square}(\kappa, \kappa') = \inf_{\kappa'' \approx \kappa'} \|\kappa - \kappa''\|_{\square}. \quad (5)$$

Note that δ_{\square} is at first only a pseudometric: one can have $\delta_{\square}(\kappa, \kappa') = 0$ for different kernels κ . To obtain a genuine metric, one can quotient by this notion of equivalence. By first mapping a graph to the corresponding kernel, as before, this metric may also be applied to graphs. [There is also a variant where rearranging is allowed only at the graph level, giving a metric $\hat{\delta}_{\square}$ on graphs with n vertices defined by $\hat{\delta}_{\square}(G, H) = \min\{\|\kappa_G - \kappa_{H'}\|_{\square} : H' \cong H\}$; see [25].]

Recall that a *standard kernel* is simply a symmetric measurable function $\kappa : [0, 1]^2 \rightarrow [0, 1]$; let \mathcal{K}_1 denote the set of all standard kernels. Writing $\kappa \sim \kappa'$ if $\delta_{\square}(\kappa, \kappa') = 0$, the metric space $(\mathcal{K}_1/\sim, \delta_{\square})$ has many nice properties. For example, as shown by Lovász and Szegedy [50], using Szemerédi’s Lemma and the Martingale Convergence Theorem one can show that $(\mathcal{K}_1/\sim, \delta_{\square})$ is compact.

More surprisingly, it turns out that $\kappa \sim \kappa'$ if and only if $s(F, \kappa) = s(F, \kappa')$ for all finite graphs F ; in one direction this is easy, the other direction is not, and was proved directly by Borgs, Chayes and Lovász [24]. (As shown by Diaconis and Janson [30], it also follows from results of Hoover and Kallenberg on exchangeable arrays.) Thus δ_{\square} and d_{sub} may be seen as two metrics on the same space \mathcal{K}_1/\sim . It is not hard to check that d_{sub} is continuous with respect to δ_{\square} , i.e., the identity mapping $i : (\mathcal{K}_1/\sim, \delta_{\square}) \rightarrow (\mathcal{K}_1/\sim, d_{\text{sub}})$ is continuous. Since the spaces are compact, it follows that the inverse is continuous, i.e., δ_{\square} and d_{sub} give rise to the same topology! This is one of the main results of Borgs, Chayes, Lovász, Sós and Vesztergombi [25]. This paper predates [24]: in [25] it was shown among other things that a sequence of graphs or kernels is Cauchy with respect to δ_{\square} if and only if it is Cauchy with respect to d_{sub} . As

pointed out by Bollobás and Riordan [21], this (together with relatively easy results about the two notions of convergence) implies (and is implied by) the uniqueness later proved directly in [24].

Borgs, Chayes, Lovász, Sós and Vesztergombi [25, 26] also showed that the cut metric is equivalent to various other natural metrics on graphs, so it seems to be *the* natural metric on dense graphs. Of course, it makes sense for arbitrary sequences of graphs, but any sequence (G_n) in which the number of edges satisfies $e(G_n) = o(|G_n|^2)$ converges in δ_\square to the zero kernel, which is not very informative.

6. Percolation on Dense Graph Sequences

Having seen kernels arise naturally in two very different contexts, it is natural to wonder if there is a connection. It turns out that the answer is yes.

One can think of $G(n, p)$ as the random subgraph of the complete graph K_n obtained by selecting each edge independently with probability p . In this setting Theorem 2.1 establishes a threshold function $p^*(n) = 1/n$ for percolation on the sequence (K_n) . What can one say if one starts from some other graph? The case of an infinite grid, for example, is the classical percolation model, but what about other finite graphs? Given any graph H , let $H(p)$ denote the subgraph of H obtained by keeping each edge independently with probability p . When does $H(p)$ have a giant component? In this form, the question does not quite make sense – for asymptotic results one needs to consider a sequence (H_n) of graphs with the number of vertices tending to infinity. For convenience, we take H_n to have n vertices.

Given an arbitrary sequence (H_n) , with no relationship between successive terms, it seems impossible to say much about the phase transition in $H_n(p_n)$. Surprisingly, however, as shown by Bollobás, Borgs, Chayes and Riordan [13], assuming only that the sequence is *dense*, i.e., that $e(H_n) = \Theta(n^2)$, one can find the critical point of the phase transition!

Theorem 6.1. *Let (H_n) be a sequence of graphs in which H_n has n vertices and $\Theta(n^2)$ edges. Let λ_n denote the largest eigenvalue of the adjacency matrix of H_n , and let $p_n = \min\{c/\lambda_n, 1\}$, where $c > 0$ is a constant.*

If $c \leq 1$, then the largest component of $H_n(p_n)$ is of size $o_p(n)$, while if $c > 1$ then the largest component of $H_n(p_n)$ has size $\Theta(n)$ whp.

Note that the size of the ‘giant’ component is not specified here; there is no equivalent of $\rho(c)$. To obtain such a result, one does need some restriction on (H_n) ; the natural restriction is convergence in the cut metric. In fact, a simplified form of the main result of [13] is the following (the full result is more general, in that it applies to ‘weighted graphs’).

Theorem 6.2. *Let (H_n) be a sequence of graphs with $|H_n| = n$ converging in δ_\square to a standard kernel κ . Let $c > 0$ be a constant, and let $C_1 = C_1(n)$ denote*

the number of vertices in the largest component of the random graph $H_n(c/n)$, and $C_2 = C_2(n)$ the number of vertices in the second largest component.

- (a) If $c \leq \|T_\kappa\|^{-1}$, then $C_1 = o_p(n)$.
- (b) If $c > \|T_\kappa\|^{-1}$, then $C_1 = \Theta(n)$ whp. More precisely, for any constant $\alpha < (c\|T_\kappa\| - 1)/c$ we have $C_1 \geq \alpha n$ whp.
- (c) If κ is irreducible, then $C_1/n \xrightarrow{P} \rho(c\kappa)$ and $C_2 = o_p(n)$.

In fact, Theorem 6.1 is an immediate consequence of Theorem 6.2 by the usual sub-subsequence argument: it suffices to prove that any subsequence (H_{n_k}) of the original sequence (H_n) has a sub-subsequence along which the conclusion of Theorem 6.1 holds. But by compactness, (H_{n_k}) has a subsequence which converges in the cut metric, to which Theorem 6.2 applies. This illustrates the general phenomenon that one may often assume convergence in the cut metric without loss of generality.

7. More General Sparse Models

Theorems 4.1 and 6.2 have a common form: after conditioning on the vertex types in the case of Theorem 4.1, in both cases the object studied is a sequence (G_n) of random graphs, where in each G_n edges are present independently. In general, the distribution of such a graph G_n is described by a symmetric n -by- n matrix A_n of edge probabilities; without some restrictions on the sequence (A_n) there is not much one can say in this generality. (But see, for example, the results of Alon [5] concerning connectedness.) It turns out to be more convenient to take the entries of A_n to be n times the edge probabilities, so the matrix with all (non-diagonal) entries equal to c corresponds to the classical $G(n, c/n)$.

Let κ_{A_n} denote the piecewise constant kernel corresponding to A_n . Then (ignoring many technical complications), the results of [15] and [13] state that if the kernels κ_{A_n} converge in some sense to a kernel κ , then the asymptotic size of the giant component is given by $\rho(\kappa)n$. In [15] κ may be (indeed usually is) unbounded and, roughly speaking, convergence is in the L^1 norm. In [13], κ is bounded, but only convergence in the cut metric is required, which is much weaker. As shown by Bollobás, Janson and Riordan [17], one can have the best of both worlds: convergence in the cut metric to an arbitrary kernel is sufficient! This is quite surprising since the cut metric is in general not at all well behaved on unbounded kernels. An interesting consequence of the main result of [17] is that if (κ_n) is a sequence of kernels converging in the cut metric to a kernel κ , then $\rho(\kappa_n) \rightarrow \rho(\kappa)$. This statement purely concerns branching process, i.e., is an analytical fact containing no combinatorics. Nevertheless, at the moment, the only known proof goes via graphs! This situation presumably will not last

long. (Assuming convergence in L^1 , which is stronger, a direct proof is not too hard; see [17].)

Although this is not a precise statement, it is quite possible that the result of [17] is the strongest possible within a certain general class: it seems reasonable to believe that convergence of (κ_{A_n}) to κ in the cut metric is the weakest condition that ensures that κ does describe the global behaviour of the random graph associated to A_n .

Although in many ways extremely general, the models discussed so far have one undesirable feature in common. In all cases, the number of short cycles is rather small (asymptotically negligible compared to the number of edges). In particular, the *clustering coefficient*, loosely defined as the probability that two common neighbours of a vertex will themselves be adjacent (i.e., the probability that two of your friends are friends) tends to 0 as the number of vertices tends to infinity. As noted by Watts and Strogatz [61], despite being sparse, many real-world networks exhibit fairly high levels of clustering.

There are many models of sparse random graphs with clustering, for example the model of Dorogovtsev and Mendes [32, Section IX C] in which each new vertex is joined to both ends of a randomly selected edge of the current graph, but again there is not that much one can say about them in general. This is perhaps not surprising, as for sparse graphs independence between edges naturally leads to low clustering, and without independence analysis of the model is very difficult. Fortunately, there is a natural extension of the BJR model with clustering that is still tractable, introduced by the same authors in [16]. Roughly speaking, the model is built out of ‘small subgraphs’ rather than directly from edges. In the simpler special case where the building blocks are all complete subgraphs, one takes vertex types from $[0, 1]$, say, and uses a sequence of kernels $\kappa_r : [0, 1]^r \rightarrow [0, \infty)$, $r = 2, 3, \dots$, to determine the probability of inserting a copy of K_r on a given set of r vertices, as a function of their types. The sequence (κ_r) may be viewed as a single *hyperkernel* on $\bigcup_r [0, 1]^r$.

Once again, the limiting behaviour of the graphs is determined by (κ_r) , and in particular by an associated branching process, now a multi-type compound Poisson process. This time the analysis is complicated by the fact that the equivalent of (1) and (2), namely $\rho = 1 - e^{-S_\kappa \rho}$, involves a non-linear integral operator S_κ . Nevertheless, the criterion for the phase transition still turns out to be relatively simple, namely that the linearized form T of this operator has norm 1.

In some sense there is a danger that the models discussed so far are *too* general: there is so much freedom in choosing the sequence of kernels that quite a significant amount of data from a real-world network may be fitted without much predictive power for further observations; we discuss this in more detail in the next section. Of course, from a mathematical point of view, the more general the better: the main result of [16] includes results such as Theorems 4.1 and 6.2 as very special cases. Finally, it turns out that one can also

work with the cut metric in the hyperkernel setting; see Bollobás, Janson and Riordan [17].

8. Sparse Quasi-random Graphs

For dense graphs, convergence in δ_{\square} to a constant kernel κ is exactly equivalent to the standard notion of quasi-randomness, introduced by Thomason [59] in 1987 (although he called it pseudo-randomness) and studied in great detail by Chung, Graham and Wilson [29] and many others. Extending this, one may think of convergence to a general kernel κ as a kind of inhomogeneous quasi-randomness, so the results of Borgs, Chayes, Lovász, Sós and Vesztergombi [25] may be viewed as stating that inhomogeneous quasi-random sequences of graphs are completely general: any sequence of graphs has such a subsequence. Moreover, the equivalence of δ_{\square} and d_{sub} in the dense case generalizes the well-known equivalence of the corresponding definitions of quasi-randomness to the inhomogeneous setting. Taking an alternative viewpoint, when $\delta_{\square}(G_n, \kappa) \rightarrow 0$ and κ is of ‘finite type’, then κ describes a (weak) Szemerédi partition of the sequence (G_n) , and the fact that $d_{\text{sub}}(G_n, \kappa) \rightarrow 0$ says that the number of copies of any fixed graph F in G_n is asymptotically given by the partition – this is a standard ‘embedding lemma’.

If we are interested in sparse graphs, then it is natural to ask whether these observations about quasi-randomness extend to this setting. There are some results in this direction, for example the results of Thomason [59, 60] on (p, α) -jumbled graphs; however, these make stronger assumptions on the error terms than we would like.

To make sense of the ideas above in the sparse setting, one first needs to adapt the definitions. For δ_{\square} there is no problem: fixing a normalizing density function $p = p(n)$, map a graph G to a piecewise constant kernel as before, but this time taking the values 0 and $1/p$; this is natural since (as is easily checked), provided $np \rightarrow \infty$, the kernels corresponding to $G(n, p)$ almost surely converge in δ_{\square} to the constant kernel 1. For d_{sub} the situation is less clear. One can simply modify the definition (4) by dividing by the expected number of copies of F in $G(n, p)$, rather than the number in K_n , obtaining a p -normalized subgraph count $s_p(F, G)$; however, as noted by Bollobás and Riordan [21], depending on the function p it may only make sense to consider certain subgraph counts, roughly corresponding to graphs F where one expects to see ‘many’ copies of F in $G(n, p)$. The simplest case is when p decreases slowly with n , more precisely when $p = n^{-o(1)}$; then all counts make sense, and one can proceed to define d_{sub} as before.

It turns out that in the sparse case, the behaviour of δ_{\square} and d_{sub} is much harder to understand, and leads to many interesting open questions. One indication that the situation is genuinely more complicated concerns kernels distinguished by their subgraph counts. Recall that for standard (or indeed

bounded) kernels, $\delta_{\square}(\kappa_1, \kappa_2) = 0$ if and only if $s(F, \kappa_1) = s(F, \kappa_2)$ for all finite graphs F . In the sparse case, there is no reason to restrict to bounded kernels, and one can construct simple examples κ_1, κ_2 with $\delta_{\square}(\kappa_1, \kappa_2) > 0$ but $s(F, \kappa_1) = s(F, \kappa_2) < \infty$ for every F . This corresponds to the fact that unbounded random variables are not in general determined by their moments; see [21].

Considering only δ_{\square} , the situation is relatively clear: under a mild additional assumption (that no part of the graph is ‘too dense’ compared to the normalizing density p) many results from the dense case carry over; a key step is that Szemerédi’s Lemma [58] carries over to this setting, as noted independently by Kohayakawa and Rödl; see [48].

Turning to d_{sub} , what are the possible limit points, i.e., limiting subgraph counts? One problem is that $s_p(F, G_n)$ need not remain bounded as $n \rightarrow \infty$. Furthermore, considering only a subset of the possible counts, limiting values are certainly possible that are not consistent with any kernel: for example, as noted in [21], one can have $s_p(K_2, G_n) \rightarrow 1$ and $s_p(C_4, G_n) \rightarrow 1$ (which would force the constant kernel $\kappa = 1$), but $s_p(C_3, G_n) \not\rightarrow 1$. A very interesting question seems to be whether imposing even very strong assumptions saying that the various counts are well behaved is enough to force the limit to be a kernel. In [21], Bollobás and Riordan conjecture that the answer is yes; one form of this conjecture is as follows.

Conjecture 8.1. *Let $p = p(n) = n^{-o(1)}$, and let $C > 0$ be constant. Suppose that (G_n) is a sequence of graphs with $|G_n| = n$ such that, for every F , $s_p(F, G_n)$ converges to some constant $0 \leq c_F \leq C^{e(F)}$. Then there is a bounded kernel κ such that $c_F = s(F, \kappa)$ for every F .*

Considering the very special case of convergence to a uniform kernel, it is shown in [21] that under assumptions similar to those of Conjecture 8.1, if $s_p(K_2, G_n) \rightarrow 1$ and $s_p(C_4, G_n) \rightarrow 1$, then $s_p(F, G_n) \rightarrow 1$ for every graph F with girth at least 4; but for F a triangle this is still open. Turning to the relationship between the metrics δ_{\square} and d_{sub} , similar difficulties arise there. It may well be that when $np \rightarrow \infty$, then under assumptions as above, the two metrics are equivalent. However, this is wide open at the moment. Assuming $\delta_{\square}(G_n, \kappa) \rightarrow 0$, partial results are given in [21], showing that $s_p(F, G_n) \rightarrow s(F, \kappa)$ does hold for many graphs F . These embedding lemmas greatly extend a result of Chung and Graham [28] for the uniform case, but still leave much to be done: in particular, the case of F a triangle is open. Given the simplicity of dense quasi-random graphs, this is surprising: one form of the question is simply whether a sparse quasi-random graph must contain a triangle! For further results and conjectures related to these, see the extensive discussion in [21]. When np is bounded, corresponding to graphs with $\Theta(n)$ edges as discussed in most of this paper, the situation is even more complicated. We return to this in the next section.

9. Models and Metrics

Suppose we are trying to use a random graph as a (presumably simplified) model of some real-world network. Then we should have some way of telling whether the model is a good fit to the network. One approach is simply to compare some parameters of interest, for example the exponent of the power-law degree distribution (if any), clustering coefficient etc. However, this is rather unsatisfactory. The models often have several parameters, and it may well happen that one can fit arbitrary values for the observed data by a suitable choice of the model parameters. But then there is no reason to expect the model to fit the network in any other way, in which case it tells us nothing new.

It would be much more satisfactory to have a single global measure of similarity, that ideally captures all the kinds of features we are interested in. If a random graph from the model is typically close to a real-world example in this metric, that is a very strong indication that it is an appropriate model. Note that this is very different from the usual distribution fitting questions that arise in statistics: there one typically has many samples from an unknown distribution, with each sample being a single number (or a small list of numbers). In the present case one typically has only one real-world sample (for example, the world-wide web), but this single sample contains a large amount of information. So the question is not whether the model has the right distribution over all possible graphs, but whether a typical graph from the model is close to the example.

This question may not appear to make much sense: for example, what is a typical graph $G(n, 1/2)$ from the distribution $\mathcal{G}(n, 1/2)$? Simply a graph chosen uniformly at random from among all graphs on $[n]$. Suppose that we have a large graph in the real world which (perhaps due to some simple physical process) forms randomly exactly according to the rule defining $G(n, 1/2)$. Is there any chance that we can recognize this? The answer is yes, at least to some extent: typically (with high probability as $n \rightarrow \infty$), such a graph is very close to the constant kernel $1/2$ in the cut metric, so for a real network also close to this kernel, $G(n, 1/2)$ is presumably a good model.

More generally, for dense graphs, the cut metric provides an answer to the vague question above. As noted in the previous section, one can think of inhomogeneous quasi-random graphs as universal among (sequences of) dense graphs. In terms of the corresponding *random* models, which are much easier to work with, for each limit point there is such a model: the W -random graph of Lovász and Szegedy [49] described near the end of Section 4. Together, from the point of view of the cut metric and its many equivalents, this family of random models is in some sense universal: any sequence of graphs has a subsequence that is well approximated by a model in the family.

Turning to the sparse case, one can ask for analogues of the results above: for example, given a metric, one would like a family of random graph models such that for each point in the completion of the space of graphs under the metric,

there is a model such that graphs from the model converge (in probability, say) to this point. Conversely, given a model, one would like a suitable metric that can tell us whether a real-world network is well approximated by the model.

A complete answer to these questions is perhaps too much to ask for. Nevertheless, partial answers may well be interesting, for two reasons. Firstly, starting from a metric gives a new way of looking, hopefully systematically, for new random graph models, some of which may be more useful for applications than current models. Secondly, the search for new metrics may lead to a better understanding of the space of sparse graphs itself, as well as a more systematic way of testing the fit between models and examples.

We finish by summarizing some early steps in these directions taken by Bollobás and Riordan in [22], concentrating on the case of graphs with $\Theta(n)$ edges. Considering first which metric to use, let us note that the cut metric is useless in this context: as noted in [22], there simply aren't any non-trivial Cauchy sequences in this metric. (This is related to the fact that non-trivial ε -regular partitions do not exist in graphs with $\Theta(n)$ edges.) From the point of view of capturing global structure, however, there is a natural alternative: the *partition metric* d_{part} . This was defined in the fully dense case (normalizing function $p = 1$) by Borgs, Chayes, Lovász, Sós and Vesztegombi [26], and in general by Bollobás and Riordan [21]. Roughly speaking, the metric is defined by mapping a graph G_n to a *set* of density matrices: for each partition of the vertices into k parts with nearly equal sizes, there is a k -by- k matrix encoding the (normalized) densities of edges in G_n between each pair of parts. Two graphs are close in the metric if (for each fixed k) the sets of possible density matrices for the two graphs are close in the Hausdorff metric. In particular, taking the simplest example, if two graphs are close in d_{part} , then their maximum balanced cuts must contain almost the same number of edges.

As shown in [21], whenever $np \rightarrow \infty$, the partition metric is equivalent to the cut metric; its advantage is that it makes very good sense when $p = 1/n$. It is natural to ask whether the partition metric distinguishes, for example, different cases of the BJR model $G(n, \kappa)$: one could hope that for each kernel κ there is a unique limit point in d_{part} such that $G(n, \kappa)$ converges to this point as $n \rightarrow \infty$. (This is problematic not least for the reason that it is hard to see which kernels give rise to genuinely different random graphs; see [22].) Less ambitiously, it is conjectured in [22] that for each kernel, the sequence $G(n, \kappa)$ is Cauchy with respect to d_{part} with probability 1. This conjecture includes as a very special case the recent result of Bayati, Gamarnik and Tetali [8] that the size of the largest independent set in the classical random graph $G(n, c/n)$ is $\beta(c)n + o_p(n)$ for some constant $\beta(c)$.

Turning to local structure, there is an extremely natural notion of similarity in the sparse case, given by simply comparing subgraph counts normalized by dividing by n . These can be combined to form a metric d_{loc} . Equivalently, for each integer t and each fixed *rooted* graph F , one considers the fraction of

vertices v of G_n whose local neighbourhood up to distance t is isomorphic to F , with v playing the role of the root. The corresponding notion of convergence has appeared in several contexts: in the work of Benjamini and Schramm [10] for a sequence of random graphs, under the name ‘distributional limit’, and in work of Aldous and Steele [4], under the name ‘local weak limit’, and Aldous and Lyons [3], where the term ‘random weak limit’ is used. Here a potential limit point can be seen as a distribution on infinite rooted graphs. Any distribution that arises as the limit of a sequence of finite graphs necessarily has a certain ‘unimodularity’ property; Aldous and Lyons [3] conjecture (in a more general context) that any unimodular distribution is a limit. If true, this conjecture has many consequences; for example, it would essentially imply that all finitely generated groups are ‘sofic’. This group property was initially introduced (in a slightly different form) by Gromov [44]; the term ‘sofic’ was coined by Weiss [62]. The key point is that several well-known conjectures in group theory have been proved for sofic groups; see Elek and Szabó [33], for example.

Although describing all local limits seems very hard, the notion still motivates the introduction of interesting new graph models. For example, there is a limit point corresponding to graphs whose local structure is an r -regular tree of triangles; for this limit point the natural model is a triangle version of the configuration model of Bollobás [11]. Such a model (in an inhomogeneous form) has recently been introduced in a completely different context by Newman [52].

Finally, although we have considered local and global structure separately, we should like to combine them, as in the dense case (where the corresponding metrics coincide, so this happens naturally). A metric that does so was defined in [22]: this is the *coloured neighbourhood metric* d_{cn} , obtained as follows. Given a graph G_n , consider all colourings of G_n with k colours (corresponding to partitions). Rather than simply recording the number of edges between each pair of parts, instead, given an integer t and a coloured rooted graph F , record the fraction of vertices of G_n whose t -neighbourhood is isomorphic to F . This gives a set of distributions on coloured graphs; the metric is obtained by taking the Hausdorff distance between these sets, and combining these distances over all t and k . This metric jointly refines d_{part} and d_{loc} : a sequence that is Cauchy in the coloured neighbourhood metric is Cauchy in both. Describing all possible limit points for this metric seems very difficult, but it may be possible to obtain interesting partial results, and one can hope that these will suggest many fruitful new random graph models.

References

- [1] W. Aiello, F. Chung and L. Lu, A random graph model for power law graphs, *Experiment. Math.* **10** (2001), 53–66.
- [2] R. Albert and A.-L. Barabási, Statistical mechanics of complex networks, *Rev. Mod. Phys.* **74** (2002), 47–97.

-
- [3] D. Aldous and R. Lyons, Processes on unimodular random networks, *Electron. J. Probab.* **12** (2007), 1454–1508 (electronic).
- [4] D. Aldous and J.M. Steele, The objective method: probabilistic combinatorial optimization and local weak convergence, in *Probability on discrete structures, Encyclopaedia Math. Sci.* **110**, Springer (2004), pp. 1–72.
- [5] N. Alon, A note on network reliability, in *Discrete probability and algorithms (Minneapolis, MN, 1993)*, *IMA Vol. Math. Appl.* **72**, Springer, New York, 1995, pp. 11–14.
- [6] M. Axenovich, A. Kézdy and R. Martin, On the editing distance of graphs, *J. Graph Theory* **58** (2008), 123–138.
- [7] A.-L. Barabási and R. Albert, Emergence of scaling in random networks, *Science* **286** (1999), 509–512.
- [8] M. Bayati, D. Gamarnik and P. Tetali, Combinatorial approach to the interpolation method and scaling limits in sparse random graphs, preprint (2009), [arXiv:0912.2444](https://arxiv.org/abs/0912.2444).
- [9] E.A. Bender and E.R. Canfield, The asymptotic number of labeled graphs with given degree sequences, *J. Combinatorial Theory Ser. A* **24** (1978), 296–307.
- [10] I. Benjamini and O. Schramm, Recurrence of distributional limits of finite planar graphs, *Electron. J. Probab.* **6** (2001), no. 23, 13 pp. (electronic).
- [11] B. Bollobás, A probabilistic proof of an asymptotic formula for the number of labelled regular graphs, *European J. Combin.* **1** (1980), 311–316.
- [12] B. Bollobás, The evolution of random graphs, *Trans. Amer. Math. Soc.* **286** (1984), 257–274.
- [13] B. Bollobás, C. Borgs, J. Chayes and O. Riordan, Percolation on dense graph sequences, *Annals of Probability* **38** (2010), 150–183.
- [14] B. Bollobás and F.R.K. Chung, The diameter of a cycle plus a random matching, *SIAM J. Discrete Math.* **1** (1988), 328–333.
- [15] B. Bollobás, S. Janson and O. Riordan, The phase transition in inhomogeneous random graphs, *Random Structures and Algorithms* **31** (2007), 3–122.
- [16] B. Bollobás, S. Janson and O. Riordan, Sparse random graphs with clustering, to appear in *Random Structures and Algorithms*. [arXiv:0807.2040](https://arxiv.org/abs/0807.2040).
- [17] B. Bollobás, S. Janson and O. Riordan, The cut metric, random graphs, and branching processes, preprint (2009). [arXiv:0901.2091](https://arxiv.org/abs/0901.2091).
- [18] B. Bollobás and O. Riordan, Robustness and vulnerability of scale-free random graphs, *Internet Mathematics* **1** (2003), 1–35.
- [19] B. Bollobás and O. Riordan, The diameter of a scale-free random graph, *Combinatorica* **24** (2004), 5–34.
- [20] B. Bollobás and O. Riordan, Random graphs and branching processes, in *Handbook of large-scale random networks, Bolyai Soc. Math. Stud* **18**, B. Bollobás, R. Kozma and D. Miklós eds (2009), pp. 15–115.
- [21] B. Bollobás and O. Riordan, Metrics for sparse graphs, in *Surveys in Combinatorics 2009, London Math. Soc. Lecture Note Series* **365**, S. Huczynska, J.D. Mitchell and C.M. Roney-Dougal eds, CUP (2009), pp. 211–287.

- [22] B. Bollobás and O. Riordan, Sparse graphs: metrics and random models, to appear in *Random Structures and Algorithms*. [arXiv:0812.2656](https://arxiv.org/abs/0812.2656).
- [23] B. Bollobás, O. Riordan, J. Spencer and G. Tusnády, The degree sequence of a scale-free random graph process, *Random Structures and Algorithms* **18** (2001), 279–290.
- [24] C. Borgs, J.T. Chayes and L. Lovász, Moments of two-variable functions and the uniqueness of graph limits, *Geom. Funct. Anal.* **19** (2010), 1597–1619.
- [25] C. Borgs, J.T. Chayes, L. Lovász, V.T. Sós and K. Vesztegombi, Convergent sequences of dense graphs I: Subgraph frequencies, metric properties and testing, *Advances in Math.* **219** (2008), 1801–1851.
- [26] C. Borgs, J.T. Chayes, L. Lovász, V.T. Sós and K. Vesztegombi, Convergent sequences of dense graphs II: Multiway cuts and statistical physics, preprint (2007). <http://www.cs.elte.hu/~lovasz/ConvRight.pdf>.
- [27] S.R. Broadbent and J.M. Hammersley, Percolation processes. I. Crystals and mazes, *Proc. Cambridge Philos. Soc.* **53** (1957), 629–641.
- [28] F. Chung and R. Graham, Sparse quasi-random graphs, *Combinatorica* **22** (2002), 217–244.
- [29] F.R.K. Chung, R.L. Graham and R.M. Wilson, Quasi-random graphs, *Combinatorica* **9** (1989), 345–362.
- [30] P. Diaconis and S. Janson, Graph limits and exchangeable random graphs, *Rendiconti di Matematica* **28** (2008), 33–61.
- [31] S.N. Dorogovtsev, Lectures on complex networks, *Oxford Master Series in Physics* **20**, Oxford University Press (2010).
- [32] S.N. Dorogovtsev and J.F.F. Mendes, Evolution of networks, *Adv. Phys.* **51** (2002), 1079–1187.
- [33] G. Elek and E. Szabó, Sofic groups and direct finiteness, *J. Algebra* **280** (2004), 426–434.
- [34] P. Erdős, On some new inequalities concerning extremal properties of graphs, in *Theory of Graphs* (Proc. Colloq., Tihany, 1966), Academic Press, New York, 1968, pp. 77–81.
- [35] P. Erdős and A. Rényi, On random graphs I., *Publicationes Mathematicae Debrecen* **5** (1959), 290–297.
- [36] P. Erdős and A. Rényi, On the evolution of random graphs, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **5** (1960), 17–61.
- [37] P. Erdős and A. Rényi, On the strength of connectedness of a random graph, *Acta Math. Acad. Sci. Hungar.* **12** (1961), 261–267.
- [38] P. Erdős and A. Rényi, On random matrices, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* **8** (1964), 455–461.
- [39] P. Erdős and A. Rényi, On the existence of a factor of degree one of a connected random graph, *Acta Math. Acad. Sci. Hungar.* **17** (1966), 359–368.
- [40] P. Erdős and A. Rényi, On random matrices. II, *Studia Sci. Math. Hungar.* **3** (1968), 459–464.
- [41] M. Faloutsos, P. Faloutsos and C. Faloutsos, On power-law relationships of the internet topology, SIGCOMM 1999, *Comput. Commun. Rev.* **29** (1999), 251–262.

- [42] A. Frieze and R. Kannan, Quick approximation to matrices and applications, *Combinatorica* **19** (1999), 175–220.
- [43] E.N. Gilbert, Random graphs, *Annals of Mathematical Statistics* **30** (1959), 1141–1144.
- [44] M. Gromov, Endomorphisms of symbolic algebraic varieties, *J. Eur. Math. Soc.* **1** (1999), 109–197.
- [45] S. Janson, On a random graph related to quantum theory, *Combin. Probab. Comput.* **16** (2007), 757–766.
- [46] S. Janson, D.E. Knuth, T. Łuczak and B. Pittel, The birth of the giant component, *Random Structures Algorithms* **4** (1993), 231–358.
- [47] J.M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan and A.S. Tomkins, The web as a graph: Measurements, models, and methods, in *Proceedings of COCOON '99, Lecture Notes in Computer Science* **1627** (1999), 1–17.
- [48] Y. Kohayakawa and V. Rödl, Szemerédi's regularity lemma and quasi-randomness, in *Recent advances in algorithms and combinatorics, CMS Books Math.* **11**, Springer (2003), pp. 289–351.
- [49] L. Lovász and B. Szegedy, Limits of dense graph sequences, *J. Combin. Theory B* **96** (2006), 933–957.
- [50] L. Lovász and B. Szegedy, Szemerédi's Lemma for the analyst, *Geom. Funct. Anal.* **17** (2007), 252–270.
- [51] T. Łuczak, Component behavior near the critical point of the random graph process, *Random Structures Algorithms* **1** (1990), 287–310.
- [52] M.E.J. Newman, Random graphs with clustering, *Phys. Rev. Lett.* **103** (2009), 058701 [4 pages].
- [53] O. Riordan, Spanning subgraphs of random graphs, *Combin. Probab. Comput.* **9** (2000), 125–148.
- [54] O. Riordan, The small giant component in scale-free random graphs, *Combin. Probab. Comput.* **14** (2005), 897–938.
- [55] M. Simonovits, A method for solving extremal problems in graph theory, stability problems, in *Theory of Graphs* (Proc. Colloq., Tihany, 1966), Academic Press, New York, 1968, pp. 279–319.
- [56] B. Söderberg, General formalism for inhomogeneous random graphs, *Phys. Rev. E* **66** (2002), 066121 [6 pages].
- [57] R. Solomonoff and A. Rapoport, Connectivity of random nets, *Bull. Math. Biophys.* **13** (1951), 107–117.
- [58] E. Szemerédi, Regular partitions of graphs, in *Problèmes combinatoires et théorie des graphes* (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976), *Colloq. Internat. CNRS* **260**, CNRS, Paris (1978), pp. 399–401.
- [59] A. Thomason, Pseudorandom graphs, in *Proceedings of Random Graphs* (M. Karonski, ed.), Poznań, 1985, *Annals of Discrete Mathematics*, **33** (1987) 307–331.

- [60] A. Thomason, Random graphs, strongly regular graphs and pseudorandom graphs, in *Surveys in Combinatorics 1987, London Math. Soc. Lecture Note Ser. 123*, Cambridge Univ. Press, Cambridge (1987), pp 173–195.
- [61] D. J. Watts and S. H. Strogatz, Collective dynamics of ‘small-world’ networks, *Nature* **393** (1998), 440–442.
- [62] B. Weiss, Sofic groups and dynamical systems, *Sankhyā Ser. A* **62** (2000), 350–359.

Recent Developments in Extremal Combinatorics: Ramsey and Turán Type Problems

Benny Sudakov*

Abstract

Extremal combinatorics is one of the central branches of discrete mathematics and has experienced an impressive growth during the last few decades. It deals with the problem of determining or estimating the maximum or minimum possible size of a combinatorial structure which satisfies certain requirements. Often such problems are related to other areas including theoretical computer science, geometry, information theory, harmonic analysis and number theory. In this paper we discuss some recent advances in this subject, focusing on two topics which played an important role in the development of extremal combinatorics: Ramsey and Turán type questions for graphs and hypergraphs.

Mathematics Subject Classification (2010). 05C35, 05C65, 05D10, 05D40

Keywords. Extremal combinatorics, Ramsey theory, Turán problems, Probabilistic methods

1. Introduction

Discrete mathematics (or combinatorics) is a fundamental mathematical discipline which focuses on the study of discrete objects and their properties. Although it is probably as old as the human ability to count, the field experienced tremendous growth during the last fifty years and has matured into a thriving area with its own set of problems, approaches and methodologies. The development of powerful techniques, based on ideas from probability, algebra, harmonic analysis and topology, is one of the main reasons for the rapid growth

*Research supported in part by NSF CAREER award DMS-0812005 and by USA-Israeli BSF grant.

Department of Mathematics, UCLA, Los Angeles, CA 90095.
E-mail: bsudakov@math.ucla.edu.

of combinatorics. Such tools play an important organizing role in combinatorics, similar to the one that deep theorems of great generality play in more classical areas of mathematics.

Extremal combinatorics is one of the central branches of discrete mathematics. It deals with the problem of determining or estimating the maximum or minimum possible cardinality of a collection of finite objects (e.g., numbers, graphs, vectors, sets, etc.) satisfying certain restrictions. Often such problems appear naturally in other areas, and one can find applications of extremal combinatorics in theoretical computer science, geometry, information theory, analysis, and number theory. Extremal combinatorics has developed spectacularly in the last few decades, and two topics which played a very important role in its development are Ramsey theory and Turán type problems for graphs and hypergraphs.

The foundations of Ramsey theory rest on the following general phenomenon: every large object, chaotic as it may be, contains a sub-object that is guaranteed to be well structured, in a certain appropriately chosen sense. This phenomenon is truly ubiquitous and manifests itself in different mathematical areas, ranging from the most basic Pigeonhole principle to intricate statements from set theory. Extremal theory of graphs and hypergraphs considers problems such as the maximum possible number of edges in a triangle-free graph with a given number of vertices. The development of this subject was instrumental in turning Graph Theory into a modern, deep and versatile field.

Both areas use a variety of sophisticated methods and arguments (for example, algebraic and probabilistic considerations, geometric constructions, the stability approach and the regularity method) and there is a considerable overlap between them. Indeed, Ramsey theory studies which configurations one can find in every finite partition of the large structure. On the other hand, extremal graph theory deals with the inevitable occurrence of some specified configuration when the edge density of graph or hypergraph exceeds a certain threshold.

In this paper we survey recent progress on some classical Ramsey and Turán type problems, focusing on the basic ideas and connections to other fields. It is of course impossible to cover everything in such a short article, and therefore the choice of results we present is inevitably biased. Yet we hope to describe enough examples, problems and techniques from this fascinating subject to appeal to researchers not only in discrete mathematics but in other areas as well.

2. Ramsey Theory

Ramsey theory refers to a large body of deep results in mathematics whose underlying philosophy is captured succinctly by the statement that “Every large system contains a large well organized subsystem.” This is an area in which a great variety of techniques from many branches of mathematics are used and whose results are important not only to combinatorics but also to logic, analysis,

number theory, and geometry. Since the publication of the seminal paper of Ramsey [81] in 1930, this subject has experienced an impressive growth, and is currently among the most active areas in combinatorics.

The *Ramsey number* $r_k(s_1, s_2, \dots, s_\ell)$ is the least integer N such that every ℓ -coloring of the unordered k -tuples of an N -element set contains a monochromatic set of size s_i in color i for some $1 \leq i \leq \ell$, where a set is called monochromatic if all k -tuples from this set have the same color. Ramsey's theorem states that these numbers exist for all values of the parameters. In the case of graphs (i.e., $k = 2$) it is customary to omit the index k and to write simply $r(s_1, \dots, s_\ell)$.

Originally, Ramsey applied his result to a problem in logic, but his theorem has many additional applications. For example, the existence of the ℓ -colored Ramsey number $r(3, 3, \dots, 3)$ can be used to deduce the classical theorem of Schur from 1916. Motivated by Fermat's last theorem, he proved that any ℓ -coloring of a sufficiently large initial segment of natural numbers contains a monochromatic solution of the equation $x + y = z$. Another application of this theorem to geometry was discovered by Erdős and Szekeres [48]. They showed that any sufficiently large set of points in the plane in general position (no 3 of which are collinear) contains the vertices of a convex n -gon. They deduced this result from Ramsey's theorem together with the simple fact that any 5 points in general position contain a convex 4-gon. Another early result of Ramsey theory is van der Waerden's theorem, which says that every finite coloring of the integers contains arbitrarily long arithmetic progressions. The celebrated density version of this theorem, proved by Szemerédi [98], has led to many deep and beautiful results in various areas of mathematics, including the recent spectacular result of Green and Tao [63] that there are arbitrarily long arithmetic progressions in the primes.

Determining or estimating Ramsey numbers is one of the central problems in combinatorics, see [62] for details. Erdős and Szekeres [48] proved a quantitative version of Ramsey's theorem showing that $r(s, n) \leq \binom{n+s-2}{s-1}$. To prove this simple statement, one can fix a vertex and, depending on the number of its neighbors in colors 1 and 2, apply an induction to one of these two sets. In particular, for the diagonal case when $s = n$ it implies that $r(n, n) \leq 2^{2n}$ for every positive integer n . The first exponential lower bound for this numbers was obtained by Erdős [32], who showed that $r(n, n) > 2^{n/2}$ for $n > 2$. His proof, which is one of the first applications of probabilistic methods in combinatorics is extremely short. The probability that a random 2-edge coloring of the complete graph K_N on $N = 2^{n/2}$ vertices contains a monochromatic set of size n is at most $\binom{N}{n} 2^{1-\binom{n}{2}} < 1$. Hence there is a coloring with the required properties. Although the proofs of both bounds for $r(n, n)$ are elementary, obtaining significantly better results appears to be notoriously difficult. Over the last sixty years, there have been several improvements on these estimates (most recently by Conlon, in [22]), but the constant factors in the above exponents remain the same. Improving these exponents is a very fundamental problem and will probably require novel techniques and ideas. Such

techniques will surely have many applications to other combinatorial problems as well.

The probabilistic proof of Erdős [32], described above, leads to another important open problem which seems very difficult. Can one explicitly construct for some fixed $\epsilon > 0$ a 2-edge coloring of the complete graph on $N > (1 + \epsilon)^n$ vertices with no monochromatic clique of size n ? *Explicit* here means that there is an algorithm which produces the coloring in polynomial time in the number of its vertices. Despite a lot of efforts this question is still open. For many years the best known result was due to Frankl and Wilson [54] who gave an elegant explicit construction of such a coloring on $n^{c \frac{\log n}{\log \log n}}$ vertices for some fixed $c > 0$. (All logarithms in this paper are in base e unless otherwise stated.) Recently a new approach to this problem and its bipartite variant was proposed in [9, 10]. In particular, for any constant C the algorithm of Barak, Rao, Shaltiel and Wigderson efficiently constructs a 2-edge coloring of the complete graph on $n^{\log^C n}$ vertices with no monochromatic clique of size n .

Off-diagonal Ramsey numbers, i.e., $r(s, n)$ with $s \neq n$, have also been intensely studied. After several successive improvements, the asymptotic behavior of $r(3, n)$ was determined by Kim [69] and by Ajtai, Komlos and Szemerédi [1].

Theorem 2.1. *There are absolute constants c_1 and c_2 such that*

$$c_1 n^2 / \log n \leq r(3, n) \leq c_2 n^2 / \log n.$$

This is an important result, which gives an infinite family of Ramsey numbers that are known up to a constant factor. The upper bound of [1] is proved by analyzing a certain randomized greedy algorithm. The lower bound construction of [69] uses a powerful *semi-random* method, which generates it through many iterations, applying probabilistic reasoning at each step. The analysis of this construction is subtle and is based on large deviation inequalities.

For $s > 4$ we only have estimates for $r(s, n)$ which are far apart. From the results of [1, 92] it follows that

$$c_3 \left(\frac{n}{\log n} \right)^{(s+1)/2} \leq r(s, n) \leq c_4 \frac{n^{s-1}}{\log^{s-2} n}, \quad (1)$$

for some absolute constants $c_3, c_4 > 0$. Recently, by analyzing the asymptotic behavior of certain random graph processes, Bohman [12] gave a new proof of the lower bound for $r(3, n)$. Together with Keevash [13], they used this approach to improve the above lower bound for $r(s, n)$ by a factor of $\log^{1/(s-2)} n$.

2.1. Hypergraphs. Although already for graph Ramsey numbers there are significant gaps between the lower and upper bounds, our knowledge of hypergraph Ramsey numbers ($k \geq 3$) is even weaker. Recall that $r_k(s, n)$ is the minimum N such that every red-blue coloring of the k -tuples of an N -element

set contains a red set of size s or a blue set of size n . Erdős, Hajnal, and Rado [43] showed that there are positive constants c and c' such that

$$2^{cn^2} < r_3(n, n) < 2^{2^{c'n}}.$$

They also conjectured that $r_3(n, n) > 2^{2^{cn}}$ for some constant $c > 0$ and Erdős offered a \$500 reward for a proof. Similarly, for $k \geq 4$, there is a difference of one exponential between the known upper and lower bounds for $r_k(n, n)$, i.e.,

$$t_{k-1}(cn^2) \leq r_k(n, n) \leq t_k(c'n),$$

where the tower function $t_k(x)$ is defined by $t_1(x) = x$ and $t_{i+1}(x) = 2^{t_i(x)}$.

The study of 3-uniform hypergraphs is particularly important for our understanding of hypergraph Ramsey numbers. This is because of an ingenious construction called the stepping-up lemma due to Erdős and Hajnal (see, e.g., Chapter 4.7 in [62]). Their method allows one to construct lower bound colorings for uniformity $k + 1$ from colorings for uniformity k , effectively gaining an extra exponential each time it is applied. Unfortunately, the smallest k for which it works is $k = 3$. Therefore, proving that $r_3(n, n)$ has doubly exponential growth will allow one to close the gap between the upper and lower bounds for $r_k(n, n)$ for all uniformities k . There is some evidence that the growth rate of $r_3(n, n)$ is closer to the upper bound, namely, that with four colors instead of two this is known to be true. Erdős and Hajnal (see, e.g., [62]) constructed a 4-coloring of the triples of a set of size $2^{2^{cn}}$ which does not contain a monochromatic subset of size n . This is sharp up to the constant factor c in the exponent. It also shows that the number of colors matters a lot in this problem and leads to the question of what happens in the intermediate case when we use three colors. In this case, Erdős and Hajnal have made some improvement on the lower bound 2^{cn^2} (see [42, 20]), showing that $r_3(n, n, n) \geq 2^{cn^2 \log^2 n}$. Recently, extending the above mentioned stepping-up lemma approach, the author, together with Conlon and Fox [25], gave a strong indication that $r_3(n, n, n)$ is probably also double-exponential.

Theorem 2.2. *There is a constant $c > 0$ such that*

$$r_3(n, n, n) \geq 2^{n^{c \log n}}.$$

A simple induction approach which was used to estimate $r(s, n)$ gives extremely poor bounds for off-diagonal hypergraph Ramsey numbers when $k \geq 3$. In 1952 Erdős and Rado [45] gave an intricate argument which shows how to bound the Ramsey numbers for uniformity k using estimates for uniformity $k - 1$. They proved that

$$r_k(s, n) \leq 2^{\binom{r_{k-1}(s-1, n-1)}{k-1}}. \quad (2)$$

Together with the upper bound in (1) this gives, for fixed s , that

$$r_3(s, n) \leq 2^{\binom{r(s-1, n-1)}{2}} \leq 2^{cn^{2s-4} / \log^{2s-6} n}.$$

Progress on this problem was slow and for several decades this was the best known bound. In [25], the authors discovered an interesting connection between the problem of bounding $r_3(s, n)$ and a new game-theoretic parameter, which we describe next.

Consider the following game, played by two players, the builder and the painter: at step $i + 1$ a new vertex v_{i+1} is revealed; then, for every existing vertex v_j , $j = 1, \dots, i$, the builder decides, in order, whether to draw the edge $v_j v_{i+1}$; if he does expose such an edge, the painter has to color it either red or blue immediately. The *vertex on-line Ramsey number* $\tilde{r}(k, l)$ is then defined as the minimum number of edges that the builder has to draw in order to force the painter to create a red K_k or a blue K_l . It appears that one can bound the Ramsey number $r_3(s, n)$ roughly by exponential in $\tilde{r}(s - 1, n - 1)$ and also provide an upper bound on $\tilde{r}(s - 1, n - 1)$ which is much smaller than the best known estimate on $\binom{r(s-1, n-1)}{2}$. These facts together with some additional ideas were used in [25] to show the following result, which improves the exponent of the upper bound by a factor of $n^{s-2}/\text{polylog } n$.

Theorem 2.3. *For fixed $s \geq 4$ and sufficiently large n , there exists a constant $c > 0$ such that*

$$r_3(s, n) \leq 2^{cn^{s-2} \log n}.$$

A similar improvement for off-diagonal Ramsey numbers of higher uniformity follows from this result together with (2).

Clearly one should also ask, how accurate are these estimates? For the first nontrivial case when $s = 4$, this problem was first considered by Erdős and Hajnal [41] in 1972. Using the following clever construction they showed that $r_3(4, n)$ is exponential in n . Consider a random tournament with vertex set $[N] = \{1, \dots, N\}$. This is a complete graph on N vertices whose edges are oriented uniformly at random. Color the triples from $[N]$ red if they form a cyclic triangle and blue otherwise. Since it is well known and easy to show that every tournament on four vertices contains at most two cyclic triangles and a random tournament on N vertices with high probability does not contain a transitive subtournament of size $c' \log N$, the resulting coloring neither has a red set of size 4 nor a blue set of size $c' \log N$. In the same paper [41], Erdős and Hajnal conjectured that $\frac{\log r_3(4, n)}{n} \rightarrow \infty$. This was recently confirmed in [25], where the authors obtained a more general result which in particular implies that $r_3(4, n) \geq 2^{cn \log n}$. This should be compared with the above upper bound that $r_3(4, n) \leq 2^{cn^2 \log n}$.

2.2. Almost monochromatic subsets. Despite the fact that Erdős [36, 20] believed $r_3(n, n)$ is closer to $2^{2^{cn}}$, he discovered together with Hajnal [42] the following interesting fact which they thought might indicate the opposite. They proved that there are $c, \epsilon > 0$ such that every 2-coloring of the triples of an N -element set contains a subset S of size $s > c(\log N)^{1/2}$ such that at least $(1/2 + \epsilon) \binom{s}{3}$ triples of S have the same color. That is, this subset deviates from

having density $1/2$ in each color by at least some fixed positive constant. Erdős ([37], page 67) further remarks that he would begin to doubt that $r_3(n, n)$ is double-exponential in n if one could prove that any 2-coloring of the triples of an N -set contains some set of size $s = c(\epsilon)(\log N)^\delta$ for which at least $(1 - \epsilon)\binom{s}{3}$ triples have the same color, where $\delta > 0$ is an absolute constant and $\epsilon > 0$ is arbitrary. Erdős and Hajnal proposed [42] that such a statement may even be true with $\delta = 1/2$. The following result in [26] shows that this is indeed the case.

Theorem 2.4. *For each $\epsilon > 0$ and ℓ , there is $c = c(\ell, \epsilon) > 0$ such that every ℓ -coloring of the triples of an N -element set contains a subset S of size $s = c\sqrt{\log N}$ such that at least $(1 - \epsilon)\binom{s}{3}$ triples of S have the same color.*

A random ℓ -coloring of the triples of an N -element set in which every triple gets one of ℓ colors uniformly at random shows that this theorem is tight up to the constant factor c . Indeed, using a standard tail estimate for the binomial distribution, one can show that in this coloring, with high probability, every subset of size $\gg \sqrt{\log N}$ has a $1/\ell + o(1)$ fraction of its triples in each color.

The above theorem shows a significant difference between the discrepancy problem in graphs and that in hypergraphs. As we already mentioned in the previous section, Erdős and Hajnal constructed a 4-coloring of the triples of an N -element set which does not contain a monochromatic subset of size $c \log \log N$. Also, by Theorem 2.2, there is a 3-coloring of the triples which does not contain a monochromatic subset of size $2^{c\sqrt{\log \log N}}$. Thus, Theorem 2.4 demonstrates (at least for $\ell \geq 3$) that the maximum almost monochromatic subset that an ℓ -coloring of the triples must contain is much larger than the corresponding monochromatic subset. This is in a striking contrast with graphs, where these two quantities have the same order of magnitude, as demonstrated by a random ℓ -coloring of the edges of a complete graph.

It would be very interesting to extend Theorem 2.4 to uniformity $k \geq 4$. In [25] the authors proved that for all k, ℓ and $\epsilon > 0$ there is $\delta = \delta(k, \ell, \epsilon) > 0$ such that every ℓ -coloring of the k -tuples of an N -element set contains a subset of size $s = (\log N)^\delta$ which contains at least $(1 - \epsilon)\binom{s}{k}$ k -tuples of the same color. Unfortunately, δ here depends on ϵ . On the other hand, this result probably holds even with $\delta = 1/(k - 1)$ (which is the case for $k = 3$).

3. Graph Ramsey Theory

The most famous question in Ramsey Theory is probably that of estimating $r(n, n)$. Since this problem remains largely unsolved with very little progress over the last 60 years, the focus of the field has shifted to the study of general graphs. Given an arbitrary fixed graph G , the Ramsey number $r(G)$ is the smallest integer N such that any 2-edge coloring of the complete graph K_N contains a monochromatic copy of G . For the classical Ramsey numbers G itself is taken to be a complete graph K_n . When ℓ colors are used to color the

edges of K_N instead of two, we will denote the corresponding value of N by $r(G; \ell)$. The original motivation to study Ramsey numbers of general graphs was the hope that it would eventually lead to methods that would give better estimates for $r(n, n)$. While this hope has not been realized, a beautiful subject has emerged with many fascinating problems and results. Graph Ramsey Theory, which started about 35 years ago, quickly became one of the most active areas of Ramsey theory. Here we discuss several problems which have played an important role in this development.

3.1. Linear Ramsey numbers. Among the most interesting questions about Ramsey numbers are the linear bounds for graphs with certain degree constraints. In 1975, Burr and Erdős [17] conjectured that, for each positive integer Δ , there is a constant $c(\Delta)$ such that every graph G with n vertices and maximum degree Δ satisfies $r(G) \leq c(\Delta)n$. This conjecture was proved by Chvatál, Rödl, Szemerédi, and Trotter [21]. Their proof is a beautiful illustration of the power of Szemerédi's celebrated regularity lemma (see, e.g., [70]). Remarkably, this means that for graphs of fixed maximum degree the Ramsey number only has a linear dependence on the number of vertices. Because the original method used the regularity lemma, it gave tower type bound on $c(\Delta)$. More precisely, $c(\Delta)$ was bounded by exponential tower of 2s with a height that is itself exponential in Δ . Since then, the problem of determining the correct order of magnitude of $c(\Delta)$ as a function of Δ has received considerable attention from various researchers.

The situation was remedied somewhat by Eaton, who proved, still using a variant of the regularity lemma, that the function $c(\Delta)$ can be taken to be of the form $2^{2^{c\Delta}}$ (here and later in this section c is some absolute constant). A novel approach of Graham, Rödl, and Rucinski [60] gave the first linear upper bound on Ramsey numbers of bounded degree graphs without using any form of the regularity lemma. Their proof implies that $c(\Delta) < 2^{c\Delta \log^2 \Delta}$. In [61], they also proved that there are bipartite graphs with n vertices and maximum degree Δ for which the Ramsey number is at least $2^{c'\Delta}n$. Recently, refining their approach further, together with Conlon and Fox [27] the author proved that

$$c(\Delta) < 2^{c\Delta \log \Delta},$$

which brings it a step closer to the lower bound.

The case of bipartite graphs with bounded degree was studied by Graham, Rödl, and Rucinski more thoroughly in [61], where they improved their upper bound, showing that $r(G) \leq 2^{c\Delta \log \Delta}n$ for every bipartite graph G with n vertices and maximum degree Δ . Using a totally different approach, Conlon [23] and, independently, Fox and Sudakov [51] have shown how to remove the $\log \Delta$ factor in the exponent, achieving an essentially best possible bound of $r(G) \leq 2^{c\Delta}n$ in the bipartite case. This gives strong evidence that in the general case $c(\Delta)$ should also be exponential in Δ . The bound proved in [51] has the following form (the estimate in [23] is slightly weaker).

Theorem 3.1. *If G is a bipartite graph with n vertices and maximum degree $\Delta \geq 1$, then*

$$r(G) \leq \Delta 2^{\Delta+5} n.$$

One family of bipartite graphs that has received particular attention are the d -cubes. The d -cube Q_d is the d -regular graph with 2^d vertices whose vertex set is $\{0, 1\}^d$ and two vertices are adjacent if they differ in exactly one coordinate. More than 30 years ago, Burr and Erdős [17] conjectured that the Ramsey number $r(Q_d)$ is linear in the number of vertices of the d -cube, i.e., there exists an absolute constant $c > 0$ such that $r(Q_d) \leq c2^d$. Since then, several authors have improved the upper bound for $r(Q_d)$, but the problem is still open. Beck [11] proved that $r(Q_d) \leq 2^{cd^2}$. The bound of Graham et al. [61] shows that $r(Q_d) \leq 8(16d)^d$. Using ideas from [72], Shi [88] proved the first exponential bound $r(Q_d) \leq 2^{cd}$, with exponent $c = (1 + o(1)) \frac{3+\sqrt{5}}{2} \approx 2.618$. A very special case of Theorem 3.1, when $G = Q_d$, gives immediately that for every positive integer d ,

$$r(Q_d) \leq d2^{2d+5},$$

which is roughly quadratic in the number of vertices of the d -cube.

Given the recent advances in developing the hypergraph regularity method it was natural to expect that linear bounds might also be provable for Ramsey numbers of bounded degree k -uniform hypergraphs. Such a result was indeed established for general k in [30] (extending two earlier proofs for $k = 3$). A short proof of this result, not using regularity and thus giving much better bounds was obtained in [24]. It is based on the approach from [23, 51] used to prove Theorem 3.1.

3.2. Sparse graphs. A graph is d -degenerate if every subgraph of it has a vertex of degree at most d . This notion nicely captures the concept of sparse graphs as every t -vertex subgraph of a d -degenerate graph has at most td edges. (Indeed, remove from the subgraph a vertex of minimum degree, and repeat this process in the remaining subgraph.) Notice that graphs with maximum degree d are d -degenerate. On the other hand, it is easy to construct a d -degenerate graph on n vertices whose maximum degree is linear in n . One of the most famous open problems in Graph Ramsey Theory is the following conjecture of Burr and Erdős [17] from 1975.

Conjecture 3.2. *For each positive integer d , there is a constant $c(d)$ such that $r(G) \leq c(d)n$ for every d -degenerate graph G on n vertices.*

This difficult conjecture is a substantial generalization of the results on Ramsey numbers of bounded degree graphs from Section 3.1 and progress on this problem was made only recently.

Kostochka and Rödl [73] gave a polynomial upper bound on the Ramsey numbers of d -degenerate graphs. The first nearly linear bound for this conjecture was obtained, by Kostochka and the author, in [74]. They proved that

d -degenerate graphs on n vertices satisfy $r(G) \leq c_d n^{1+\epsilon}$ for any fixed $\epsilon > 0$. The following is the best current estimate, which appeared in [52].

Theorem 3.3. *For each positive integer d there is a constant c_d such that every d -degenerate graph G with order n satisfies $r(G) \leq 2^{c_d \sqrt{\log n}} n$.*

In the past two decades Conjecture 3.2 was also proved for some special families of d -degenerate graphs (see, e.g., [2, 18, 85]). For example, we know that planar graphs and more generally graphs which can be drawn on a surface of bounded genus have linear Ramsey numbers. One very large and natural family of d -degenerate graphs are sparse random graphs. The *random graph* $G_{n,p}$ is the probability space of labeled graphs on n vertices, where every edge appears independently with probability p . When $p = d/n$ it is easy to show using standard large deviation estimates for binomial distribution that with high probability $G_{n,p}$ is $O(d)$ -degenerate. Hence it is natural to test the above conjecture on random graphs. This was done in [52], where it was proved that sparse random graphs do indeed have typically linear Ramsey numbers.

3.3. Maximizing the Ramsey number. Another related problem on Ramsey numbers of general graphs was posed in 1973 by Erdős and Graham. Among all graphs with m edges, they wanted to find a graph G with maximum Ramsey number. Since the results we mentioned so far clearly show that sparse graphs have slowly growing Ramsey numbers, one would probably like to make such a G as dense as possible. Indeed, Erdős and Graham [40] conjectured that among all the graphs with $m = \binom{n}{2}$ edges (and no isolated vertices), the complete graph on n vertices has the largest Ramsey number. This conjecture is very difficult and so far there has been no progress on this problem. Because of the lack of progress, in the early 80s Erdős [35] (see also [20]) asked whether one could at least show that the Ramsey number of any graph with m edges is not substantially larger than that of the complete graph with the same size. Since the number of vertices in a complete graph with m edges is a constant multiple of \sqrt{m} , Erdős conjectured that $r(G) \leq 2^{c\sqrt{m}}$ for every graph G with m edges and no isolated vertices. The authors of [3] showed that for all graphs with m edges $r(G) \leq 2^{c\sqrt{m} \log m}$ and also proved this conjecture in the special case when G is bipartite. Recently, Erdős' conjecture was established in full generality in [95].

Theorem 3.4. *If G is a graph on m edges without isolated vertices, then $r(G) \leq 2^{250\sqrt{m}}$.*

This theorem is best possible up to a constant factor in the exponent, since a complete graph with m edges has Ramsey number at least $2^{\sqrt{m/2}}$. Based on the results from Section 3.1, it seems plausible that the following strengthening of Conjecture 3.2 holds as well. For all d -degenerate graphs G on n vertices, $r(G) \leq 2^{cd}n$. Such a bound would be a far-reaching generalization of the estimates on Ramsey numbers of bounded-degree graphs and also of Theorem 3.4. Indeed, it is easy to check that every graph with m edges is $\sqrt{2m}$ -degenerate.

3.4. Methods. The result of Chvatál et al. [21] which gave the first linear bound on Ramsey numbers of bounded degree graphs (see Section 3.1), was proved using the regularity lemma. This is a surprising and extremely powerful result proved by Szemerédi that has numerous applications in various areas including combinatorial number theory, computational complexity, and mainly extremal graph theory. The regularity lemma was an essential tool in the proof of the celebrated theorem of Szemerédi that any dense subset of integers contains long arithmetic progressions. The precise statement of the lemma is somewhat technical and can be found in [70] together with the description of several of its famous applications.

Roughly this lemma states that the vertices of every large enough graph can be partitioned into a finite number of parts such that the edges between almost all of the parts behave like a random graph. The strength of the regularity lemma is that it applies to every graph and provides a good approximation of its structure which enables one to extract a lot of information about it. It is also known that there is an efficient algorithm for finding such a regular partition. Although the regularity lemma is a great tool for proving qualitative statements, the quantitative bounds which one usually gets from such proofs are rather weak. This is because the number of parts M in the partition of the graph given by the regularity lemma may be very large, more precisely of tower type. Moreover, Gowers [57] constructed examples of graphs for which M has to grow that fast. Therefore, to obtain good quantitative estimates, one should typically use a different approach.

One such approach was proposed by Graham, Rödl, and Rucinski [60] (see also [50] for some extensions). They noticed that in some applications, instead of having tight control on the distribution of edges (which the regularity lemma certainly gives), it is enough to satisfy a bi-density condition, i.e., to have a lower bound on the density of edges between any two sufficiently large disjoint sets. Using this observation one can show that in every red-blue edge coloring of K_N , either the red color satisfies a certain bi-density condition or there is a large set in which the proportion of blue edges is very close to 1. Then, for example, one can find a blue copy of any bounded-degree graph in this almost blue set. On the other hand, this approach is highly specific to the 2-color case and it would be of considerable interest to make it work for k colors.

Another basic tool used to prove several results mentioned in Sections 3.1-3.3 as well as some other recent striking results in extremal combinatorics is a simple and yet surprisingly powerful lemma, whose proof is probabilistic. Early variants of this lemma, have been proved and applied by various researchers starting with Rödl, Gowers, Kostochka and Sudakov (see [72], [58], [94]).

The lemma asserts, roughly, that every graph with sufficiently many edges contains a large subset U in which every set of d vertices has many common neighbors. The proof uses a process that may be called a *dependent random choice* for finding the set U ; U is simply the set of all common neighbors of an appropriately chosen random set R . Intuitively, it is clear that if some set of

d vertices has only few common neighbors, it is unlikely all the members of R will be chosen among these neighbors. Hence, we do not expect U to contain any such subset of d vertices.

The main idea of this approach is that in the course of a probabilistic proof, it is often better not to make the choices uniformly at random, but to try and make them depend on each other in a way tailored to the specific argument needed. While this sounds somewhat vague, this simple reasoning and its various extensions have already found many applications to extremal graph theory, additive combinatorics, Ramsey theory and combinatorial geometry. For more information about this technique and its applications we refer the interested reader to the recent survey [53].

4. Turán Numbers

Extremal problems are at the heart of graph theory. These problems were extensively studied during the last half century. One of the central questions from which extremal graph theory originated can be described as follows. Given a *forbidden graph* H , determine $\text{ex}(n, H)$, the maximal number of edges in a graph on n vertices that does not contain a copy of H . This number is also called the *Turán number of H* . Instances of this problem appear naturally in discrete geometry, additive number theory, probability, analysis, computer science and coding theory. In this section we describe classical results in this area, mention several applications and report on some recent progress on the problem of determining $\text{ex}(n, H)$ for bipartite graphs.

4.1. Classical results. How dense can a graph G on n vertices be if it contains no triangles? One way to obtain such a graph is to split the vertices into two nearly equal parts A and B and to connect every vertex in A with every vertex in B by an edge. This graph clearly has no triangles and is also very dense. Moreover, it is maximal triangle-free graph, since adding any other edge to G creates a triangle. But is it the densest triangle-free graph on n vertices? More than a hundred years ago Mantel [78] proved that this is indeed the case and therefore $\text{ex}(n, K_3) = \lfloor n^2/4 \rfloor$. This, earliest extremal result in graph theory already has an interesting application, found by Katona [65].

Consider v_1, \dots, v_n , vectors in \mathbb{R}^d of length $|v_i| \geq 1$. How many pairs of these vectors have sum of length less than 1? Suppose we have v_i, v_j, v_k such that all three pairwise sums have length less than 1. Then an easy computation shows that

$$|v_i + v_j + v_k|^2 = |v_i + v_j|^2 + |v_i + v_k|^2 + |v_j + v_k|^2 - |v_i|^2 - |v_j|^2 - |v_k|^2 < 0.$$

This contradiction together with Mantel's theorem shows that the number of pairs i, j with $|v_i + v_j| < 1$ is at most $\lfloor n^2/4 \rfloor$. Suppose now we have two independent identical copies X and Y of some arbitrary distribution with values

in \mathbb{R}^d . By sampling many independent copies of this distribution and using the above claim on the vectors in \mathbb{R}^d one can prove the following general inequality

$$\Pr[|X + Y| \geq 1] \geq \frac{1}{2}(\Pr[|X| \geq 1])^2.$$

The starting point of extremal graph theory is generally considered to be the following celebrated theorem of Turán [99]. Partition n vertices into r parts V_1, \dots, V_r of nearly equal size, i.e., $||V_i| - |V_j|| \leq 1$. Let the *Turán graph* $T_{n,r}$ be the complete r -partite graph obtained by putting the edges between all the pairs of vertices in different parts. In 1941 Turán proved that the largest n -vertex graph, not containing a clique K_{r+1} is precisely $T_{n,r}$. In addition, he posed the problem of determining $\text{ex}(n, H)$ for general graphs and also for hypergraphs (see Section 6).

A priori one might think that the answer to Turán's problem would be messy and that to deal with every particular graph might require each time a new approach. The important and deep theorem of Erdős and Stone [47] together with an observation of Erdős and Simonovits [46] shows that this is not the case. Their very surprising result says that for most graphs there is a single parameter, the chromatic number, which determines the asymptotic behavior of $\text{ex}(n, H)$. The *chromatic number* of a graph H is the minimal number of colors needed to color the vertices of H such that adjacent vertices get different colors. Erdős, Stone and Simonovits proved that for a fixed H and large n

$$\text{ex}(n, H) = \left(1 - \frac{1}{\chi(H) - 1}\right) \frac{n^2}{2} + o(n^2).$$

A moment's thought shows that this determines the asymptotics of $\text{ex}(n, H)$ for all graphs H with chromatic number at least 3. For example, if H is a graph formed by the edges of the icosahedron, then it is easy to check that the chromatic number of H is 4 and therefore $\text{ex}(n, H) = (1 + o(1))n^2/3$.

4.2. Bipartite graphs. As we already mentioned, the theorem of Erdős, Stone and Simonovits determines asymptotically $\text{ex}(n, H)$ for all graphs with chromatic number at least 3. However, for bipartite graphs it only gives $\text{ex}(n, H) = o(n^2)$. The determination of Turán numbers for bipartite graphs remains a challenging project with many open problems. In fact, even the order of magnitude of $\text{ex}(n, H)$ is not known for quite simple bipartite graphs, such as the complete bipartite graph with four vertices in each part, the cycle of length eight, and the 3-cube graph. Here we describe some partial results obtained so far, which use a variety of techniques from different fields including probability, number theory and algebraic geometry.

Let $t \leq s$ be positive integers and let $K_{t,s}$ denote the complete bipartite graph with parts of size t and s . For every fixed t and $s \geq t$, Kővári, Sós and Turán [75] proved, more than 60 years ago, that

$$\text{ex}(n, K_{t,s}) \leq \frac{1}{2}(s-1)^{1/t}n^{2-1/t} + \frac{1}{2}(t-1)n.$$

It is conjectured that the right hand side gives the correct order of magnitude of $\text{ex}(n, K_{t,s})$. However, progress on this problem was slow and despite several results by various researchers this is known only for $s > (t-1)!$ (see [4] and its references). In particular, in the most interesting case $s = t$ the Turán number of $K_{4,4}$ is already unknown. All constructions for this problem are algebraic and the more recent ones require some tools from elementary algebraic geometry.

The Turán numbers for $K_{t,s}$ appear naturally in problems in other areas of mathematics. For example, in 1946, Erdős [31] asked to determine the maximum possible number of unit distances among n points on the plane. One might think that potentially the number of such distances may be even quadratic. Given such a set of n points, consider a graph whose vertices are the points and two of them are adjacent if the distance between them is one. Since on the plane for any two fixed points p and p' there are precisely two other points whose distance to both p, p' is one, the resulting graph has no $K_{2,3}$. Therefore, by the above result there are at most $O(n^{3/2})$ unit distances. Erdős conjectured that the number of such distances is always at most $n^{1+o(1)}$, but the best current bound for this problem, obtained in [93], is $O(n^{4/3})$.

Suppose we have a set of integers A such that $A + A = \{a + a' \mid a, a' \in A\}$ contains all numbers $1^2, 2^2, \dots, n^2$. How small can the set A be? This is a special case of the question asked by Wooley [100] at the AIM conference on additive combinatorics in 2004. Clearly, A has size at least \sqrt{n} but the truth is probably $n^{1-o(1)}$. It appears that Erdős and Newman [44] already considered this problem earlier and noticed that using extremal graph theory one can show that $|A| \geq n^{2/3-o(1)}$. Consider a graph whose vertices are elements of A and for every $1 \leq x \leq n$ choose some pair a, a' such that $x^2 = a + a'$ and connect them by an edge. Erdős and Newman use bounds on $\text{ex}(n, K_{2,s})$ to conclude that if $|A| = n^{2/3-\epsilon}$ then this graph must contain two vertices a_1 and a_2 with at least n^δ common neighbors. Thus one can show that $a_1 - a_2$ can be written as a difference of two squares in n^δ different ways and therefore will have too many divisors, a contradiction.

Not much is known for Turán numbers of general bipartite graphs. Moreover, we do not even have a good guess what parameter of a bipartite graph might determine the order of growth of its Turán number. Some partial answers to this question were proposed by Erdős. Recall that a graph is t -degenerate if its every subgraph contains a vertex of degree at most t . In 1966 Erdős [33] (see also [20]) conjectured that every t -degenerate bipartite graph H satisfies $\text{ex}(n, H) \leq O(n^{2-1/t})$. Recently, progress on this conjecture was obtained in [3]. One of the results in this paper says that the conjecture holds for every bipartite graph H in which the degrees of all vertices in one part are at most t . This result, which can be also derived from an earlier result of Füredi [55], is a far reaching generalization of the above estimate of Kővári, Sós and Turán. It is tight for every fixed t as was shown, e.g., by constructions in [4]. Another result in [3] gives the first known estimate on the Turán numbers of degenerate bipartite graphs.

Theorem 4.1. *Let H be a bipartite t -degenerate graph on h vertices. Then for all $n \geq h$*

$$ex(n, H) \leq h^{1/2t} n^{2 - \frac{1}{4t}}.$$

The proof of this theorem and also of the first result from [3] mentioned above is based on the dependent random choice approach, which we briefly discussed in Section 3.4.

4.3. Subgraph multiplicity. Turán's theorem says that any graph with $m > (1 - \frac{1}{r}) \frac{n^2}{2}$ edges contains at least one copy of K_{r+1} . The question of how many such copies $f_r(m, n)$ must exist in an n -vertex graph with m edges received quite a lot of attention and has turned out to be notoriously difficult. When m is very close to the $ex(n, K_{r+1})$ this function was computed by Erdős. Let $m = p \binom{n}{2}$, where the edge density p (the fraction of the pairs which are edges) is a fixed constant strictly greater than $1 - 1/r$. One very interesting open question is to determine the asymptotic behavior of $f_r(m, n)$ as a function of p only. Further results in this direction were obtained by Goodman, Lovász, Simonovits, Bollobás, and Fisher (for more details see [82, 80] and their references). Recently Razborov [82] and Nikiforov [80] resolved this problem for the cases $r = 2$ and $r = 3$, respectively. It appears that in these cases the solution corresponds to the complete $(t + 1)$ -partite graph in which t parts are roughly equal and are larger than the remaining part, and the integer t is such that $p \in [1 - \frac{1}{t}, 1 - \frac{1}{t+1}]$.

For bipartite graphs the situation seems to be very different. The beautiful conjectures of Erdős and Simonovits [90] and of Sidorenko [89] suggest that for any bipartite H there is $\gamma(H) > 0$ such that the number of copies of H in any graph G on n vertices and edge density $p > n^{-\gamma(H)}$ is asymptotically at least the same as in the n -vertex random graph with edge density p . The original formulation of the conjecture by Sidorenko is in terms of graph homomorphisms. A homomorphism from a graph H to a graph G is a mapping $f : V(H) \rightarrow V(G)$ such that for each edge (u, v) of H , $(f(u), f(v))$ is an edge of G . Let $h_H(G)$ denote the number of homomorphisms from H to G . We also consider the normalized function $t_H(G) = h_H(G)/|G|^{|H|}$, which is the fraction of mappings $f : V(H) \rightarrow V(G)$ which are homomorphisms. Sidorenko's conjecture states that for every bipartite graph H with q edges and every graph G ,

$$t_H(G) \geq t_{K_2}(G)^q.$$

This conjecture also has the following appealing analytical form. Let μ be the Lebesgue measure on $[0, 1]$ and let $h(x, y)$ be a bounded, symmetric, non-negative and measurable function on $[0, 1]^2$. Let H be a bipartite graph with vertices u_1, \dots, u_t in the first part and vertices v_1, \dots, v_s in the second part. Denote by E the set of edges of H , i.e., all the pairs (i, j) such that u_i and v_j are adjacent, and let $|E| = q$.

Conjecture 4.2.

$$\int \prod_{(i,j) \in E} h(x_i, y_j) d\mu^{s+t} \geq \left(\int h d\mu^2 \right)^q.$$

The expression on the left hand side of this inequality is quite common. Such integrals are called Feynman integrals in quantum field theory and they also appear in classical statistical mechanics. Unsurprisingly then, Sidorenko's conjecture has connections to a broad range of topics, such as matrix theory, Markov chains, graph limits and quasirandomness. So far this conjecture was established only in very special cases, e.g., for complete bipartite graphs, trees, even cycles (see [89]), and also for cubes [64].

Recently, Sidorenko's conjecture was proved for a new class of graphs. In [28], it was shown that the conjecture holds for every bipartite graph H which has a vertex adjacent to all the vertices in the other part. Using this result, one can easily deduce an approximate version of Sidorenko's conjecture for all graphs. For a connected bipartite graph H with parts V_1, V_2 , define the bipartite graph \bar{H} with parts V_1, V_2 such that $(v_1, v_2) \in V_1 \times V_2$ is an edge of \bar{H} if and only if it is not an edge of H . Define the *width* of H to be the minimum degree of \bar{H} . If H is not connected, the width of H is the sum of the widths of the connected components of H . Note that the width of a connected bipartite graph is 0 if and only if it has a vertex that is complete to the other part. Moreover, the width of a bipartite graph with h vertices is always at most $h/2$.

Theorem 4.3. *If H is a bipartite graph with q edges and width w , then $t_H(G) \geq t_{K_2}(G)^{q+w}$ holds for every graph G .*

5. Generalizations

Turán's theorem, which determines the maximum number of edges in a K_{r+1} -free graph on n vertices, is probably the most famous result in extremal combinatorics and there are many interesting generalizations and extensions of this theorem. In this section we discuss several such results.

5.1. Local density. A generalization of Turán's theorem that takes into account edge distribution, or local density, was introduced by Erdős [34] in 1975. He asked the following question. Suppose that G is a K_{r+1} -free graph on n vertices in which every set of αn vertices spans at least βn^2 edges for some $0 \leq \alpha, \beta \leq 1$. How large can β be as a function of α ? Erdős, Faudree, Rousseau and Schelp [39] studied this problem and conjectured that for α sufficiently close to 1 the Turán graph $T_{n,r}$ has the highest local density. They proved this for triangle-free graphs ($r = 2$) and the general case of this conjecture was established in [66]. It is easy to check that for $\alpha \geq \frac{r-1}{r}$ every subset of $T_{n,r}$ of size αn contains at least $\frac{r-1}{2r}(2\alpha - 1)n^2$ edges. The result in [66] says that if G

is a K_{r+1} -free graph on n vertices and $1 - \frac{1}{2r^2} \leq \alpha \leq 1$, then G contains a set of αn vertices spanning at most $\frac{r-1}{2r}(2\alpha - 1)n^2$ edges and equality holds only when G is a Turán graph.

For triangle-free graphs and general α it was conjectured in [39] that β is determined by a family of extremal triangle-free graphs. Besides the complete bipartite graph $T_{n,2}$ already mentioned, another important graph is $C_{n,5}$, which is obtained from a 5-cycle by replacing each vertex i by an independent set V_i of size $n/5$ (assuming for simplicity that n is divisible by 5), and each edge ij by a complete bipartite graph joining V_i and V_j (this operation is called a ‘blow-up’). Erdős et al. conjectured that for α above $17/30$ the Turán graph has the highest local density and for $1/2 \leq \alpha \leq 17/30$ the best graph is $C_{n,5}$. On the other hand, for $r \geq 3$ Chung and Graham [19] conjectured that the Turán graph has the best local density even for α as low as $1/2$. When α is a small constant the situation is unclear and there are no natural conjectures.

The case $r = 2$ and $\alpha = 1/2$ is one of the favorite questions of Erdős that he returned to often and offered a \$250 prize for its solution. Here the conjecture is that any triangle-free graph on n vertices should contain a set of $n/2$ vertices that spans at most $n^2/50$ edges. This conjecture is one of several important questions in extremal graph theory where the optimal graph is suspected to be the blow-up of the 5-cycle $C_{n,5}$. So far these problems are completely open and we need new techniques to handle them.

Another question, that is similar in spirit, is to determine how many edges one may need to delete from a K_{r+1} -free graph on n vertices in order to make it bipartite. This is an instance of the well known Max-Cut problem, which asks for the largest bipartite subgraph of a given graph G . This problem has been the subject of extensive research both from the algorithmic perspective in computer science and the extremal perspective in combinatorics.

A long-standing conjecture of Erdős [34] says that one needs to delete at most $n^2/25$ edges from a triangle-free graph to make it bipartite, and $C_{n,5}$ shows that this estimate would be the best possible. This problem is still open and the best known bound is $(1/18 - \epsilon)n^2$ for some constant $\epsilon > 0$, obtained by Erdős, Faudree, Pach and Spencer [38]. Erdős also conjectured that for K_4 -free graphs on n vertices the answer for this problem is at most $n^2/9$. This was recently proved in [96].

Theorem 5.1. *Every K_4 -free graph on n vertices can be made bipartite by deleting at most $n^2/9$ edges. Moreover, the only extremal graph which requires deletion of so many edges is the Turán graph $T_{n,3}$.*

It is also plausible to conjecture that, for all $r > 3$, the K_{r+1} -free n -vertex graph that requires the most edge deletions in order to make it bipartite is the Turán graph $T_{n,r}$.

It was observed in [76] that for regular graphs, a bound for the local density problem implies a related bound for the problem of making the graph bipartite. Indeed, suppose n is even, G is a d -regular graph on n vertices and S is a

set of $n/2$ vertices. Then $dn/2 = \sum_{s \in S} d(s) = 2e(S) + e(S, \bar{S})$ and $dn/2 = \sum_{s \notin S} d(s) = 2e(\bar{S}) + e(S, \bar{S})$. This implies that $e(S) = e(\bar{S})$, i.e., S and \bar{S} span the same number of edges. Deleting the $2e(S)$ edges within S and \bar{S} makes the graph bipartite. Thus, for example, if in a regular triangle-free graph G one can find a set S with $|S| = n/2$ which spans at most $n^2/50$ edges, then G can be made bipartite by deleting at most $n^2/25$ edges. This relation, together with Theorem 5.1, gives some evidence that indeed for $r \geq 3$ the Turán graph should have the best local density for all $0 \leq \alpha \leq 1$.

5.2. Graphs with large minimum degree. Clearly, for any graph G , the largest K_{r+1} -free subgraph of G has at least as many edges as does the largest r -partite subgraph. For which graphs do we have an equality? This question was raised by Erdős [35], who noted that by Turán's theorem there is an equality for the complete graph K_n . In [7] and [16] it was shown that the equality holds with high probability for sufficiently dense random graphs. Recently, a general criteria implying equality was obtained in [5], where it is proved that a minimum degree condition is sufficient. Given a fixed graph H and a graph G let $e_r(G)$ and $e_H(G)$ denote the number of edges in the largest r -partite and the largest H -free subgraphs of G , respectively. The following theorem shows that both Turán's and Erdős-Stone-Simonovits' theorems holds not only for K_n but also for any graph with large minimum degree.

Theorem 5.2. *Let H be a graph with chromatic number $r+1 \geq 3$. Then there are constants $\gamma = \gamma(H) > 0$ and $\mu = \mu(H) > 0$ such that if G is a graph on n vertices with minimum degree at least $(1 - \mu)n$, then*

$$e_r(G) \leq e_H(G) \leq e_r(G) + O(n^{2-\gamma}).$$

Moreover, if $H = K_{r+1}$ then $e_H(G) = e_r(G)$.

The assertion of this theorem for the special case when H is a triangle is proved in [14] and in a stronger form in [8].

As well as being interesting in its own right, this theorem was motivated by the following question in computer science. Given some *property* \mathcal{P} and a graph G , it is a fundamental computational problem to find the smallest number of edge deletions and additions needed to turn G into a graph satisfying this property. We denote this quantity by $E_{\mathcal{P}}(G)$. Specific instances of graph modification problems arise naturally in several fields, including molecular biology and numerical algebra. A graph property is *monotone* if it is closed under removal of vertices and edges. Note that, when trying to turn a graph into one satisfying a monotone property, we will only need to use edge deletions and therefore in these cases the problem is called an edge-deletion problem. Two examples of interesting monotone properties are k -colorability and the property of not containing a copy of a fixed graph H . It appears that using combinatorial methods it is possible to give a nearly complete answer to the question of

how accurately one can approximate (up to additive error) the solution of the edge-deletion problem for monotone properties.

For any fixed $\epsilon > 0$ and any monotone property \mathcal{P} there is a deterministic algorithm, obtained in [5], which, given a graph G on n vertices, approximates $E_{\mathcal{P}}(G)$ within an additive error of ϵn^2 (i.e., it computes a number X such that $|X - E_{\mathcal{P}}(G)| \leq \epsilon n^2$). Moreover, the running time of the algorithm is linear in the size of the graph. This algorithm uses a strengthening of Szemerédi's regularity lemma which implies that every graph G can be approximated by a small (fixed size) weighted graph W , so that $E_{\mathcal{P}}(G)$ is an approximate solution of a related problem on W . Since W has a fixed size, we can now resort to a brute force solution. Given the above, a natural question is for which monotone properties one can obtain better additive approximations of $E_{\mathcal{P}}$. Another result in [5] essentially resolves this problem by giving a precise characterization of the monotone graph properties for which such approximations exist.

On the one hand, if there is a bipartite graph that does not satisfy property \mathcal{P} , then there is a $\delta > 0$ for which it is possible to approximate $E_{\mathcal{P}}$ within an additive error of $n^{2-\delta}$ in polynomial time. On the other hand, if all bipartite graphs satisfy \mathcal{P} , then for any $\delta > 0$ it is NP -hard to approximate distance to \mathcal{P} within an additive error of $n^{2-\delta}$. The proof of this result, among the other tools, uses Theorem 5.2 together with spectral techniques. Interestingly, prior to [5], it was not even known that computing $E_{\mathcal{P}}$ *precisely* for most properties satisfied by all bipartite graphs (e.g., being triangle-free) is NP -hard. It thus answers (in a strong form) a question of Yannakakis, who asked in 1981 if it is possible to find a large and natural family of graph properties for which computing $E_{\mathcal{P}}$ is NP -hard.

5.3. Spectral Turán theorem. Given an arbitrary graph G , consider a partition of its vertices into r parts which maximizes the number of edges between the parts. Then the degree of each vertex within its own part is at most $1/r$ -times its degree in G , since otherwise we can move this vertex to some other part and increase the total number of edges connecting different parts. This simple construction shows that the largest r -partite and hence also largest K_{r+1} -free subgraph of G has at least a $\frac{r-1}{r}$ -fraction of its edges. We say that a graph G (or rather a family of graphs) is r -Turán if this trivial lower bound is essentially an upper bound as well, i.e., the largest K_{r+1} -free subgraph of G has at most $(1 + o(1))\frac{r-1}{r}|E(G)|$ edges. Note that Turán's theorem says that this holds when G is a complete graph on n vertices. Thus it is very natural to ask, which other graphs are r -Turán?

It has been shown that for any fixed r , there exists $p(r, n)$ such that for all $p \gg p(r, n)$ with high probability the random graph $G_{n,p}$ is r -Turán. The value of p for which this result holds was improved several times by various researches. Recently, resolving the longstanding conjecture, Conlon and Gowers [29] and independently Schacht [87] established the optimal value of $p(r, n) = n^{-2/(r+2)}$. These results about random graphs do not yet provide a deterministic sufficient

condition for a graph to be r -Turán. However, they suggest that one should look at graphs whose edges are distributed sufficiently evenly. It turns out that under certain circumstances, such an edge distribution can be guaranteed by a simple assumption about the spectrum of the graph.

For a graph G , let $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ be the eigenvalues of its adjacency matrix. The quantity $\lambda(G) = \max\{\lambda_2, -\lambda_n\}$ is called the *second eigenvalue* of G . A graph $G = (V, E)$ is called an (n, d, λ) -graph if it is d -regular, has n vertices and the second eigenvalue of G is at most λ . It is well known (see [6, 77] for more details) that if λ is much smaller than the degree d , then G has certain random-like properties. Thus, λ could serve as some kind of “measure of randomness” in G . The following recent result from [97] shows that Turán’s theorem holds asymptotically for graphs with small second eigenvalue.

Theorem 5.3. *Let $r \geq 2$ be an integer and let $G = (V, E)$ be an (n, d, λ) -graph. If $d^r/n^{r-1} \gg \lambda$ then the largest K_{r+1} -free subgraph of G has at most $(1 + o(1)) \frac{r-1}{r} |E(G)|$ edges*

This result generalizes Turán’s theorem, since the second eigenvalue of the complete graph K_n is 1 and thus it satisfies the above condition. Theorem 5.3 is also part of the fast-growing comprehensive study of graph theoretical properties of (n, d, λ) -graphs, which has recently attracted lots of attention both in combinatorics and theoretical computer science. For a recent survey about these fascinating graphs and their properties, we refer the interested reader to [77].

6. Turán-type Problems for Hypergraphs

Given a k -uniform hypergraph H , the Turán number $\text{ex}(n, H)$ is the maximum number of edges in a k -uniform hypergraph on n vertices that does not contain a copy of H . Determining these numbers is one of the main challenges in extremal combinatorics. For ordinary graphs (the case $k = 2$), a rich theory has been developed, whose highlights we described in Section 4. In 1941, Turán also posed the question of finding $\text{ex}(n, K_s^{(k)})$ for complete k -uniform hypergraphs with $s > k > 2$ vertices, but to this day not one single instance of this problem has been solved. It seems very hard even to determine the *Turán density*, which is defined as $\pi(H) = \lim_{n \rightarrow \infty} \text{ex}(n, H) / \binom{n}{k}$.

The most famous problem in this area is the conjecture of Turán that $\text{ex}(n, K_4^{(3)})$ is given by the following construction, which we denote by T_n . Partition n vertices into 3 sets V_0, V_1, V_2 of equal size. Consider all triples which either intersect all these sets or contain two vertices in V_i and one in $V_{i+1 \pmod{3}}$. This hypergraph has density $5/9$ and every 4 vertices span at most 3 edges. In memory of Turán, Erdős offered \$1000 for proving that $\pi(K_4^{(3)}) = 5/9$. Despite several results giving rather close estimates for the Turán density of $K_4^{(3)}$, this problem remains open. One of the main difficulties is that Turán’s conjecture, if

it is true, has exponentially many non-isomorphic extremal configurations (see [71]).

Recently the problem of finding the numbers $\text{ex}(n, H)$ got a lot of attention and these numbers were determined for various hypergraphs. One such example is the Fano plane $PG_2(2)$, which is the projective plane over the field with 2 elements. It is the unique 3-uniform hypergraph with 7 vertices and 7 edges, in which every pair of vertices is contained in a unique triple and triples corresponds to the lines in the projective plane. A hypergraph is *2-colorable* if its vertices can be labeled as red or blue so that no edge is monochromatic. It is easy to check that the Fano plane is not 2-colorable, and therefore any 2-colorable hypergraph cannot contain the Fano plane. Partition an n -element set into two almost equal parts, and take all the triples that intersect both of them. This is clearly the largest 2-colorable 3-uniform hypergraph on n vertices. In 1976 Sós conjectured that this construction gives the exact value of $\text{ex}(n, PG_2(2))$. This was proved independently in [67] and [56], where it was also shown that the extremal construction, which we described above, is unique.

The strategy of the proof is first to obtain an approximate structure theorem, and then to show that any imperfection in the structure leads to a suboptimal configuration. This is the so-called “stability approach” which was first introduced for graphs by Simonovits. Following the above two papers, this approach has become a standard tool for attacking extremal problems for hypergraphs as well, and was used successfully to determine several hypergraph Turán numbers.

Let $\mathcal{C}_r^{(2k)}$ be the $2k$ -uniform hypergraph obtained by letting P_1, \dots, P_r be pairwise disjoint sets of size k and taking as edges all sets $P_i \cup P_j$ with $i \neq j$. This can be thought of as the ‘ k -expansion’ of the complete graph K_r : each vertex has been replaced with a set of size k . The Turán problem for $\mathcal{C}_r^{(2k)}$ was first considered by Frankl and Sidorenko, as a possible generalization of Turán’s theorem for graphs. Using a clever reduction of this problem to the case of graphs they showed that the Turán density of $\mathcal{C}_r^{(2k)}$ is at most $\frac{r-2}{r-1}$. Frankl and Sidorenko also gave a matching lower bound construction, which was essentially algebraic but existed only when $r = 2^a + 1$. In [68], among other results, it was shown that, surprisingly, when r is not of the form $2^a + 1$ then the Turán density of $\mathcal{C}_r^{(4)}$ is strictly smaller than $\frac{r-2}{r-1}$. Interestingly, this result, showing that certain constructions do not exist, also uses a stability argument. By studying the properties of a $\mathcal{C}_r^{(4)}$ -free hypergraph with density close to $\frac{r-2}{r-1}$ the authors show that it gives rise to an edge coloring of the complete graph K_{r-1} with special properties. Next they show that for such an edge-coloring there is a natural $GF(2)$ vector space structure on the colors. Of course, such a space has cardinality 2^a , for some integer a , so one gets a contradiction unless $r = 2^a + 1$.

It is interesting to note that T_n , the conjectured extremal example for $K_4^{(3)}$, also does not contain 4 vertices which span a single edge. Thus, there is a 3-uniform hypergraph with edge density $5/9$ in which every 4 vertices span either

zero or two edges. In [83], Razborov showed that $5/9$ is the maximum possible density for such a hypergraph. Combining his result with the stability approach (described above) Pikhurko proved that the unique extremal configuration for this problem is T_n . Razborov's proof uses the formalism of flag algebras, which, roughly speaking, allows one to computerize the search for inequalities which should be satisfied by various statistics of the extremal hypergraph. Then the "right inequalities" can be proved using Cauchy-Schwarz type arguments. This approach works for various other extremal problems as well. For example, one can use it to improve the best known bounds for Turán's original conjecture (see [83]).

6.1. Hypergraphs and arithmetic progressions. Extremal problems for hypergraphs have many connections to other areas of mathematics. Here we describe one striking application of hypergraphs to number theory.

An old question of Brown, Erdős and Sós asks to determine the maximum number of edges in the k -uniform hypergraph which has no s edges whose union has at most t vertices. This is a very difficult question which is solved only for few specific values of parameters. One such special case is the so-called $(6, 3)$ -problem. Here, one wants to maximize the number of edges in a 3-uniform hypergraph such that every 6 vertices span at most 2 edges. In 1976, Ruzsa and Szemerédi [86] proved that such a hypergraph can have only $o(n^2)$ edges. Surprisingly, this purely combinatorial result has a tight connection with number theory. Using it one can give a short proof of the well-known theorem of Roth that every $A \subset [n]$ of size ϵn (for constant ϵ and large n) contains a 3-term arithmetic progression. Indeed, consider a 3-uniform hypergraph whose vertex set is the disjoint union of $[n]$, $[2n]$ and $[3n]$ and whose edges are all the triples $x, x + a, x + 2a$ with $x \in [n]$ and $a \in A$. This hypergraph has $O(n)$ vertices, $n|A|$ edges and, one can check that every 3-term arithmetic progression in A corresponds to 6 vertices spanning at least 3 edges and vice versa.

The $(6, 3)$ -theorem of Ruzsa and Szemerédi is closely related to the triangle removal lemma, which says that for every ϵ there is a δ such that every graph on n vertices with at most δn^3 triangles can be made triangle-free by removing ϵn^2 edges. The original proof of both results used the regularity lemma and therefore gave a very poor dependence of δ on ϵ . Very recently, this result was substantially improved by Fox [49]. Still, the dependence of δ on ϵ in [49] is of tower-type and compared with the Fourier-analytical approach it gives much weaker bounds for the number-theoretic applications.

A remarkable extension of the triangle removal lemma to hypergraphs was obtained by Gowers [59] and independently by Nagle, Rödl, Schacht and Skokan [84, 79]. They proved that if a k -uniform hypergraph on n vertices has at most δn^{k+1} copies of the complete hypergraph $K_{k+1}^{(k)}$, then all these copies can be destroyed by removing ϵn^k edges. This result was obtained by developing a new, very useful and important tool: the hypergraph analogue of the regularity

lemma. The hypergraph removal lemma can be used to give a short proof of Szemerédi's theorem that dense subsets of integers contain long arithmetic progressions (see [91]).

7. Conclusion

We mentioned several specific problems of extremal combinatorics throughout this paper. Many of them are of a fundamental nature, and we believe that any progress on these questions will require the development of new techniques which will have wide applicability. We also gave examples of connections between extremal combinatorics and other areas of mathematics. In the future it is safe to predict that the number of such examples will only grow. Combinatorics will employ more and more advanced tools from algebra, topology, analysis and geometry and, on the other hand, there will be more applications of purely combinatorial techniques to non-combinatorial problems. One spectacular instance of such an interaction is a series of recent results on approximate subgroups and expansion properties of linear groups which combine combinatorics, number theory, algebra and model theory (see, e.g., [15] and its references).

The open problems which we mentioned, as well as many more additional ones which we skipped due to the lack of space, will provide interesting challenges for future research in extremal combinatorics. These challenges, the fundamental nature of the area and its tight connection with other mathematical disciplines will ensure that in the future extremal combinatorics will continue to play an essential role in the development of mathematics.

Acknowledgments

I would like to thank Conlon, Fox, Keevash, Krivelevich and Markman for carefully reading this manuscript and many valuable comments, which greatly improved the presentation.

References

- [1] M. Ajtai, J. Komlós, and E. Szemerédi, A note on Ramsey numbers, *J. Combinatorial Theory, Ser. A* **29** (1980), 354–360.
- [2] N. Alon, Subdivided graphs have linear Ramsey numbers, *J. Graph Theory* **18** (1994), 343–347.
- [3] N. Alon, M. Krivelevich, B. Sudakov, Turán numbers of bipartite graphs and related Ramsey-type questions, *Combin. Probab. Comput.* **12** (2003), 477–494.
- [4] N. Alon, L. Rónyai and T. Szabó, Norm-graphs: variations and applications, *J. Combinatorial Theory, Ser. B* **76** (1999), 280–290.

-
- [5] N. Alon, A. Shapira and B. Sudakov, Additive approximation for edge-deletion problems, *Annals of Mathematics* **170** (2009), 371–411.
- [6] N. Alon and J. Spencer, **The probabilistic method**, 3rd Ed., Wiley, New York, 2008.
- [7] L. Babai, M. Simonovits and J. Spencer, Extremal subgraphs of random graphs, *J. Graph Theory* **14** (1990), 599–622.
- [8] J. Balogh, P. Keevash and B. Sudakov, On the minimal degree implying equality of the largest triangle-free and bipartite subgraphs, *J. Combinatorial Theory Ser. B* **96** (2006), 919–932.
- [9] B. Barak, G. Kindler, R. Shaltiel, B. Sudakov and A. Wigderson, Simulating Independence: New Constructions of Condensers, Ramsey Graphs, Dispersers and Extractors, *Proc. of the 37th Symposium on Theory of Computing (STOC)*, ACM (2005), 1–10.
- [10] B. Barak, A. Rao, R. Shaltiel and A. Wigderson, 2-source dispersers for sub-polynomial entropy and Ramsey graphs beating the Frankl-Wilson construction, *Proc. 38th Symposium on Theory of Computing (STOC)*, ACM (2006), 671–680.
- [11] J. Beck, An upper bound for diagonal Ramsey numbers, *Studia Sci. Math. Hungar* **18** (1983), 401–406.
- [12] T. Bohman, The Triangle-Free Process, *Advances in Mathematics* **221** (2009), 1653–1677.
- [13] T. Bohman and P. Keevash, The early evolution of the H-free process, preprint.
- [14] J. Bondy, J. Shen, S. Thomassé and C. Thomassen, Density conditions implying triangles in k-partite graphs, *Combinatorica* **26** (2006), 121–131.
- [15] E. Breuillard, B. Green and T. Tao, Linear Approximate Groups, preprint.
- [16] G. Brightwell, K. Panagiotou and A. Steger, On extremal subgraphs of random graphs, *Proc. of the 18th Symposium on Discrete Algorithms (SODA)*, ACM-SIAM (2007), 477–485.
- [17] S. A. Burr and P. Erdős, On the magnitude of generalized Ramsey numbers for graphs, in: *Infinite and Finite Sets I*, 10, Colloq. Math. Soc. Janos Bolyai, North-Holland, Amsterdam, 1975, 214–240.
- [18] G. Chen and R. H. Schelp, Graphs with linearly bounded Ramsey numbers, *J. Combin. Theory Ser. B* **57** (1993), 138–149.
- [19] F. Chung and R. Graham, On graphs not containing prescribed induced subgraphs, in: *A tribute to Paul Erdős*, Cambridge Univ. Press, Cambridge, 1990, 111–120.
- [20] F. Chung and R. Graham, **Erdős on Graphs. His Legacy of Unsolved Problems**, A K Peters, Ltd., Wellesley, MA, 1998.
- [21] V. Chvátal, V. Rödl, E. Szemerédi, and W. T. Trotter, Jr., The Ramsey number of a graph with bounded maximum degree, *J. Combin. Theory Ser. B* **34** (1983), 239–243.
- [22] D. Conlon, A new upper bound for diagonal Ramsey numbers, *Annals of Mathematics* **170** (2009), 941–960.

- [23] D. Conlon, Hypergraph packing and sparse bipartite Ramsey numbers, *Combin. Probab. Comput.*, **18** (2009), 913–923.
- [24] D. Conlon, J. Fox and B. Sudakov, Ramsey numbers of sparse hypergraphs, *Random Structures Algorithms* **35** (2009), 1–14.
- [25] D. Conlon, J. Fox and B. Sudakov, Hypergraph Ramsey numbers, *J. Amer. Math. Soc.* **23** (2010), 247–266.
- [26] D. Conlon, J. Fox and B. Sudakov, Large almost monochromatic subsets in hypergraphs, *Israel Journal of Mathematics*, to appear.
- [27] D. Conlon, J. Fox and B. Sudakov, On two problems in graph Ramsey theory, submitted.
- [28] D. Conlon, J. Fox and B. Sudakov, An approximate version of Sidorenko’s conjecture, preprint.
- [29] D. Conlon and T. Gowers, Combinatorial theorems relative to a random set, preprint.
- [30] O. Cooley, N. Fountoulakis, D. Kühn and D. Osthus, Embeddings and Ramsey numbers of sparse k -uniform hypergraphs, *Combinatorica* **28** (2009), 263–297.
- [31] P. Erdős, On sets of distances of n points. *Amer. Math. Monthly* **53** (1946), 248–250.
- [32] P. Erdős, Some remarks on the theory of graphs, *Bull. Amer. Math. Soc.* **53** (1947), 292–294.
- [33] P. Erdős, Some recent results on extremal problems in graph theory, in: *em Theory of Graphs* (Rome, 1966), Gordon and Breach, New York, (1967), 117–123.
- [34] P. Erdős, Problems and results in graph theory and combinatorial analysis. Proceedings of the Fifth British Combinatorial Conference (Univ. Aberdeen, Aberdeen, 1975), pp. 169–192. *Congressus Numerantium*, No. XV, Utilitas Math., Winnipeg, Man., 1976.
- [35] P. Erdős, On some problems in graph theory, combinatorial analysis and combinatorial number theory, in: *Graph theory and combinatorics (Cambridge, 1983)*, Academic Press, London, 1984, 1–17.
- [36] P. Erdős, Problems and results on graphs and hypergraphs: similarities and differences, in *Mathematics of Ramsey theory*, Algorithms Combin., Vol. 5 (J. Nešetřil and V. Rödl, eds.) 12–28. Berlin: Springer-Verlag, 1990.
- [37] P. Erdős, Problems and results in discrete mathematics, *Discrete Math.* **136** (1994), 53–73.
- [38] P. Erdős, R. Faudree, J. Pach and J. Spencer, How to make a graph bipartite, *J. Combin. Theory Ser. B* **45** (1988), 86–98.
- [39] P. Erdős, R. Faudree, C. Rousseau and R. Schelp, A local density condition for triangles, *Discrete Math.* **127** (1994), 153–161.
- [40] P. Erdős and R. Graham, On partition theorems for finite graphs, in *Infinite and finite sets (Colloq., Keszthely, 1973)*, Vol. I; Colloq. Math. Soc. János Bolyai, Vol. 10, North-Holland, Amsterdam, 1975, 515–527.

- [41] P. Erdős and A. Hajnal, On Ramsey like theorems, Problems and results, Combinatorics (Proc. Conf. Combinatorial Math., Math. Inst., Oxford, 1972) , pp. 123–140, Inst. Math. Appl., Southend-on-Sea, 1972.
- [42] P. Erdős and A. Hajnal, Ramsey-type theorems, *Discrete Appl. Math.* **25** (1989), 37–52.
- [43] P. Erdős, A. Hajnal, and R. Rado, Partition relations for cardinal numbers, *Acta Math. Acad. Sci. Hungar.* **16** (1965), 93–196.
- [44] P. Erdős and D.J. Newman, Bases for sets of integers, *J. Number Theory* **9** (1977), 420–425.
- [45] P. Erdős and R. Rado, Combinatorial theorems on classifications of subsets of a given set, *Proc. London Math. Soc.* **3** (1952), 417–439.
- [46] P. Erdős and M. Simonovits, A limit theorem in graph theory, in *Studia Sci. Math. Hungar* **1** (1966), 51–57.
- [47] P. Erdős and A.H. Stone, On the structure of linear graphs, *Bull. Amer. Math. Soc.* **52** (1946), 1087–1091.
- [48] P. Erdős and G. Szekeres, A combinatorial problem in geometry, *Compositio Math.* **2** (1935), 463–470.
- [49] J. Fox, A new proof of the triangle removal lemma, preprint.
- [50] J. Fox and B. Sudakov, Induced Ramsey-type theorems, *Advances in Mathematics* **219** (2008), 1771–1800.
- [51] J. Fox and B. Sudakov, Density theorems for bipartite graphs and related Ramsey-type results, *Combinatorica* **29** (2009), 153–196.
- [52] J. Fox and B. Sudakov, Two remarks on the Burr-Erdős conjecture, *European J. Combinatorics*, **30** (2009), 1630–1645.
- [53] J. Fox and B. Sudakov, Dependent Random Choice, *Random Structures and Algorithms*, to appear.
- [54] P. Frankl and R. Wilson, Intersection theorems with geometric consequences, *Combinatorica* **1** (1981), 357–368.
- [55] Z. Füredi, On a Turán type problem of Erdős, *Combinatorica* **11** (1991), 75–79.
- [56] Z. Füredi and M. Simonovits, Triple systems not containing a Fano configuration *Combin. Probab. Comput.* **14** (2005), 467–484.
- [57] W.T. Gowers, Lower bounds of tower type for Szemerédi’s uniformity lemma, *Geom. Funct. Anal.* **7** (1997), 322–337.
- [58] W.T. Gowers, A new proof of Szemerédi’s theorem for arithmetic progressions of length four, *Geom. Funct. Anal.* **8** (1998), 529–551.
- [59] W.T. Gowers, Hypergraph regularity and the multidimensional Szemerédi theorem, *Ann. of Math.* **166** (2007), 897–946.
- [60] R. Graham, V. Rödl, and A. Ruciński, On graphs with linear Ramsey numbers, *J. Graph Theory* **35** (2000), 176–192.
- [61] R. Graham, V. Rödl, and A. Ruciński, On bipartite graphs with linear Ramsey numbers, *Combinatorica* **21** (2001), 199–209.

- [62] R. L. Graham, B. L. Rothschild, and J. H. Spencer, **Ramsey theory**, 2nd edition, John Wiley & Sons (1980).
- [63] B. Green and T. Tao, The primes contain arbitrarily long arithmetic progressions, *Annals of Math.* **167** (2008), 481–547.
- [64] H. Hatami, Graph norms and Sidorenko’s conjecture, *Israel Journal of Mathematics*, to appear.
- [65] G. Katona, Graphs, vectors and inequalities in probability theory (in Hungarian), *Mat. Lapok* **20** (1969), 123–127.
- [66] P. Keevash and B. Sudakov, Local density in graphs with forbidden subgraphs, *Combinatorics, Probability and Computing* **12** (2003), 139–153.
- [67] P. Keevash and B. Sudakov, The Turán number of the Fano Plane, *Combinatorica* **25** (2005), 561–574.
- [68] P. Keevash and B. Sudakov, On a hypergraph Turán problem of Frankl, *Combinatorica* **25** (2005), 673–706.
- [69] J. H. Kim, The Ramsey number $R(3, t)$ has order of magnitude $t^2 / \log t$, *Random Structures and Algorithms* **7** (1995), 173–207.
- [70] J. Komlós and M. Simonovits, Szemerédi’s regularity lemma and its applications in graph theory, in: *Combinatorics, Paul Erdős is eighty, Vol. 2* (Keszthely, 1993), 295–352, Bolyai Soc. Math. Stud., 2, Jnos Bolyai Math. Soc., Budapest, 1996.
- [71] A. Kostochka, A class of constructions for Turán’s $(3, 4)$ -problem, *Combinatorica* **2** (1982), 187–192.
- [72] A. Kostochka and V. Rödl, On graphs with small Ramsey numbers, *J. Graph Theory* **37** (2001), 198–204.
- [73] A. Kostochka and V. Rödl, On graphs with small Ramsey numbers II, *Combinatorica* **24** (2004), 389–401.
- [74] A. Kostochka and B. Sudakov, On Ramsey numbers of sparse graphs, *Combin. Probab. Comput.* **12** (2003), 627–641.
- [75] T. Kővári, V.T. Sós and P. Turán, On a problem of K. Zarankiewicz, *Colloquium Math.* **3** (1954), 50–57.
- [76] M. Krivelevich, On the edge distribution in triangle-free graphs, *J. Combin. Theory Ser. B* **63** (1995), 245–260.
- [77] M. Krivelevich and B. Sudakov, Pseudo-random graphs, in: *More Sets, Graphs and Numbers*, Bolyai Society Mathematical Studies 15, Springer, 2006, 199–262.
- [78] W. Mantel, Problem 28, *Wiskundige Opgaven* **10** (1907), 60–61.
- [79] B. Nagle, V. Rödl and M. Schacht, The counting lemma for regular k -uniform hypergraphs, *Random Structures and Algorithms* **28** (2006), 113–179.
- [80] V. Nikiforov, The number of cliques in graphs of given order and size, *Transactions of AMS*, to appear.
- [81] F.P. Ramsey, On a problem of formal logic, *Proc. London Math. Soc. Ser. 2* **30** (1930), 264–286.

- [82] A. Razborov, On the minimal density of triangles in graphs, *Combin. Probab. Comput.* **17** (2008), 603–618.
- [83] A. Razborov, On 3-hypergraphs with forbidden 4-vertex configurations, preprint.
- [84] V. Rödl and J. Skokan, Regularity lemma for k -uniform hypergraphs, *Random Structures and Algorithms* **25** (2004), 1–42.
- [85] V. Rödl and R. Thomas, Arrangeability and clique subdivisions, in: *The mathematics of Paul Erdős, II*, Algorithms Combin. 14, Springer, Berlin, 1997, 236–239.
- [86] I. Ruzsa and E. Szemerédi, Triples systems with no six points carrying three triangles, in: *Combinatorics (Proc. Fifth Hungarian Colloq., Keszthely, 1976)*, Vol. II, Colloq. Math. Soc. Bolyai 18, North-Holland, Amsterdam, 1978, 939–945.
- [87] M. Schacht, Extremal results for random discrete structures, preprint.
- [88] L. Shi, Cube Ramsey numbers are polynomial, *Random Structures & Algorithms*, **19** (2001), 99–101.
- [89] A. F. Sidorenko, A correlation inequality for bipartite graphs, *Graphs Combin.* **9** (1993), 201–204.
- [90] M. Simonovits, Extremal graph problems, degenerate extremal problems and super-saturated graphs, in: *Progress in graph theory* (A. Bondy ed.), Academic, New York, 1984, 419–437.
- [91] J. Solymosi, Note on a generalization of Roth’s theorem, in: *Discrete and computational geometry*, Algorithms Combin. 25, Springer, Berlin, 2003, 825–827.
- [92] J. Spencer, Asymptotic lower bounds for Ramsey functions, *Discrete Math.* **20** (1977/78), 69–76.
- [93] J. Spencer, E. Szemerédi and W. Trotter, Unit distances in the Euclidean plane, in: *Graph theory and combinatorics (Cambridge, 1983)*, Academic Press, London, 1984, 293–303.
- [94] B. Sudakov, Few remarks on the Ramsey-Turan-type problems, *J. Combinatorial Theory Ser. B* **88** (2003), 99–106.
- [95] B. Sudakov, A conjecture of Erdos on graph Ramsey numbers, submitted.
- [96] B. Sudakov, Making a K_4 -free graph bipartite, *Combinatorica* **27** (2007), 509–518.
- [97] B. Sudakov, T. Szabo and V. Vu, A generalization of Turán’s theorem, *J. Graph Theory* **49** (2005), 187–195.
- [98] E. Szemerédi, On sets of integers containing no k elements in arithmetic progression, *Acta Arith.* **27** (1975), 199–245.
- [99] P. Turán, On an extremal problem in graph theory (in Hungarian), *Mat. Fiz. Lapok* **48**, (1941) 436–452.
- [100] T. Wooley, Problem 2.8, *Problem presented at the workshop on Recent Trends in Additive Combinatorics*, AIM, Palo Alto, 2004, <http://www.aimath.org/WWN/additivecomb/additivecomb.pdf>.

Section 15

**Mathematical Aspects of
Computer Science**

This page is intentionally left blank

Smoothed Analysis of Condition Numbers

Peter Bürgisser*

Abstract

We present some recent results on the probabilistic behaviour of interior point methods for the convex conic feasibility problem and for homotopy methods solving complex polynomial equations. As suggested by Spielman and Teng, the goal is to prove that for all inputs (even ill-posed ones), and all slight random perturbations of that input, it is unlikely that the running time will be large. These results are obtained through a probabilistic analysis of the condition of the corresponding computational problems.

Mathematics Subject Classification (2010). 65H20, 65Y20, 68Q25, 90C31

Keywords. Condition number, distance to ill-posedness, analysis of algorithms, smoothed analysis, volume of tubes, convex conic feasibility problem, Renegar's condition number, interior point methods, polynomial equation solving, homotopy methods, polynomial time, Smale's 17th problem

1. Introduction

In computer science, the most common theoretical approach to understanding the behaviour of algorithms is *worst-case analysis*. This means proving a bound on the worst possible performance an algorithm can have. In many situations this gives satisfactory answers. However, there are cases of algorithms that perform exceedingly well in practice and still have a provably bad worst-case behaviour. A famous example is Dantzig's simplex algorithm. In an attempt to rectify this discrepancy, researchers have introduced the concept of *average-case analysis*, which means bounding the expected performance of an algorithm on random inputs. For the simplex algorithm, average-case analyses have been

*Supported by DFG grants BU 1371/2-1, BU 1371/3-1, and Paderborn Institute for Scientific Computation (PaSCo).

Institute of Mathematics, University of Paderborn, D-33098 Paderborn, Germany.
E-mail: pbuerg@upb.de.

first given by Borgwardt [13] and Smale [63]. However, while a proof of good average performance yields an indication of a good performance in practice, it can rarely explain it convincingly. The problem is that the results of an average-case analysis strongly depend on the distribution of the inputs, which is unknown, and usually assumed to be Gaussian for rendering the mathematical analysis feasible.

Spielman and Teng [67] suggested in 2001 the concept of *smoothed analysis* as a new form of analysis of algorithms that arguably blends the best of both worst-case and average-case. They used this new framework to give a more compelling explanation of the simplex method (for the shadow vertex pivot rule), see [69].

The general idea of smoothed analysis is easy to explain. Let $T: \mathbb{R}^p \supseteq \mathcal{D} \rightarrow \mathbb{R}_+ \cup \{\infty\}$ be any function (measuring running time etc). Instead of showing “it is unlikely that $T(a)$ will be large,” one shows that “for all \bar{a} and all slight random perturbations a of \bar{a} , it is unlikely that $T(a)$ will be large.” We model the perturbation a by a normal distribution $N(\bar{a}, \sigma^2 \mathbf{I})$ with center \bar{a} and covariance matrix $\sigma^2 \mathbf{I}$, given by the density

$$\rho(a) = \left(\frac{1}{\sigma\sqrt{2\pi}} \right)^p \cdot \exp \left(-\frac{\|a - \bar{a}\|^2}{2\sigma^2} \right).$$

The goal of a smoothed analysis of T is to give good estimates of

$$\sup_{\bar{a} \in \mathcal{D}} \text{Prob}_{a \sim N(\bar{a}, \sigma^2 \mathbf{I})} \{T(a) \geq \varepsilon^{-1}\}.$$

In a first approach, one may focus on expectations, that is, on bounding

$$\sup_{\bar{a} \in \mathcal{D}} \mathbb{E}_{a \sim N(\bar{a}, \sigma^2 \mathbf{I})} T(a).$$

Figure 1 succinctly summarizes the three types of analysis of algorithms.

Worst-case analysis	Average-case analysis	Smoothed analysis
$\sup_{a \in \mathcal{D}} T(a)$	$\mathbb{E}_{a \sim N(0, \mathbf{I})} T(a)$	$\sup_{\bar{a} \in \mathcal{D}} \mathbb{E}_{a \sim N(\bar{a}, \sigma^2)} T(a)$

Figure 1. Three types of analysis of algorithms.

Smoothed analysis is not only useful for analyzing the simplex algorithm, but can be applied to a wide variety of numerical algorithms. For doing so, understanding the concept of condition numbers is an important intermediate step.

A distinctive feature of the computations considered in numerical analysis is that they are affected by errors. A main character in the understanding of

the effects of these errors is the *condition number* of the input. This is a positive number which, roughly speaking, quantifies the errors when computations are performed with infinite precision but the input has been modified by a small perturbation. The condition number depends only on the data and the problem at hand (but not on the algorithm). The best known condition number is that for matrix inversion and linear equation solving. For a square matrix A it takes the form $\kappa(A) = \|A\| \cdot \|A^{-1}\|$ and was independently introduced by von Neumann and Goldstine [46] and Turing [71].

Condition numbers are omnipresent in round-off analysis. They also appear as a parameter in complexity bounds for a variety of efficient iterative algorithms in linear algebra, linear and convex optimization, as well as homotopy methods for solving systems of polynomial equations. The running time $T(a, \varepsilon)$ of these algorithms, measured as the number of arithmetic operations, can often be bounded in the form

$$T(a, \varepsilon) \leq (\text{size}(a) + \mu(a) + \log \varepsilon^{-1})^c, \quad (1)$$

with some universal constant $c > 0$. Here the input is a vector $a \in \mathbb{R}^n$ of real numbers, $\text{size}(a) = n$ is the dimension of a , the positive parameter ε measures the required accuracy, and $\mu(a)$ is some measure of conditioning of a . (Depending on the situation, $\mu(a)$ may be either a condition number or its logarithm. Moreover, $\log \varepsilon^{-1}$ might be replaced by $\log \log \varepsilon^{-1}$.)

Smale [65] proposed a two-part scheme for dealing with *complexity upper bounds* in numerical analysis. The first part consists of establishing bounds of the form (1). The second part of the scheme is to analyze the distribution of $\mu(a)$ under the assumption that the inputs a are random with respect to some probability distribution. More specifically, we aim at tail estimates of the form

$$\text{Prob} \{ \mu(a) \geq \varepsilon^{-1} \} \leq \text{size}(a)^c \varepsilon^\alpha \quad (\varepsilon > 0)$$

with universal constants $c, \alpha > 0$. In a first attempt, one may try to show upper bounds on the expectation of $\mu(a)$ (or $\log \mu(a)$, depending on the situation). Combining the two parts of the scheme, we arrive at upper bounds for the average running time of our specific numerical algorithms considered. So if we content ourselves with statements about the probabilistic average-case, we can eliminate the dependence on $\mu(a)$ in (1). This approach was elaborated upon for average-case complexity by Blum and Shub [11], Renegar [47], Demmel [29], Kostlan [40], Edelman [33, 34], Shub [54], Shub and Smale [59, 60], Cheung and Cucker [23], Cucker and Wschebor [26], Cheung et al. [24], Beltrán and Pardo [6], Bürgisser et al. [20], and others.

Spielman and Teng in their ICM 2002 paper [68] proposed to refine part two of Smale's scheme by performing a smoothed analysis of the condition number $\mu(a)$ involved for obtaining more meaningful probabilistic upper complexity bounds. The implementation of this idea has been a success story. The goal of

this survey is to present some of the recent results in this direction. Beside the original papers the interested reader may also consult the survey [14] and the forthcoming book [17].

Acknowledgments. I thank Dennis Amelunxen, Felipe Cucker, and Javier Peña for constructive comments on the manuscript.

2. Conic Condition Numbers

Often, a probabilistic analysis of condition numbers can be done in a systematic way by geometric tools. Let us explain this approach for Turing’s condition number $\kappa(A) = \|A\| \cdot \|A^{-1}\|$ of a matrix $A \in \mathbb{R}^{n \times n}$. This quantity measures the sensitivity or errors for the tasks of inverting A or of solving the linear system $Ax = b$. We interpret $\Sigma = \{B \in \mathbb{R}^{n \times n} \mid \det B = 0\}$ as the “set of ill-posed inputs” for these tasks. It is mathematically convenient to measure distances between matrices with the Euclidean or Frobenius norm $\|A\|_F := (\text{trace}(AA^T))^{1/2}$. We replace the spectral norm $\|A\|$ by the larger Frobenius norm $\|A\|_F$ and, instead of $\kappa(A)$, study the larger quantity $\kappa_F(A) := \|A\|_F \cdot \|A^{-1}\|$. The Eckart-Young Theorem [32] states that $\kappa_F(A)$ is inversely proportional to the distance of A to Σ . More specifically, we have

$$\kappa_F(A) = \frac{\|A\|_F}{\text{dist}(A, \Sigma)}. \quad (2)$$

If the entries of A are independent standard normal distributed, then $A/\|A\|_F$ is uniformly distributed on the unit sphere S^{n^2-1} . Since κ_F is scale-invariant, we may assume that the inputs A are chosen uniformly at random in S^{n^2-1} . We also write $\Sigma_S := \Sigma \cap S^{n^2-1}$. The ε -neighborhood of Σ_S , for $0 < \varepsilon \leq 1$, is defined as

$$T(\Sigma_S, \varepsilon) := \{A \in S^{n^2-1} \mid d_S(A, \Sigma_S) < \arcsin \varepsilon\}, \quad (3)$$

where $d_S(A, \Sigma_S) := \inf\{d_S(A, B) \mid B \in \Sigma_S\}$ and $d_S(A, B)$ denotes the angular distance of A and B in S^{n^2-1} . Using $\text{dist}(A, \Sigma) = \sin d_S(A, \Sigma_S)$ we obtain from (2) for $0 < \varepsilon \leq 1$

$$\text{Prob}\{\kappa_F(A) \geq \varepsilon^{-1}\} = \frac{\text{vol } T(\Sigma_S, \varepsilon)}{\text{vol } S^{n^2-1}}.$$

The task is therefore to compute or to estimate the volume of neighborhoods of Σ_S .

This approach applies to a much more general context than just the matrix condition number. Assume that \mathbb{R}^{p+1} is the data space of a computational problem under consideration and the set of “ill-posed inputs” $\Sigma \subseteq \mathbb{R}^{p+1}$ is an algebraic cone, i.e., a real algebraic set that is closed by multiplications with

scalars. We associate with Σ the *conic condition number* function defined as

$$\mathcal{C}: \mathbb{R}^{p+1} \setminus \{0\} \rightarrow \mathbb{R}, \quad \mathcal{C}(a) := \frac{\|a\|}{\text{dist}(a, \Sigma)},$$

where $\|\cdot\|$ and dist refer to the Euclidean norm. For instance the matrix condition number κ_F is conic due to the Eckart-Young Theorem (2). The homogeneity of \mathcal{C} allows us to restrict to inputs a lying in the unit sphere S^p , so that the conic condition number $\mathcal{C}(a)$ takes the form

$$\mathcal{C}(a) = \frac{1}{\text{dist}(a, \Sigma)} = \frac{1}{\sin d_S(a, \Sigma_S)},$$

where $\Sigma_S := \Sigma \cap S^p$ and d_S refers to the angular distance on S^p .

Demmel [29] derived a general result giving an average-case analysis for conic condition numbers in terms of geometric invariants of the corresponding set of ill-posed inputs Σ . This is based on general estimates on the volume of neighborhoods of Σ_S obtained with integral-geometric tools. The core of these ideas, in the context of one variable polynomial equation solving, can already be found in Smale's early AMS bulletin article [62] dating from 1981.

Bürgisser et al. [18, 19] recently extended Demmel's result from average-case analysis to a natural geometric framework of smoothed analysis of conic condition numbers, called *uniform smoothed analysis*. Suppose that \mathcal{C} is a conic condition number as above associated with the set Σ of ill-posed inputs. For $0 \leq \sigma \leq 1$ let $B(\bar{a}, \sigma)$ denote the spherical cap in the sphere S^p centered at $\bar{a} \in S^p$ and having angular radius $\arcsin \sigma$. Moreover, we define for $0 < \varepsilon \leq 1$ the ε -neighborhood of Σ_S as in (3). The task of a uniform smoothed analysis of \mathcal{C} consists of providing good upper bounds on

$$\sup_{\bar{a} \in S^p} \text{Prob}_{a \in B(\bar{a}, \sigma)} \{\mathcal{C}(a) \geq \varepsilon^{-1}\},$$

where a is assumed to be chosen uniformly at random in $B(\bar{a}, \sigma)$. The probability occurring here has an immediate geometric meaning:

$$\text{Prob}_{a \in B(\bar{a}, \sigma)} \{\mathcal{C}(a) \geq \varepsilon^{-1}\} = \frac{\text{vol}(T(\Sigma_S, \varepsilon) \cap B(\bar{a}, \sigma))}{\text{vol}(B(\bar{a}, \sigma))}. \quad (4)$$

Thus uniform smoothed analysis means to provide bounds on the relative volume of the intersection of ε -neighborhoods of Σ_S with small spherical disks, see Figure 2. We note that uniform smoothed analysis interpolates transparently between worst-case and average-case analysis. Indeed, when $\sigma = 0$ we get worst-case analysis, while for $\sigma = 1$ we obtain average-case analysis.

The following result from Bürgisser et al. [19] extends the previously mentioned result by Demmel [29] from average-case to smoothed analysis.

Theorem 2.1. *Let \mathcal{C} be a conic condition number with set Σ_S of ill-posed inputs. Assume that Σ_S is contained in a real algebraic hypersurface, given as*

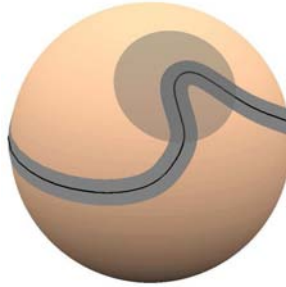


Figure 2. Neighborhood of the curve Σ_S intersected with a spherical disk.

the zero set of a homogeneous polynomial of degree d . Then, for all $0 < \sigma \leq 1$ and all $0 < \varepsilon \leq \sigma/(p(2d+1))$ we have

$$\sup_{\bar{a} \in S^p} \text{Prob}_{a \in B(\bar{a}, \sigma)} \{ \mathcal{C}(a) \geq \varepsilon^{-1} \} \leq 26 dp \frac{\varepsilon}{\sigma},$$

$$\sup_{\bar{a} \in S^p} \mathbb{E}_{a \in B(\bar{a}, \sigma)} (\ln \mathcal{C}(a)) \leq 2 \ln(dp) + 2 \ln \frac{1}{\sigma} + 4.7.$$

The proof relies on a classical paper by Weyl [75] in which a formula for the volume of the ε -neighborhood of a submanifold M of the sphere S^p was derived. In this formula, integrals of (absolute) curvature of M enter. In [19], the integrals of absolute curvature of a smooth algebraic hypersurface M of S^p were bounded in terms of the degree d of M by means of Chern's principal kinematic formula of integral geometry [21] and Bézout's Theorem. The smoothness assumption can then be removed by a perturbation argument.

In Cucker et al. [25] it was shown that Theorem 2.1 is quite robust with respect to the assumption on the distribution modeling the perturbations. The bound on the expectation extends (in order of magnitude) to any radially symmetric probability distributions supported on a spherical disk of radius σ whose density may even have a mild singularity at the center of the perturbation.

The setting of conic condition numbers has a natural counterpart over the complex numbers. In this setting, a result similar to Theorem 2.1 was obtained in Bürgisser et al. [18]. The critical parameter entering the estimates is again the degree but now algebraic varieties of higher codimension are taken into account as well.

Demmel's paper [29] also dealt with both complex and real problems. For complex problems he provided complete proofs. For real problems, Demmel's bounds rely on an unpublished (and apparently unavailable) result by Ocneanu on the volumes of tubes around real algebraic varieties. A second goal of Bürgisser et al. [18] was to prove a result akin to Ocneanu's.

Theorem 2.1 has a wide range of applications to linear equation solving, eigenvalue computation, and polynomial equation solving. It easily gives the following uniform smoothed analysis of the condition number of a matrix $A \in$

$\mathbb{R}^{n \times n}$:

$$\sup_{\|\bar{A}\|_F=1} \mathbb{E}_{A \in B(\bar{A}, \sigma)} (\ln \kappa(A)) = \mathcal{O}\left(\frac{n}{\sigma}\right).$$

Sharper results (for Gaussian perturbations) were obtained by Wschebor [79] and Sankar et al. [53]. A paper by Tao and Vu [70] deals with the condition number of integer matrices under random discrete perturbations.

3. Convex Conic Feasibility Problem

For simplicity we focus on the complexity of feasibility problems and leave out the discussion of the related convex optimization problems.

3.1. Renegar's condition number. Let X and Y be real finite-dimensional vector spaces endowed with norms. Further, let $K \subseteq X$ be a closed convex cone that is assumed to be regular, that is $K \cap (-K) = \{0\}$ and K has nonempty interior. We denote by $L(X, Y)$ the space of linear maps from X to Y endowed with the operator norm. Given $A \in L(X, Y)$, consider the feasibility problem in primal form of deciding

$$\exists x \in X \setminus \{0\} \quad Ax = 0, \quad x \in K. \quad (5)$$

Two special cases of this general framework should be kept in mind. For $K = \mathbb{R}_+^n$, the nonnegative orthant in $X = \mathbb{R}^n$, one obtains the homogeneous *linear programming feasibility problem*. The feasibility version of homogeneous *semidefinite programming* corresponds to the cone $K = \mathcal{S}_+^n$ consisting of the positive semidefinite matrices in $X = \{x \in \mathbb{R}^{n \times n} \mid x = x^T\}$.

The feasibility problem dual to (5) is

$$\exists y^* \in Y^* \setminus \{0\} \quad A^*y^* \in K^*. \quad (6)$$

Here X^*, Y^* are the dual spaces of X, Y , respectively, $A^* \in L(Y^*, X^*)$ denotes the map adjoint to A , and $K^* := \{x^* \in X^* \mid \forall x \in K \langle x^*, x \rangle \geq 0\}$ denotes the cone dual to K .

We denote by \mathcal{P} the set of instances $A \in L(X, Y)$ for which the primal problem (5) is feasible. Likewise, we denote by \mathcal{D} the set of $A \in L(X, Y)$ for which the dual problem (6) is feasible.

\mathcal{P} and \mathcal{D} are closed subsets of $L(X, Y)$ and the separation theorem implies that $L(X, Y) = \mathcal{P} \cup \mathcal{D}$, cf. Rockafellar [52]. One can show that $\Sigma := \mathcal{P} \cap \mathcal{D}$ is the common boundary of both of the sets \mathcal{P} and \mathcal{D} . The *conic feasibility problem for K* is to decide for given $A \in L(X, Y)$ whether $A \in \mathcal{P}$ or $A \in \mathcal{D}$. The set Σ can be considered as the *set of ill-posed instances* for the conic feasibility problem. Indeed, for given $A \in \Sigma$, arbitrarily small perturbations of A may yield instances in both \mathcal{P} and \mathcal{D} . We note that Σ is a cone that is neither convex nor an algebraic set.

Jim Renegar [49, 50, 51] defined the *condition number* $\mathcal{C}_R(A)$ of an instance A of the conic feasibility problem for K by

$$\mathcal{C}_R(A) := \frac{\|A\|}{\text{dist}(A, \Sigma)}, \quad (7)$$

where $\|\cdot\|$ denotes the spectral norm and dist refers to the corresponding metric. This definition can be rephrased as follows. Suppose $A \in \mathcal{P}$. Then $1/\mathcal{C}_R(A)$ is the supremum over all $\delta > 0$ such that

$$\forall A' \in \mathbb{R}^{m \times n} : \frac{\|A' - A\|}{\|A\|} < \delta \implies A' \in \mathcal{P}. \quad (8)$$

Roughly, $1/\mathcal{C}_R(A)$ is the largest normwise relative error of A that makes A stay in \mathcal{P} . An analogous characterization applies for $A \in \mathcal{D}$.

The most efficient known algorithms for solving convex optimization problems in theory and practice are interior-point methods, cf. Nesterov and Nemirovskii [44]. Renegar [50, 51] was the first to realize that the number of steps of interior-point algorithms solving the conic feasibility problem can be effectively bounded in terms of $\mathcal{C}_R(A)$. Early work related to this is Vavasis and Ye [72] and Nesterov et al. [45]. Condition-based analyses also exist for other algorithms in convex optimization. For Khachiyan's ellipsoid method [41] such analysis was performed by Freund and Vera [35]. For the perceptron algorithm, condition-based analyses were given for linear programming by Dunagan and Vempala [31] and for general convex conic systems by Belloni, Freund and Vempala [5].

Vera et al. [73] recently showed the following general result. One can relax the above pair of conic feasibility problems (5), (6), to a primal-dual pair of conic optimization problems. When K is a self-scaled cone with a known self-scaled barrier function, the conic programming relaxation can be solved via a primal-dual interior-point algorithm. Moreover, for a well-posed instance A , a strict solution to one of the two original conic systems can be obtained in $\mathcal{O}(\sqrt{\nu} \log(\nu \mathcal{C}_R(A)))$ interior-point iterations. Here ν is a complexity parameter of the self-scaled barrier function of K that equals n in the interesting cases $K = \mathbb{R}_+^n$ and $K = \mathcal{S}_+^n$. An important feature of this algorithm is that the condition of the systems of equations that arise at each interior-point iteration grows in a controlled manner and remains bounded by a constant factor times $\mathcal{C}_R(A)^2$ throughout the entire algorithm.

We specialize now the discussion to the case of linear programming. That is, we consider the cone $K = \mathbb{R}_+^n$ in $X = \mathbb{R}^n$. Note that K is self-dual, i.e., $K^* = K$ when identifying X with its dual space. We set $Y = \mathbb{R}^m$ with $n \geq m$ and view $A \in L(X, Y)$ as an $m \times n$ -matrix with the columns $a_1, \dots, a_n \in \mathbb{R}^m$.

The primal feasibility problem (5) now reads as

$$\exists x \in \mathbb{R}^n \setminus \{0\} \ Ax = 0, x \geq 0.$$

Geometrically, this means that 0 lies in the interior of the convex hull Δ of a_1, \dots, a_n . The dual feasibility problem (6) translates to

$$\exists y \in \mathbb{R}^m \setminus \{0\} \quad A^T y \geq 0$$

meaning that Δ lies in some closed halfspace H (with $0 \in \partial H$). Since $\Sigma = \mathcal{P} \cap \mathcal{D}$, an instance A is ill-posed iff 0 lies in the convex hull Δ of a_1, \dots, a_n and Δ is contained in some closed halfspace.

We note that individual scaling of the columns a_i does not change membership of A in \mathcal{P} or \mathcal{D} , respectively. It therefore makes sense to measure relative errors of A componentwise. The resulting *GCC-condition number* $\mathcal{C}(A)$ has been introduced and investigated by Goffin [38] and Cheung and Cucker [22]. Formally, $1/\mathcal{C}(A)$ is defined as the supremum over all $\delta > 0$ such that (8) holds with $\|A' - A\|/\|A\|$ replaced by $\max_i \|a'_i - a_i\|/\|a_i\|$. We remark that $\mathcal{C}(A)$ differs from $\mathcal{C}_R(A)$ by at most a factor of \sqrt{n} if the a_i have equal norms.

In the following we will assume the normalization $\|a_i\| = 1$. Hence we can interpret the matrix A with columns a_1, \dots, a_n as an element in the product $\mathbb{S} := S^{m-1} \times \dots \times S^{m-1}$ of the spheres S^{m-1} . An advantage of the GCC-condition number is that it has various nice geometric characterizations that greatly facilitate its probabilistic analysis. When introducing the metric $d_{\mathbb{S}}$ on \mathbb{S} by $d(A, B) := \max_i d_S(a_i, b_i)$ with d_S denoting angular distance on S^{m-1} , and writing $\Sigma_{\mathbb{S}} := \Sigma \cap \mathbb{S}$, the definition of $\mathcal{C}(A)$ can be rephrased as

$$\mathcal{C}(A) = \frac{1}{\sin d_{\mathbb{S}}(A, \Sigma_{\mathbb{S}})}.$$

This characterization can be turned into a more specific form. Let $\rho(A)$ be the angular radius of a *spherical cap of minimal radius* containing $a_1, \dots, a_n \in S^{m-1}$. It is easy to see that $\rho(A) \leq \frac{\pi}{2}$ iff $A \in \mathcal{D}$. Hence, $\rho(A) = \frac{\pi}{2}$ iff $A \in \Sigma$. The following characterization is due to Cheung and Cucker [22]

$$d_{\mathbb{S}}(A, \Sigma_{\mathbb{S}}) = \begin{cases} \frac{\pi}{2} - \rho(A) & \text{if } A \in \mathcal{D} \\ \rho(A) - \frac{\pi}{2} & \text{if } A \in \mathcal{P}. \end{cases}$$

It follows that $\mathcal{C}(A)^{-1} = \sin d_{\mathbb{S}}(A, \Sigma_{\mathbb{S}}) = |\cos \rho(A)|$.

3.2. Average and smoothed analysis. The average-case analysis of the GCC condition number is intimately related to a classical question on covering a sphere by random spherical caps.

Suppose that the entries of the matrix $A \in \mathbb{R}^{m \times n}$ are independent standard Gaussian random variables. After normalization this means that each column a_i is independently chosen from the uniform distribution on the sphere S^{m-1} . Let $p(n, m, \alpha)$ denote the probability that randomly chosen spherical caps with

centers a_1, \dots, a_n and angular radius α do *not* cover the sphere S^{m-1} . We claim that

$$p(n, m, \alpha) = \text{Prob} \{ \rho(A) \leq \pi - \alpha \}.$$

Indeed, the caps of radius α with center a_1, \dots, a_n do not cover S^{m-1} iff there exists $y \in S^{m-1}$ having distance greater than α from all a_i . The latter means that the cap of radius $\pi - \alpha$ centered at $-y$ contains all the a_i , which implies $\rho(A) \leq \pi - \alpha$ and vice versa.

The problem of determining the coverage probabilities $p(n, m, \alpha)$ is classical and completely solved only for $m \leq 2$ (Gilbert [37], Miles [43]). For $m > 2$ little was known except

$$p(n, m, \pi/2) = \frac{1}{2^{n-1}} \sum_{k=0}^m \binom{n-1}{k}$$

due to Wendel [74] and asymptotic formulas for $p(n, m, \alpha)$ for $\alpha \rightarrow 0$ due to Janson [39]. Bürgisser et al. [20] recently discovered a *closed formula* for $p(n, m, \alpha)$ in the case $\alpha \geq \pi/2$ and an upper bound for $p(n, m, \alpha)$ in the case $\alpha \leq \pi/2$. In particular, this implies

$$\mathbb{E}(\ln \mathcal{C}(A)) \leq 2 \ln m + 3.31. \tag{9}$$

A smoothed analysis of a condition number of linear programming was first obtained by Dunagan et al. [30]. They obtained the following excellent result for Renegar’s condition number $\mathcal{C}_R(A)$ of a matrix $A \in \mathbb{R}^{m \times n}$ with $n \geq m$:

$$\sup_{\|\bar{A}\|=1} \mathbb{E}_{A \sim N(\bar{A}, \sigma^2 I)} (\ln \mathcal{C}_R(A)) = \mathcal{O}\left(\frac{n}{\sigma}\right). \tag{10}$$

This implies the bound $\mathcal{O}(\sqrt{n} \ln \frac{n}{\sigma})$ on the smoothed expected number of iterations of the above mentioned interior-point algorithms for the conic feasibility problem in the LP-case $K = \mathbb{R}_+^n$.

For the GCC condition number a similar result can be obtained in the model of uniform smoothed analysis by different methods. More specifically, fix $\bar{a}_i \in S^{m-1}$ for $i = 1, \dots, n$ and, independently for each i , choose a_i uniformly at random in the spherical cap $B(\bar{a}_i, \sigma)$ of S^{m-1} centered at \bar{a}_i with angular radius $\arcsin \sigma$. That is, we choose $A \in B(\bar{A}, \sigma) := \prod_i B(\bar{a}_i, \sigma)$ uniformly at random. Amelunxen and Bürgisser [2] showed the following uniform tail bound: for $0 < \varepsilon \leq \sigma/(2m(m+1))$ we have

$$\sup_{\bar{A} \in \mathbb{S}} \text{Prob} \{ A \in \mathcal{D}, \mathcal{C}(A) \geq \varepsilon^{-1} \} \leq 6.5 nm^2 \frac{\varepsilon}{\sigma}.$$

For the primal feasible case ($A \in \mathcal{P}$) a slightly worse tail estimate was obtained. This implies a bound on the expectation similar to (10)

$$\sup_{\bar{A} \in \mathbb{S}} \mathbb{E}_{A \in B(\bar{A}, \sigma)} (\ln \mathcal{C}(A)) = \mathcal{O}\left(\ln \frac{n}{\sigma}\right). \tag{11}$$

The proof of this result is based on similar ideas as for Theorem 2.1. One of the points of [2] was to show that the bound (11) is *robust* in the sense of [25]: it extends to radially symmetric probability distributions supported on $B(\bar{a}_i, \sigma)$ whose density may even have a mild singularity at the center of the perturbation.

3.3. Grassmann condition number. In view of the great relevance of semidefinite programming [77] it would be desirable to have a smoothed analysis of Renegar's condition number for the cone of semidefinite matrices. However, the proofs of (10) as well as of (11) crucially rely on the product structure of the cone $\mathbb{R}_+ \times \cdots \times \mathbb{R}_+$. We therefore try to address the problem for a general regular closed convex cone $K \subseteq \mathbb{R}^n$ in a different, coordinate-free way, following Amelunxen's PhD thesis [1].

We assign to an instance $A \in \mathbb{R}^{m \times n}$ of full rank $m < n$ its kernel $W := \ker A$. This is an element of the Grassmann manifold $\mathbb{G} := \text{Gr}(n - m, n)$, which is defined as the set of $(n - m)$ -dimensional linear subspaces of \mathbb{R}^n . We note that $\text{im} A^T$ equals the orthogonal complement W^\perp of W . The conic feasibility problem for K on instance A can thus be rephrased as deciding the alternative

$$(P) \quad W \cap K \neq 0 \quad \text{or} \quad (D) \quad W^\perp \cap K^* \neq 0,$$

for given W , compare (5) and (6). Since A enters this decision problem only through W , we view A as a particular way of representing the object W in the Grassmann manifold of inputs. In this setting we define the set $P_{\mathbb{G}}$ of primal feasible instances and the set $D_{\mathbb{G}}$ of dual feasible instances by

$$P_{\mathbb{G}} := \{W \in \mathbb{G} \mid W \cap K \neq 0\}, \quad D_{\mathbb{G}} := \{W \in \mathbb{G} \mid W^\perp \cap K^* \neq 0\}.$$

Let us point out that, unlike in the conic feasibility problem, we have here perfect symmetry with regard to switching from the primal to the dual given by the isometry $\text{Gr}(n - m, n) \rightarrow \text{Gr}(m, n), W \mapsto W^\perp$. The set of ill-posed instances, defined as $\Sigma_{\mathbb{G}} := P_{\mathbb{G}} \cap D_{\mathbb{G}}$, can be shown to be a hypersurface in \mathbb{G} . It is easily seen that W is ill-posed iff W touches the cone K .

The Grassmann manifold \mathbb{G} is a compact manifold with a well-defined Riemannian metric that is orthogonally invariant. Therefore the (geodesic) distance between two elements of \mathbb{G} is well-defined. In analogy with the previous developments, Amelunxen defined the *Grassmann condition number* of $W \in \mathbb{G}$ as

$$\mathcal{C}_{\mathbb{G}}(W) := \frac{1}{\sin d(W, \Sigma_{\mathbb{G}})},$$

where d denotes the geodesic distance in \mathbb{G} . In the case $W \cap K = 0$ the distance $d(W, \Sigma_{\mathbb{G}})$ has a more intuitive interpretation: it equals the angular distance between the subsets $W \cap S^{n-1}$ and $K \cap S^{n-1}$ of the sphere S^{n-1} .

The following result cleanly separates Renegar's condition number into the intrinsic Grassmann condition $\mathcal{C}_{\mathbb{G}}(W)$ and the representation-dependent matrix

condition number $\kappa(A) = \|A\| \cdot \|A^\dagger\|$ (where A^\dagger denotes the Moore-Penrose inverse of A). For $A \in \mathbb{R}^{m \times n}$ of rank m and $W = \ker A$ we have

$$\mathcal{C}_{\mathbb{G}}(W) \leq \mathcal{C}_R(A) \leq \kappa(A) \cdot \mathcal{C}_{\mathbb{G}}(W). \quad (12)$$

This was shown by Belloni and Freund [4] for the dual feasible case and extended to the primal feasible case by Amelunxen [1].

The Grassmann manifold \mathbb{G} has an orthogonally invariant volume form that defines a probability measure on \mathbb{G} . In particular it makes sense to talk about the uniform distribution on \mathbb{G} . This distribution arises naturally for $W = \ker A$ when we assume that the entries of $A \in \mathbb{R}^{m \times n}$ are independent standard Gaussian.

Amelunxen and Bürgisser [1, 3] obtained the following average-case analysis of the Grassmann condition number, which holds for *any* regular closed convex cone $K \subseteq \mathbb{R}^n$:

$$\begin{aligned} \text{Prob}_{W \in \mathbb{G}} \{ \mathcal{C}_{\mathbb{G}}(W) \geq \varepsilon^{-1} \} &\leq 6n\varepsilon \quad \text{if } \varepsilon < n^{-\frac{3}{2}}, \\ \mathbb{E}_{W \in \mathbb{G}} (\ln \mathcal{C}_{\mathbb{G}}(W)) &\leq 2.5 \ln n + 2.8. \end{aligned}$$

Here is a very brief indication of the ideas of proof. Showing the first statement means bounding the volume of the ε -neighborhood of $\Sigma_{\mathbb{G}}$ in \mathbb{G} . By a perturbation argument, it suffices to consider cones with smooth boundary ∂K of positive Gaussian curvature so that K is strictly convex. Then $M := \partial K \cap S^{n-1}$ is a smooth hypersurface in the sphere S^{n-1} . By assumption, each $W \in \Sigma_{\mathbb{G}}$ touches the cone K along a unique ray $\mathbb{R}_+ p_W$ determined by a point $p_W \in M$. The fiber over $p \in M$ of the map $\Sigma_{\mathbb{G}} \rightarrow M, W \mapsto p_W$ consists of the $(n-m)$ -dimensional subspaces W of the tangent space of $T_p \partial K$ containing the line $\mathbb{R}p$. The set of these W can be identified with the Grassmann manifold of $(n-m-1)$ -dimensional subspaces of $T_p M$. This way, one sees that $\Sigma_{\mathbb{G}} \rightarrow M$ has the structure of a Grassmann bundle over M . With some work it is possible to extend Weyl's formula [75] for the volume of ε -neighborhoods of M in S^{n-1} to obtain a formula for the ε -neighborhood of $\Sigma_{\mathbb{G}}$ in \mathbb{G} .

This approach should also yield a uniform smoothed analysis of the Grassmann condition number and we are currently elaborating the details.

We close this section with a few further remarks. In the case $K = \mathbb{R}_+^n$ we get the better bounds $\mathbb{E} (\ln \mathcal{C}_{\mathbb{G}}(A)) \leq 1.5 \ln m + 6$ only depending on m as in (9).

In view of the inequality $\ln \mathcal{C}_R(A) \leq \ln \kappa(A) + \ln \mathcal{C}_{\mathbb{G}}(W)$ resulting from (12) one may ask about the contribution of the data-dependent $\ln \kappa(A)$. Surprisingly, this contribution turns out to be bounded if m/n is bounded away from 1. It is a known fact (Geman [36], Silverstein [61]) that for standard Gaussian matrices A_n of size $m_n \times n$ with m_n/n converging to a fixed number $q \in (0, 1)$, the condition number $\kappa(A_n)$ converges to $\frac{1+\sqrt{q}}{1-\sqrt{q}}$ almost surely. Recently, this average-case analysis was complemented by a smoothed analysis by Bürgisser

and Cucker [16] who showed

$$\sup_{\|\bar{A}\|=1} \mathbb{E}_{A \sim N(\bar{A}, \sigma^2 I)} (\kappa(A)) \leq \frac{20.1}{1-q}$$

for $q \in (0, 1)$, $m/n \leq q$, and sufficiently large n . As in the average case, the bound is independent of n . Interestingly, it is also independent of σ for large n .

4. Solving Complex Polynomial Equations

4.1. Smale’s 17th problem. In 2000, Steve Smale published a list of mathematical problems for the 21st century [66]. The 17th problem in the list reads as follows:

Can a zero of n complex polynomial equations in n unknowns be found approximately, on the average, in polynomial time with a uniform algorithm?

This is the guiding problem underlying the series of papers [56, 57, 58, 60, 59] —commonly referred to as “the Bézout series” — written by Shub and Smale during the first half of the 1990s, a collection of ideas, methods, and results that pervade all the research done in Smale’s 17th problem since it was proposed.

We make now precise the different notions intervening in Smale’s 17th problem. Fix a degree pattern $\mathbf{d} = (d_1, \dots, d_n)$. The input space is the vector space $\mathcal{H}_{\mathbf{d}}$ of polynomial systems $f = (f_1, \dots, f_n)$ with $f_i = \sum_{\alpha} a_{\alpha}^i X^{\alpha} \in \mathbb{C}[X_0, \dots, X_n]$ homogeneous of degree d_i . We endow $\mathcal{H}_{\mathbf{d}}$ with the *Bombieri-Weyl Hermitian inner product* that is associated with the norm

$$\|f\|^2 := \sum_{|\alpha|=d_i} |a_{\alpha}^i|^2 \binom{d_i}{\alpha}^{-1}.$$

The reason to do so is that this inner product is invariant under the natural action of the unitary group $U(n+1)$. The quantity $N := \dim_{\mathbb{C}} \mathcal{H}_{\mathbf{d}}$ measures the size of the input system f and we further put $D := \max_i d_i$ and let $\mathcal{D} = \prod_i d_i$ be the Bézout number.

We look for solutions ζ of the equation $f(\zeta) = 0$ in the complex projective space $\mathbb{P}^n := \mathbb{P}(\mathbb{C}^{n+1})$. The expression “on the average” in Smale’s 17th problem refers to the expectation with respect to the uniform distribution on the unit sphere $S(\mathcal{H}_{\mathbf{d}})$ of $\mathcal{H}_{\mathbf{d}}$. For $f, g \in \mathcal{H}_{\mathbf{d}} \setminus \{0\}$, we denote by $d_{\mathbb{S}}(f, g)$ the angle between f and g . Similarly we define $d_{\mathbb{P}}(x, y)$ for $x, y \in \mathbb{P}^n$.

In [54], Mike Shub introduced the following *projective version of Newton’s method*. Let $Df(\zeta)|_{T_{\zeta}}$ denote the restriction of the derivative of $f: \mathbb{C}^{n+1} \rightarrow \mathbb{C}^n$ at ζ to the tangent space $T_{\zeta} := \{v \in \mathbb{C}^{n+1} \mid \langle v, \zeta \rangle = 0\}$ of \mathbb{P}^n at ζ . We associate to $f \in \mathcal{H}_{\mathbf{d}}$ a map $N_f: \mathbb{C}^{n+1} \setminus \{0\} \rightarrow \mathbb{C}^{n+1} \setminus \{0\}$ defined (almost everywhere) by

$$N_f(x) = x - Df(x)|_{T_x}^{-1} f(x).$$

Note that $N_f(x)$ is homogeneous of degree 0 in f so that N_f induces a rational map from \mathbb{P}^n to \mathbb{P}^n .

The expression “approximate zero” in Smale’s 17th problem has the following precise meaning. By an *approximate zero* of $f \in \mathcal{H}_{\mathbf{d}}$ associated with a zero $\zeta \in \mathbb{P}^n$ of f we understand a point $z \in \mathbb{P}^n$ such that the sequence of Newton iterates $z_{i+1} := N_f(z_i)$ with initial point $z_0 := z$ converges immediately quadratically to ζ , i.e., $d_{\mathbb{P}}(z_i, \zeta) \leq 2^{-(2^i-1)} d_{\mathbb{P}}(z_0, \zeta)$ for all $i \in \mathbb{N}$.

The *condition number* of f at the zero ζ measures how much does ζ change when we perturb f a little. More specifically, we consider the *solution variety* $V_{\mathbb{P}} := \{(f, \zeta) \mid f(\zeta) = 0\} \subseteq \mathcal{H}_{\mathbf{d}} \times \mathbb{P}^n$, which is a smooth Riemannian submanifold. By the implicit function theorem, the projection map $V_{\mathbb{P}} \rightarrow \mathcal{H}_{\mathbf{d}}$, $(g, x) \mapsto g$ can be locally inverted around (f, ζ) if ζ is a simple solution of f : let us denote by G its local inverse. The *condition number* $\mu(f, \zeta)$ of (f, ζ) is defined as the operator norm of the derivative of G at ζ . After some rescaling it takes the following form:

$$\mu(f, \zeta) = \|f\| \cdot \|M^\dagger\|, \quad (13)$$

where (choosing a representative of ζ with $\|\zeta\| = 1$)

$$M := \text{diag}(\sqrt{d_1}, \dots, \sqrt{d_n})^{-1} Df(\zeta) \in \mathbb{C}^{n \times (n+1)}.$$

Here $M^\dagger = M^*(MM^*)^{-1}$ denotes the Moore-Penrose inverse of M and $\|M^\dagger\|$ its spectral norm.

We remark that before Shub and Smale’s work, condition numbers for finding the roots of polynomials in one variable were defined and studied by Wilkinson [76], Woźniakowski [78], and Demmel [28].

Smale’s α -theory [64] shows that the size of the basin of attraction of a simple zero ζ for Newton’s operator N_f is controlled by $\mu(f, \zeta)$. More specifically, for z being an approximate zero of f associated with ζ , it is sufficient to have (cf. [55, 15])

$$d_{\mathbb{P}}(z, \zeta) \leq \frac{0.3}{D^{3/2} \mu(f, \zeta)}. \quad (14)$$

Finally, the notion of “uniform polynomial time algorithm” in Smale’s 17th problem refers to the so-called BSS-model [10], which is essentially a model of a random access machine operating with real numbers with infinite precision and at unit cost.

The overall idea in the Bézout series is to use a *linear homotopy*. Given a *start system* $(g, \zeta) \in V_{\mathbb{P}}$ and an input $f \in \mathcal{H}_{\mathbf{d}}$ we consider the line segment $[g, f]$ connecting g and f that consists of the systems

$$q_t := (1-t)g + tf \quad \text{for } t \in [0, 1].$$

If $[g, f]$ does not meet the discriminant variety (i.e., none of the q_t has a multiple

zero), then there exists a unique lifting to the solution variety $V_{\mathbb{P}}$

$$\gamma: [0, 1] \rightarrow V, t \mapsto (q_t, \zeta_t) \quad (15)$$

such that $q_0 = g$. The root of f we are looking for is ζ_1 (note $q_1 = f$).

The idea is to follow the path γ numerically: we choose a partition $t_0 = 0, t_1, \dots, t_k = 1$ and, writing $q_i := q_{t_i}$ and $\zeta_i := \zeta_{t_i}$, we successively compute approximations z_i of ζ_i by Newton's method starting with $z_0 := \zeta$. More specifically, we compute

$$z_{i+1} := N_{q_{i+1}}(z_i).$$

Two questions arise: how do we choose the start system (g, ζ) and how do we find the subdivision points t_i ?

The state of the art at the end of the Bézout series, i.e., in [59], showed an incomplete picture. For the choice of the subdivision, the rule consisted of taking a regular subdivision of $[g, f]$ for a given k , executing the path-following procedure, and repeating with k replaced by $2k$ if the final point could not be shown to be an approximate zero of f (a criterion for checking this follows from (14)).

As for the question of the choice of the start system (g, ζ) , Shub and Smale proved in [59] that good start systems (g, ζ) existed for each degree pattern \mathbf{d} (in the sense that the average number of iterations for the rule above was polynomial in the input size N), but they could not exhibit a procedure to generate one such start system. They conjectured in [59] that the system $g \in \mathcal{H}_{\mathbf{d}}$ given by $g_i = X_0^{d_i-1} X_i$ is a good start system. While this conjecture is supported by numerical experiments, a proof remains elusive.

After the Bézout series, the next breakthrough took a decade to come. Beltrán and Pardo proposed in [7, 8, 9] that the start system (g, ζ) should be randomly chosen. We consider the following probability distribution ρ_{st} on $V_{\mathbb{P}}$ for the start system (g, ζ) . It consists of drawing g in the sphere $S(\mathcal{H}_{\mathbf{d}}) := \{g \in \mathcal{H}_{\mathbf{d}} \mid \|g\| = 1\}$ from the uniform distribution and then choosing one of the (almost surely) \mathcal{D} zeros of g from the uniform distribution on $\{1, \dots, \mathcal{D}\}$. This procedure is clearly non-constructive, as computing a zero of a system is the problem we wanted to solve in the first place. One of the major contributions in [7] was to show that the distribution ρ_{st} can be efficiently sampled.

4.2. Average and smoothed analysis. Following a result in Shub [55], the following specific adaptive choice of the subdivision was proposed in [15]. We reparametrize the curve γ from (15) using a parameter $0 \leq \tau \leq 1$ which measures a ratio of angles. More specifically, let $\alpha = d_{\mathbb{S}}(g, f)$ and $\alpha\tau(t)$ be the angle between $g/\|g\|$ and $q_t/\|q_t\|$. As the stepsize we choose, with the parameter $\lambda = 7.53 \cdot 10^{-3}$,

$$\alpha(\tau_{i+1} - \tau_i) = \frac{\lambda}{D^{3/2} \mu(q_i, z_i)^2}$$

and call the resulting algorithm ALH (*Adaptive Linear Homotopy*). An analysis [55, 15] shows that ALH finds an approximate zero of $f = q_1$ with a number $K(f, g, \zeta)$ of steps bounded by

$$K(f, g, \zeta) \leq 217 D^{3/2} d_{\mathbb{S}}(f, g) \int_0^1 \mu(q_\tau, \zeta_\tau)^2 d\tau. \tag{16}$$

Consider the *Las Vegas algorithm LV* that on input $f \in \mathcal{H}_{\mathbf{d}}$ draws the start system $(g, \zeta) \in V_{\mathbb{P}}$ at random from the distribution $\rho_{\mathbf{st}}$ and then runs ALH on input (f, g, ζ) . The algorithm LV either outputs an approximate zero z of f or loops forever. We write

$$K(f) := \mathbb{E}_{(g, \zeta) \sim \rho_{\mathbf{st}}} (K(f, g, \zeta))$$

for the *expected number of iterations* of LV on input f . The expected running time (i.e., number of arithmetic operations) of LV is given by $K(f)$ times the cost of one iteration, the latter being dominated by that of computing one Newton iterate (which is $\mathcal{O}(N + n^3)$).

Beltrán and Pardo [9] performed an average-case analysis of LV showing that

$$E_{f \in S(\mathcal{H}_{\mathbf{d}})} K(f) = \mathcal{O}(nND^{3/2}).$$

We note that in this result, randomness enters in two ways: as a computational technique (choice of the start system) and as a way of analyzing the algorithm (average over all inputs).

Bürgisser and Cucker [15] succeeded in giving a smoothed analysis of the algorithm LV. For making such analysis possible it was essential to model random perturbations by Gaussians. For $\bar{f} \in \mathcal{H}_{\mathbf{d}}$ and $\sigma > 0$ we denote by $\rho_{\bar{f}, \sigma}$ the density of the Gaussian distribution $N(\bar{f}, \sigma^2 \mathbf{I})$ on $\mathcal{H}_{\mathbf{d}}$ with mean \bar{f} and covariance matrix $\sigma^2 \mathbf{I}$. For technical simplicity, the smoothed analysis of LV assumes that the local perturbations follow a *truncated Gaussian distribution* $N_A(\bar{f}, \sigma^2 \mathbf{I})$ with center $\bar{f} \in \mathcal{H}_{\mathbf{d}}$ that is defined by the following density

$$\rho(f) = \begin{cases} \frac{\rho_{\bar{f}, \sigma}(f)}{P_{A, \sigma}} & \text{if } \|f - \bar{f}\| \leq A \\ 0 & \text{otherwise.} \end{cases}$$

Here $A := \sqrt{2N}$ and $P_{A, \sigma} := \text{Prob}\{\|g\| \leq A \mid g \sim N(0, \sigma^2 \mathbf{I})\}$: one can show that $P_{A, \sigma} \geq \frac{1}{2}$ for all $\sigma \leq 1$.

Here is the smoothed analysis result for LV from Bürgisser and Cucker [15]:

Theorem 4.1. *For any $0 < \sigma \leq 1$, the algorithm LV satisfies*

$$\sup_{\bar{f} \in S(\mathcal{H}_{\mathbf{d}})} \mathbb{E}_{f \sim N_A(\bar{f}, \sigma^2 \mathbf{I})} K(f) = \mathcal{O}\left(\frac{nND^{3/2}}{\sigma}\right).$$

Average (or smoothed) complexity results do not provide information on the running time of an algorithm for the instance at hand. In [15] a condition based analysis for LV was achieved. It bounds the number of iterations on input f in terms of the *maximum condition* of f defined in [56] as

$$\mu_{\max}(f) := \max_{\zeta | f(\zeta)=0} \mu(f, \zeta).$$

Theorem 4.2. *The expected number of iterations of Algorithm LV with input $f \in S(\mathcal{H}_{\mathbf{d}})$ is bounded as*

$$K(f) = \mathcal{O}(D^3 n N \mu_{\max}^2(f)).$$

All previously known complexity bounds depended also on the condition of the intermediate systems q_t encountered along the homotopy.

The polynomials occurring in practice often have a special structure. For instance one might be interested in polynomial systems lying in a certain fixed linear subspace of $\mathcal{H}_{\mathbf{d}}$. Important examples are provided by *sparse polynomial systems*, where the set of occurring monomials is prescribed (and usually small). It is a challenging research problem to analyze the behaviour of homotopy algorithms for sparse random input systems in a meaningful way. For work in this direction we refer to Dedieu [27] and Malajovich and Rojas [42].

4.3. A near solution to Smale's 17th problem. Even though randomized algorithms are efficient in theory and reliable in practice they do not offer an answer to the question of the existence of a *deterministic* algorithm computing approximate zeros of complex polynomial systems in average polynomial time. We shall exhibit a deterministic algorithm finding an approximate zero of a given polynomial system that works in nearly-polynomial average time, more precisely in average time $N^{\mathcal{O}(\log \log N)}$.

In the case $D \leq n$ we apply algorithm ALH with the starting system (U, \mathbf{z}) , where $U_i = X_i^{d_i} - X_0^{d_i}$ and $\mathbf{z} = (1 : 1 : \dots : 1)$. Let $K_U(f)$ denote the number of iterations of the resulting deterministic algorithm. One can show that $\mu_{\max}(U)^2 \leq 2(n+1)^D$. Using this and employing the same technique as for Theorem 4.2 one can show that (cf. [15])

$$\mathbb{E}_{f \in S(\mathcal{H}_{\mathbf{d}})} K_{\overline{U}}(f) = \mathcal{O}(D^3 N n^{D+1}).$$

For $D > n$ we use another approach, namely, a real number algorithm designed by Renegar [48] which in this case has a performance similar to the above algorithm when $D \leq n$. Putting both pieces together Bürgisser and Cucker [15] obtained a near solution to Smale's 17th problem.

Theorem 4.3. *There is a deterministic real number algorithm that on input $f \in \mathcal{H}_{\mathbf{d}}$ computes an approximate zero of f in average time $N^{\mathcal{O}(\log \log N)}$,*

where $N = \dim \mathcal{H}_{\mathbf{d}}$ measures the size of the input f . Moreover, if we restrict data to polynomials satisfying

$$D \leq n^{\frac{1}{1+\varepsilon}} \quad \text{or} \quad D \geq n^{1+\varepsilon},$$

for some fixed $\varepsilon > 0$, then the average time of the algorithm is polynomial in the input size N .

4.4. Some ideas of the proofs. It is essential that, for fixed t , $q_t = (1-t)g + tf$ follows a Gaussian law if f and g do so. Note that the variance σ_t^2 of q_t is given by $\sigma_t^2 = (1-t)^2\sigma_g^2 + t^2\sigma_f^2$, where σ_f^2 and σ_g^2 denote the variances of f and g , respectively. By a change of parameter, the integral in (16) bounding the number of steps of ALH can be estimated as

$$d_{\mathbb{S}}(f, g) \int_0^1 \mu_2(q_\tau)^2 d\tau \leq \int_0^1 \|f\| \|g\| \frac{\mu_2^2(q_t)}{\|q_t\|^2} dt, \tag{17}$$

where the mean square condition number $\mu_2(q)$ of $q \in \mathcal{H}_{\mathbf{d}}$ is defined as

$$\mu_2(q) := \left(\frac{1}{\mathcal{D}} \sum_{\zeta|q(\zeta)=0} \mu(q, \zeta)^2 \right)^{1/2}.$$

The factor $\|f\| \|g\|$ in (17) can be easily bounded and factored out the expectation. So by exchanging the expectation (over f and/or g) with the integral over t we face the problem of estimating expectations of $\mu_2^2(q_t)/\|q_t\|^2$ for different choices of the mean \bar{q}_t and the variance σ_t^2 . This is achieved by the following smoothed analysis of the mean square condition number, which is the technical heart of the proofs in [15].

Theorem 4.4. *Let $\bar{q} \in \mathcal{H}_{\mathbf{d}}$ and $\sigma > 0$. For $q \in \mathcal{H}_{\mathbf{d}}$ drawn from $N(\bar{q}, \sigma^2\mathbf{I})$ we have*

$$\mathbb{E}_q \left(\frac{\mu_2^2(q)}{\|q\|^2} \right) \leq \frac{e(n+1)}{2\sigma^2}.$$

Sketch of Proof. We distinguish points $[\zeta] \in \mathbb{P}^n$ from their representatives ζ in the sphere $\mathbb{S}^n := \{\zeta \in \mathbb{C}^{n+1} \mid \|\zeta\| = 1\}$. Note that $[\zeta] \cap \mathbb{S}^n$ is a circle with radius one. We work with the “lifting”

$$V := \{(q, \zeta) \in \mathcal{H}_{\mathbf{d}} \times \mathbb{S}^n \mid q(\zeta) = 0\}$$

of the solution variety $V_{\mathbb{P}}$, which is a vector bundle over \mathbb{S}^n with respect to the projection $\pi_2: V \rightarrow \mathbb{S}^n, (q, \zeta) \mapsto \zeta$.

For $\zeta \in \mathbb{S}^n$ we consider the following subspace R_{ζ} of $\mathcal{H}_{\mathbf{d}}$ consisting of systems h that vanish at ζ of higher order in the following sense:

$$R_{\zeta} := \{h \in \mathcal{H}_{\mathbf{d}} \mid h(\zeta) = 0, Dh(\zeta) = 0\}.$$

We further decompose the orthogonal complement R_ζ^\perp of R_ζ in \mathcal{H}_d (defined with respect to the Bombieri-Weyl Hermitian inner product). Let L_ζ denote the subspace of R_ζ^\perp consisting of the systems vanishing at ζ and let C_ζ denote its orthogonal complement in R_ζ^\perp . Then we have an orthogonal decomposition

$$\mathcal{H}_d = C_\zeta \oplus L_\zeta \oplus R_\zeta \tag{18}$$

parameterized by $\zeta \in \mathbb{S}^n$. In fact, this can be interpreted as an orthogonal decomposition of the trivial Hermitian vector bundle $\mathcal{H}_d \times \mathbb{S}^n \rightarrow \mathbb{S}^n$ into sub-bundles C , L , and R over \mathbb{S}^n . Moreover, the vector bundle V is the orthogonal sum of L and R : we have $V_\zeta = L_\zeta \oplus R_\zeta$ for all ζ .

Let \mathcal{M} denote the space $\mathbb{C}^{n \times (n+1)}$ of matrices. In the special case, where all the degrees d_i are one, the solution manifold V specializes to the manifold

$$W := \{(M, \zeta) \in \mathcal{M} \times \mathbb{S}^n \mid M\zeta = 0\}$$

and π_2 specializes to the vector bundle $p_2: W \rightarrow \mathbb{S}^n, (M, \zeta) \mapsto \zeta$ with the fibers

$$W_\zeta := \{M \in \mathcal{M} \mid M\zeta = 0\}.$$

One can show that we have *isometrical* linear maps

$$W_\zeta \rightarrow L_\zeta, M = (m_{ij}) \mapsto g_{M,\zeta} := (\sqrt{d_i} \langle X, \zeta \rangle^{d_i-1} \sum_j m_{ij} X_j). \tag{19}$$

In other words, the Hermitian vector bundles W and L over \mathbb{S}^n are isometric.

We compose the orthogonal bundle projection $V_\zeta = L_\zeta \oplus R_\zeta \rightarrow L_\zeta$ with the bundle isometry $L_\zeta \simeq W_\zeta$ obtaining the map of vector bundles

$$\Psi: V \rightarrow W, (g_{M,\zeta} + h, \zeta) \mapsto (M, \zeta)$$

whose fibers $\Psi^{-1}(M, \zeta)$ are isometric to R_ζ . The map Ψ provides the link to the condition number: by the definition (13) we have

$$\frac{\mu(q, \zeta)}{\|q\|} = \|M^\dagger\|, \quad \text{where } (M, \zeta) = \Psi(q, \zeta). \tag{20}$$

(For showing this use $Dg_{M,\zeta}(\zeta) = \text{diag}(\sqrt{d_1}, \dots, \sqrt{d_n})M$.)

Let $\rho_{\mathcal{H}_d}$ denote the density of the Gaussian $N(\bar{q}, \sigma^2 \mathbf{I})$ on \mathcal{H}_d , where $\bar{q} \in \mathcal{H}_d$ and $\sigma > 0$. For fixed $\zeta \in \mathbb{S}^n$ we decompose the mean \bar{q} as

$$\bar{q} = \bar{k}_\zeta + \bar{g}_\zeta + \bar{h}_\zeta \in C_\zeta \oplus L_\zeta \oplus R_\zeta$$

according to (18). If we denote by ρ_{C_ζ} , ρ_{L_ζ} , and ρ_{R_ζ} the densities of the Gaussian distributions in the spaces C_ζ , L_ζ , and R_ζ with covariance matrices $\sigma^2 \mathbf{I}$ and means \bar{k}_ζ , \bar{M}_ζ , and \bar{h}_ζ , respectively, then the density $\rho_{\mathcal{H}_d}$ factors as

$$\rho_{\mathcal{H}_d}(k + g + h) = \rho_{C_\zeta}(k) \cdot \rho_{L_\zeta}(g) \cdot \rho_{R_\zeta}(h). \tag{21}$$

The Gaussian density ρ_{L_ζ} on L_ζ induces a Gaussian density ρ_{W_ζ} on the fiber W_ζ with the covariance matrix $\sigma^2\mathbf{I}$ via the isometrical linear map (19), so that we have $\rho_{W_\zeta}(M) = \rho_{L_\zeta}(g_{M,\zeta})$.

Think now of choosing (q, ζ) at random from V by first choosing $q \in \mathcal{H}_d$ from $N(\bar{q}, \sigma^2\mathbf{I})$, then choosing one of its \mathcal{D} zeros $[\zeta] \in \mathbb{P}^n$ at random from the uniform distribution on $\{1, \dots, \mathcal{D}\}$, and finally choosing a representative ζ in the unit circle $[\zeta] \cap \mathbb{S}^n$ uniformly at random. We denote the resulting probability density on V by ρ_V (this is a natural extension of ρ_{st}). Then we have

$$\mathbb{E}_{\mathcal{H}_d} \left(\frac{\mu_2(q)^2}{\|q\|^2} \right) = \mathbb{E}_V \left(\frac{\mu(q, \zeta)^2}{\|q\|^2} \right), \tag{22}$$

where $\mathbb{E}_{\mathcal{H}_d}$ and \mathbb{E}_V refer to the expectations with respect to the distribution $N(\bar{q}, \sigma^2\mathbf{I})$ on \mathcal{H}_d and the probability density ρ_V on V , respectively.

To estimate the right-hand side in (22) we reduce the problem to one in the space \mathcal{M} of matrices via the map Ψ . Equation (20) implies that

$$\mathbb{E}_V \left(\frac{\mu(q, \zeta)^2}{\|q\|^2} \right) = \mathbb{E}_W (\|M^\dagger\|^2), \tag{23}$$

where \mathbb{E}_W denotes the expectation with respect to the *pushforward density* ρ_W of the density ρ_V via the map Ψ .

We have thus reduced our problem to a probability analysis of $\|M^\dagger\|$, the latter being a quantity closely related to the matrix condition number $\kappa(M) = \|M\| \cdot \|M^\dagger\|$. In order to proceed, we need to get some understanding of the probability density ρ_W .

The probability density ρ_W defines a pushforward density $\rho_{\mathbb{S}^n}$ on \mathbb{S}^n via the projection $p_2: W \rightarrow \mathbb{S}^n$, as well as *conditional probability densities* $\tilde{\rho}_{W_\zeta}$ on the fibers W_ζ , and we have

$$\mathbb{E}_W (\|M^\dagger\|^2) = \mathbb{E}_{\zeta \sim \rho_{\mathbb{S}^n}} \left(\mathbb{E}_{M \sim \tilde{\rho}_{W_\zeta}} (\|M^\dagger\|^2) \right). \tag{24}$$

(This is made formal by means of the coarea formula or Fubini’s theorem for Riemannian manifolds.) For proving Theorem 4.4 it is therefore enough to show that for all $\zeta \in \mathbb{S}^n$

$$\mathbb{E}_{M \sim \tilde{\rho}_{W_\zeta}} (\|M^\dagger\|^2) \leq \frac{e(n+1)}{2\sigma^2}. \tag{25}$$

The analysis of the situation reveals that the density $\tilde{\rho}_{W_\zeta}$ is closely related to a Gaussian, namely it has the form (c_ζ denoting a normalization factor)

$$\tilde{\rho}_{W_\zeta}(M) = c_\zeta^{-1} \cdot \det(MM^*) \rho_{W_\zeta}(M).$$

This finding allows one to prove tail bounds similarly as it was done in Sankar et al. [53, §3]. □

References

- [1] D. Amelunxen (2010), ‘Geometric analysis of the condition of the convex feasibility problem’, PhD thesis, University of Paderborn, in preparation.
- [2] D. Amelunxen and P. Bürgisser (2008), ‘Robust Smoothed Analysis of a Condition Number for Linear Programming’, accepted for *Math. Program. Ser. A*, available at arXiv:0803.0925.
- [3] D. Amelunxen and P. Bürgisser, ‘Probabilistic analysis of the Grassmann condition number (working title)’, in preparation.
- [4] A. Belloni and R. M. Freund (2009), ‘A geometric analysis of Renegar’s condition number, and its interplay with conic curvature’, *Math. Program. Ser. A* **119**, 95–107.
- [5] A. Belloni, R. M. Freund and S. Vempala (2009), ‘An efficient re-scaled perceptron algorithm for conic systems’, *Math. Oper. Res.* **34** (3), 621–641.
- [6] C. Beltrán and L. M. Pardo (2007), ‘Estimates on the distribution of the condition number of singular matrices’, *Found. Comput. Math.* **7**(1), 87–134.
- [7] C. Beltrán and L. M. Pardo (2008), ‘On Smale’s 17th problem: a probabilistic positive solution’, *Found. Comput. Math.* **8**(1), 1–43.
- [8] C. Beltrán and L.M. Pardo (2009), ‘Smale’s 17th problem: average polynomial time to compute affine and projective solutions’, *J. Amer. Math. Soc.* **22**(2), 363–385.
- [9] C. Beltrán and L. M. Pardo, ‘Fast linear homotopy to find approximate zeros of polynomial systems’, submitted.
- [10] L. Blum, F. Cucker, M. Shub and S. Smale (1998), *Complexity and Real Computation*, Springer.
- [11] L. Blum and M. Shub (1986), ‘Evaluating rational functions: infinite precision is finite cost and tractable on average’, *SIAM J. Comput.* **15**(2), 384–398.
- [12] L. Blum, M. Shub and S. Smale (1989), ‘On a theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines’, *Bull. Amer. Math. Soc. (N.S.)* **21**(1), 1–46.
- [13] K.-H. Borgwardt (1982), ‘The average number of pivot steps required by the simplex-method is polynomial’, *Z. Oper. Res. Ser.* **26**(5), 157–177.
- [14] P. Bürgisser (2009), ‘Smoothed Analysis of Condition Numbers’, in *Foundations of Computational Mathematics, Hong Kong 2008*, edited by F. Cucker, A. Pinkus, and M. J. Todd, London Mathematical Society Lecture Note Series 363, Cambridge University Press.
- [15] P. Bürgisser, and F. Cucker (2009), ‘Solving Polynomial Equations in Smoothed Polynomial Time and a Near Solution to Smale’s 17th Problem’, accepted for 42nd annual ACM Symposium on Theory of Computing (STOC 2010). Full version available at <http://arxiv.org/abs/0909.2114>.
- [16] P. Bürgisser, and F. Cucker ‘Smoothed Analysis of Moore-Penrose Inversion’, submitted. Available at arXiv:1002.4690.
- [17] P. Bürgisser, and F. Cucker ‘Condition’, book in preparation.

-
- [18] P. Bürgisser, F. Cucker and M. Lotz (2006), ‘Smoothed analysis of complex conic condition numbers’, *J. Math. Pures et Appl.* **86**, 293–309.
- [19] P. Bürgisser, F. Cucker and M. Lotz (2008), ‘The probability that a slightly perturbed numerical analysis problem is difficult’, *Math. Comp.* **77**, 1559–1583.
- [20] P. Bürgisser, F. Cucker, and M. Lotz (2010), ‘Coverage processes on spheres and condition numbers for linear programming’, *Ann. Probab.* **38**(2), 570–604.
- [21] S.-S. Chern (1966), ‘On the kinematic formula in integral geometry’, *J. Math. Mech.* **16**, 101–118.
- [22] D. Cheung and F. Cucker (2001), ‘A new condition number for linear programming’, *Math. Program.* **91**(1, Ser. A), 163–174.
- [23] D. Cheung and F. Cucker (2002), ‘Probabilistic analysis of condition numbers for linear programming’, *J. Optim. Theory Appl.* **114**(1), 55–67.
- [24] D. Cheung, F. Cucker, and R. Hauser (2005), ‘Tail decay and moment estimates of a condition number for random linear conic systems’, *SIAM J. Optim.* **15**(4), 1237–1261.
- [25] F. Cucker, R. Hauser, and M. Lotz (2010), ‘Adversarial smoothed analysis’, to appear in *J. of Complexity*.
- [26] F. Cucker and M. Wschebor (2003), ‘On the expected condition number of linear programming problems’, *Numer. Math.* **94**(3), 419–478.
- [27] J.-P. Dedieu (1999), ‘Condition number analysis for sparse polynomial systems’, *Foundations of computational mathematics (Rio de Janeiro, 1997)*, 75–101, Springer, Berlin, 1997.
- [28] J. Demmel (1987), ‘On condition numbers and the distance to the nearest ill-posed problem’, *Numer. Math.* **51**, 251–289.
- [29] J. Demmel (1988), ‘The probability that a numerical analysis problem is difficult’, *Math. Comp.* **50**, 449–480.
- [30] J. Dunagan, D. A. Spielman, and S.-H. Teng (2009), ‘Smoothed analysis of condition numbers and complexity implications for linear programming’, *Math. Program.* Ser. A, DOI 10.1007/s10107-009-0278-5.
- [31] J. Dunagan and S. Vempala (2008), ‘A simple polynomial-time rescaling algorithm for solving linear programs’, *Math. Program.* **114** (1, Ser. A), 101–114.
- [32] C. Eckart and G. Young (1936), ‘The approximation of one matrix by another of lower rank’, *Psychometrika* **1**, 211–218.
- [33] A. Edelman (1988), ‘Eigenvalues and condition numbers of random matrices’, *SIAM J. of Matrix Anal. and Applic.* **9**, 543–556.
- [34] A. Edelman (1992), ‘On the distribution of a scaled condition number’, *Math. Comp.* **58**, 185–190.
- [35] R.M. Freund and J.R. Vera (1999), ‘Condition-based complexity of convex optimization in conic linear form via the ellipsoid algorithm’, *SIAM J. Optim.* **10**(1), 155–176.
- [36] S. Geman (1980), ‘A limit theorem for the norm of random matrices’, *Ann. Probab.* **8**(2), 252–261.

- [37] E.N. Gilbert (1965), ‘The probability of covering a sphere with N circular caps’, *Biometrika* **52**, 323–330.
- [38] J.-L. Goffin (1980), ‘The relaxation method for solving systems of linear inequalities’, *Math. Oper. Res.* **5**(3), 388–414.
- [39] S. Janson (1986), ‘Random coverings in several dimensions’, *Acta Math.* **156**, 83–118.
- [40] E. Kostlan (1988), ‘Complexity theory of numerical linear algebra’, *J. Comput. Appl. Math.* **22**, 219–230.
- [41] L.G. Khachiyan (1979), ‘A polynomial time algorithm in linear programming’, *Dokl. Akad. Nauk. SSSR* **244**, 1093–1096.
- [42] G. Malajovich and J.M. Rojas (2004), ‘High probability analysis of the condition number of sparse polynomial systems’, *Theoret. Comput. Sci.* **315** (2-3), 524–555.
- [43] R.E. Miles (1969), ‘The asymptotic values of certain coverage probabilities’, *Biometrika*, **56**, 661–680.
- [44] Y. Nesterov and A. Nemirovskii (1994), *Interior-point polynomial algorithms in convex programming*, volume 13 of *SIAM Studies in Applied Mathematics*. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA.
- [45] Y. Nesterov and M. Todd and Y. Ye (1999), ‘Infeasible-start primal-dual methods and infeasibility detectors for nonlinear programming problems’, *Math. Program.* **84** (2, Ser. A), 227–267.
- [46] J. von Neumann and H. H. Goldstine (1947), ‘Numerical inverting of matrices of high order’, *Bull. Amer. Math. Soc.* **53**, 1021–1099.
- [47] J. Renegar (1987), ‘On the efficiency of Newton’s method in approximating all zeros of systems of complex polynomials’, *Math. of Oper. Research* **12**, 121–148.
- [48] J. Renegar (1989), ‘On the worst-case arithmetic complexity of approximating zeros of systems of polynomials’, *SIAM J. Comput.* **18**, 350–370.
- [49] J. Renegar (1994), ‘Some perturbation theory for linear programming’, *Math. Program.* **65** (1, Ser. A), 73–91.
- [50] J. Renegar (1995a), ‘Incorporating condition measures into the complexity theory of linear programming’, *SIAM J. Optim.* **5**(3), 506–524.
- [51] J. Renegar (1995b), ‘Linear programming, complexity theory and elementary functional analysis’, *Math. Program.* **70**(3, Ser. A), 279–351.
- [52] R.T. Rockafellar (1970), *Convex Analysis*, Princeton University Press.
- [53] A. Sankar, D. A. Spielman, and S.-H. Teng (2006), ‘Smoothed analysis of the condition numbers and growth factors of matrices’, *SIAM J. Matrix Anal. Appl.* **28**(2), 446–476.
- [54] M. Shub (1993), ‘Some remarks on Bézout’s theorem and complexity theory’, In *From Topology to Computation: Proceedings of the Smalefest (Berkeley, CA, 1990)*, pages 443–455. Springer, New York.
- [55] M. Shub (2009), ‘Complexity of Bézout’s theorem VI: Geodesics in the condition (number) metric’, *Found. Comput. Math.* **9**(2), 171–178.
- [56] M. Shub and S. Smale (1993a), ‘Complexity of Bézout’s theorem. I. Geometric aspects’, *J. Amer. Math. Soc.* **6**(2), 459–501.

- [57] M. Shub and S. Smale (1993b), ‘Complexity of Bézout’s theorem II: volumes and probabilities’, in *Computational Algebraic Geometry*, F. Eyssette and A. Galligo, editors, volume 109 of *Progress in Mathematics*, pages 267–285, Birkhäuser.
- [58] M. Shub and S. Smale (1993), ‘Complexity of Bézout’s theorem III: condition number and packing’, *J. Complexity* **9**, 4–14.
- [59] M. Shub and S. Smale (1994), ‘Complexity of Bézout’s theorem V: polynomial time’, *Theoretical Computer Science* **133**, 141–164.
- [60] M. Shub and S. Smale (1996), ‘Complexity of Bézout’s theorem IV: probability of success; extensions’, *SIAM J. of Numer. Anal.* **33**, 128–148.
- [61] J.W. Silverstein (1985), ‘The smallest eigenvalue of a large-dimensional Wishart matrix’, *Ann. Probab.* **13**(4), 1364–1368.
- [62] S. Smale (1981), ‘The fundamental theorem of algebra and complexity theory’, *Bull. Amer. Math. Soc.* **4**, 1–36.
- [63] S. Smale (1983), ‘On the average number of steps of the simplex method of linear programming’, *Math. Programming* **27**(3), 241–262.
- [64] S. Smale (1986), ‘Newton’s method estimates from data at one point’, in *The merging of disciplines: new directions in pure, applied, and computational mathematics (Laramie, Wyo., 1985)*, pages 185–196. Springer, New York, 1986.
- [65] S. Smale (1997), ‘Complexity theory and numerical analysis’, In A. Iserles, editor, *Acta Numerica*, pages 523–551. Cambridge University Press.
- [66] S. Smale (2000), ‘Mathematical problems for the next century’, In *Mathematics: frontiers and perspectives*, pages 271–294, Amer. Math. Soc., Providence, RI.
- [67] D. A. Spielman and S.-H. Teng (2001), ‘Smoothed analysis of algorithms: why the simplex algorithm usually takes polynomial time’, *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*, pages 296–305, New York, ACM.
- [68] D. A. Spielman and S.-H. Teng (2002), ‘Smoothed analysis of algorithms’, *Proceedings of the International Congress of Mathematicians*, volume I, pages 597–606.
- [69] D. A. Spielman and S.-H. Teng (2004), ‘Smoothed analysis: Why the simplex algorithm usually takes polynomial time’, *Journal of the ACM* **51**(3), 385–463.
- [70] T. Tao and V. Vu (2007), ‘The condition number of a randomly perturbed matrix’, *Proceedings 39th annual ACM Symposium on Theory of computing*, pages 248–255.
- [71] A. M. Turing (1948), ‘Rounding-off errors in matrix processes’, *Quart. J. Mech. Appl. Math.* **1**, 287–308.
- [72] S. A. Vavasis and Y. Ye (1995), ‘Condition numbers for polyhedra with real number data’, *Oper. Res. Lett.* **17**(5), 209–214.
- [73] J. C. Vera, J. C. Rivera, J. Peña, and Y. Hui (2007), ‘A primal-dual symmetric relaxation for homogeneous conic systems’, *J. Complexity* **23**, 245–261.
- [74] J. G. Wendel (1962), ‘A problem in geometric probability’, *Math. Scand.* **11**, 109–111.
- [75] H. Weyl (1939), ‘On the Volume of Tubes’, *Amer. J. Math.* **61**(2), 461–472.

-
- [76] J. H. Wilkinson (1963), *Rounding errors in algebraic processes*, Prentice-Hall Inc., Englewood Cliffs, N.J.
- [77] H. Wolkowicz, R. Saigal, and L. Vandenberghe (2000), *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, Kluwer Academic Publishers.
- [78] H. Woźniakowski (1977), ‘Numerical stability for solving nonlinear equations’, *Numer. Math.* **27**(4), 373–390.
- [79] M. Wschebor (2004), ‘Smoothed analysis of $\kappa(a)$ ’, *J. of Complexity* **20**, 97–107.

Privacy Against Many Arbitrary Low-sensitivity Queries

Cynthia Dwork*

Abstract

We consider privacy-preserving data analysis, in which a trusted curator, holding an n -row database filled with personal information, is presented with a large set \mathcal{Q} of queries about the database. Each query is a function, mapping the database to a real number. The curator's task is to return relatively accurate responses to all queries, while simultaneously protecting the privacy of the individual database rows.

An active area of research on this topic seeks algorithms ensuring *differential privacy*, a powerful notion of privacy that protects against all possible linkage attacks and composes automatically and obliviously, in a manner whose worst-case behavior is easily understood. Highly accurate differentially private algorithms exist for many types of datamining tasks and analyses, beginning with *counting queries* of the form "How many rows in the database satisfy Property P ?" Accuracy must decrease as the number of queries grows. For the special case of counting queries known techniques permit distortion whose dependence on n and $|\mathcal{Q}|$ is $\Theta(n^{2/3} \log |\mathcal{Q}|)$ [1] or $\Theta(\sqrt{n} \log^2 |\mathcal{Q}|)$ [8]. This paper describes the first solution for large sets \mathcal{Q} of *arbitrary* queries for which the presence or absence of a single datum has small effect on the outcome.

Mathematics Subject Classification (2010). Primary 68Q99; Secondary 68P99.

Keywords. Privacy, private data analysis, differential privacy, boosting, learning theory

1. Introduction

Private data analysis is a topic of intense study in multiple disciplines, with relevance to analysis of medical, financial, educational, and social data. In this

*Joint work with Guy Rothblum and Salil Vadhan.

Microsoft Research, 1065 La Avenida, Mountain View, CA 94043, USA.
E-mail: dwork@microsoft.com.

paper, we focus on a common setting in the literature, in which a trusted curator, holding an n -row database filled with personal information, is presented with a large set \mathcal{Q} of queries about the database. The curator’s task is to return relatively accurate responses to all queries, while simultaneously protecting the privacy of the individual database rows.

An active area of research on this topic seeks algorithms ensuring ϵ -*differential privacy*, a powerful notion of privacy that protects against all possible linkage attacks and composes automatically and obliviously, in a manner whose worst-case behavior is easily understood. Roughly speaking, differential privacy says that probability of seeing any outcome of an analysis or sequence of analyses is essentially the same, independent of whether any individual joins, or refrains from joining, the database. The probability space is over random choices made by the curator, and “essentially the same” is made mathematically rigorous and is quantified by a parameter typically known as ϵ ; smaller ϵ means better privacy.

Definition 1. [3, 5] [ϵ -Differential Privacy] A randomized function \mathcal{K} gives ϵ -*differential privacy* if for all data sets D and D' differing on at most one row, and all $S \subseteq \text{Range}(\mathcal{K})$,

$$\Pr[\mathcal{K}(D) \in S] \leq \exp(\epsilon) \times \Pr[\mathcal{K}(D') \in S], \tag{1}$$

where the probability space in each case is over the coin flips of \mathcal{K} .

Highly accurate differentially private algorithms exist for many types of queries, beginning with *counting queries* of the form “How many rows in the database satisfy Property P ?” and including many datamining, learning, and statistical tasks, as well as the generation of synthetic data sets of various types, and several on-line tasks such as maintaining data structures and minimizing regret. This paper describes the first solution for large sets \mathcal{Q} of arbitrary (not necessarily counting) queries.

An important quantity in designing a differentially private algorithm for a query q is the *sensitivity* of the function to be computed, denoted Δq .

Definition 2. [5] [Sensitivity] For $q : \mathcal{D} \rightarrow \mathbf{R}^d$, the L_1 *sensitivity* of q is

$$\begin{aligned} \Delta f &= \max_{D, D'} \|q(D) - q(D')\|_1 \\ &= \max_{D, D'} \sum_{i=1}^d |q(D)_i - q(D')_i| \end{aligned} \tag{2}$$

for all D, D' differing in at most one row.

In particular, when $d = 1$ the sensitivity of q is the maximum difference in the values that the function q may take on a pair of databases that differ in only one row. This is the difference the curator must hide, and the principal

techniques involve adding appropriately generated noise to the output of the function.

Suppose for a moment that the adversary, posing questions to the curator, is extremely knowledgeable; for example, the adversary may know that the database is in the set $\{D, D'\}$, consisting of a pair of databases that differ in a single row (say, the data of 1 person, or the presence or absence of a single individual), and the adversary's goal is to determine which of D and D' is the actual database. Speaking intuitively, each noisy response to a query $q \in \mathcal{Q}$ will reveal just a small amount of statistical information about the database, but these amounts accumulate. It is thus important to scale the noise to the "complexity" of the entire query sequence \mathcal{Q} . In fact, we can think of \mathcal{Q} as a single query whose dimension is the sum, over all $q \in \mathcal{Q}$ of the dimension (arity) of q . In this way we can talk about the sensitivity of a query sequence \mathcal{Q} .

Dwork, McSherry, Nissim, and Smith defined differential privacy and showed that adding noise generated according to the Laplace distribution $\text{Lap}((\Delta\mathcal{Q})/\varepsilon)$, freshly generated for *each* of the components of the output, is sufficient to ensure differential privacy for the query set \mathcal{Q} [5, 3]¹. When \mathcal{Q} has low sensitivity this may well yield acceptable results. For example, when \mathcal{Q} consists of counting queries, the distortion for each query will be roughly on the order of $|\mathcal{Q}|/\varepsilon$, which may be perfectly acceptable for small $|\mathcal{Q}|$, say, $|\mathcal{Q}| \in O(n^{1/2})$ or even $|\mathcal{Q}| \in O(n^{1-c})$ for some $0 < c < 1$. Formally,

Theorem 1.1. [5] *For $q : \mathcal{D} \rightarrow \mathbf{R}^d$, the mechanism \mathcal{K} that adds independently generated noise with distribution $\text{Lap}(\Delta q/\varepsilon)$ to each of the d output terms enjoys ε -differential privacy.*

Differential privacy is a property of the mechanism, and it holds regardless of what any possible adversary might know about the database, members of the database, or the rest of the world. For this reason differentially private mechanisms are automatically immune to so-called *linkage attacks*. The literature and popular press abound with these attacks, in which supposedly "anonymized" records containing both "sensitive" and "insensitive" attributes are linked by matching sets of "insensitive" attributes with those of identified records in a second database containing only "insensitive" attributes.

It is immediate from Definition 1 that differentially private mechanisms compose automatically:

Theorem 1.2. [5] *Let \mathcal{K}_1 and \mathcal{K}_2 be, respectively, ε_1 and ε_2 differentially private mechanisms. Then their (parallel or sequential) composition is at worst $(\varepsilon_1 + \varepsilon_2)$ -differentially private.*

For this reason we say that differential privacy composes *obliviously and automatically*: the curators operating the two mechanisms do not need to coordinate in order to control the cumulative damage done by the two mechanisms.

¹The Laplace distribution with parameter b , denoted $\text{Lap}(b)$, has density function $P(z|b) = \frac{1}{2b} \exp(-|z|/b)$; its variance is $2b^2$.

Continuing with our example of counting queries, when $|Q| \in \Omega(n)$ Theorem 1.1 no longer gives meaningful results. However, if Q is of moderate size, say, for example, $|Q| \in O(n^{2-c})$, we can use an earlier technique due to Dinur, Dwork, and Nissim [2, 7]. In this approach, fresh binomial or Gaussian noise of sufficient variance is added to each count. This yields a slightly weaker notion of privacy. The weakness comes from two sources. First, a very unlucky choice in the random noise can lead to an outcome, say, a response to a specific counting query, that violates the differential privacy requirement that the ratio of the probabilities of seeing this response on adjacent databases² is at most e^ϵ . Second, rather than considering the worst-case privacy loss under composition, this technique considered *likely* loss under composition. More precisely, the argument examined the expected gain in confidence that an adversary would experience in trying to distinguish between two databases differing in a single row. These are called *evolution of confidence* arguments. Formally, this technique yields what is known as (ϵ, δ) -differential privacy (see Section 2).

These two sources of weakness have been blurred a bit in the literature. As we will show (Section 3.1), we can apply the evolution of confidence argument for *any* composition of differentially private mechanisms [8]. Thus, one can obtain essentially the same effect using Laplacian noise. The required distortion for this technique grows as roughly $O(\sqrt{|Q| \ln(1/\delta)})$, where δ is the probability of failure (the “unlikely” compositions mentioned above).

So far, nothing we have said is specific to counting queries, although we have used them as a running example. For very large sets of *counting queries only*, Blum, Liggett, and Roth allow estimation of the answers to all queries with error roughly (we are suppressing several parameters) $O(n^{2/3} VCDim(Q))$, where the second term denotes the Vapnik-Chervonenkis dimension of the class Q . This is at most $\log_2(Q)$ (we always assume Q is finite) [1]. This remarkable result produces a small *synthetic database*; to answer a query $q \in Q$, one simply runs q against the synthetic database and “scales up” the answer.

The Principal Contribution of this Paper. We describe a method, using very different techniques, of handling large numbers of *arbitrary* queries. We are particularly interested in low-sensitivity queries q , for example, queries q such that $\Delta q \leq \rho$ for some fixed ρ . Our errors will be on the order of $O(\rho\sqrt{n} \log^2 |Q|)$.³

1.1. Additional Related Work. Inspired by the results of Blum *et al.*, [1], Dwork, Naor, Reingold, Rothblum, and Vadhan investigated the computational complexity of producing synthetic databases [6]. One contribution was an alternative approach to handling large numbers of counting (or other

²*i.e.*, databases differing in at most one row

³These results are joint work with Guy Rothblum and Salil Vadhan, and represent a strengthening and generalization of the results in [8].

linear) queries. Their algorithm runs in time that is polynomial in the size of the set \mathcal{Q} of counting queries and the size of the universe \mathcal{X} of possible database rows (the algorithm in [1] requires superpolynomial time). The error bounds obtained in [6] had a less pleasant dependence on $|\mathcal{Q}|$ than those in [1], and the results described herein stem from the (successful) effort by Dwork, Rothblum, and Vadhan to improve this dependence [8].

2. Additional Definitions

A database is a set of *rows*. We have used the expression “differing in at most one row” without formally defining it. There are two natural definitions. In the first, the databases have the same number of rows and agree on all but one, but they differ in the last. We prefer the second:

Definition 3. Databases D and D' are said to be *adjacent*, or to *differ in a single row*, if one database is contained in the other, and the larger database has exactly one additional row.

The Laplace distribution with parameter b , denoted $\text{Lap}(b)$, has density function $\Pr(z|b) = \frac{1}{2b} \exp(-|z|/b)$; its variance is $2b^2$. Taking $b = \Delta q/\varepsilon$ we have that the density at z is proportional to $e^{-\varepsilon|z|}$. This distribution has highest density at 0 (good for accuracy), and for any z, z' such that $|z - z'| \leq \Delta q$ the density at z is at most e^ε times the density at z' . It is also symmetric about 0, and this is important. We cannot, for example, have a distribution that only yields non-negative noise. Otherwise the only databases on which a counting query could return a response of 0 would be those in which no row satisfies the query. Letting D be such a database, and letting $D' = D \cup \{r\}$ for some row r satisfying the query, the pair D, D' would violate ε -differential privacy. Finally, the distribution gets flatter as ε decreases. This is correct: smaller ε means better privacy, so the noise density should be less “peaked” at 0 and change more gradually as the magnitude of the noise increases.

The following definition is a natural relaxation of pure ε -differentially private. When δ is small, for example, negligible in the size of the database, the definition is still quite powerful.

Definition 4 ((ε, δ) -Differential Privacy). [4] A randomized function \mathcal{K} gives (ε, δ) -*differential privacy* if for all data sets D and D' differing on at most one row, and all $S \subseteq \text{Range}(\mathcal{K})$,

$$\Pr[\mathcal{K}(D) \in S] \leq \exp(\varepsilon) \times \Pr[\mathcal{K}(D') \in S] + \delta, \quad (3)$$

where the probability space in each case is over the coin flips of \mathcal{K} .

3. Three Essential Elements

As discussed in the Introduction, without loss of generality we will assume \mathcal{Q} contains only 1-dimensional queries.

We will adapt three results from the literature, creating three building blocks from which we will construct our privacy-preserving mechanism for answering many arbitrary low-sensitivity queries with relatively small error per query. The first result was discussed in the Introduction: modifying the evolution of confidence argument to handle differentially private distributions in general, and Laplacian noise in particular.

The second result is a minor adaptation of an argument in [6] relating to generalization bounds.

The third result is the principal contribution of [8]. This is a general method for answering many queries in a privacy-preserving fashion using an adaptation of *boosting*, a wildly successful technique from learning theory [11] that has developed into a field of its own. The novelty in [8] is in applying boosting to *queries* rather than to data points.

3.1. Small to Moderate Numbers of Counting Queries.

Theorem 3.1. *For $\varepsilon, \delta \in [0, 1]$, let z satisfy $\exp^{-z^2/2} < \delta$ and let ε' satisfy $z\sqrt{k}(\varepsilon' + (2\varepsilon')^2) + k(2(\varepsilon')^2) < \varepsilon$. The mechanism \mathcal{M} that adds independently generated noise distributed according to $\text{Lap}(\rho/\varepsilon')$ to each answer in a set of k sensitivity ρ queries is (ε, δ) -differentially private.*

Theorem 3.1 is an immediate corollary of a useful generalization of the evolution of confidence argument of Dinur, Dwork, and Nissim [2, 7], described next.

Theorem 3.2. [8] *Let \mathcal{M} be an ε -differentially private mechanism. Then for any k invocations of \mathcal{M} and any pair of adjacent databases D, D' and any potential sequence of events $\mathcal{E}_1, \dots, \mathcal{E}_k, \mathcal{E}_i \subseteq \text{Range}(\mathcal{M})$,*

$$\Pr \left[\left| \sum_{i=1}^k \ln \frac{\Pr[\mathcal{E}_i]}{\Pr'[\mathcal{E}_i]} \right| > z\sqrt{k}(\varepsilon + 2\varepsilon^2) + k(2\varepsilon^2) \right] < e^{-z^2/2}$$

Here, for any event \mathcal{E} , $\Pr[\mathcal{E}]$ and $\Pr'[\mathcal{E}]$ denote the probability of event \mathcal{E} when the database is D , respectively, D' . The probabilities are over the random choices made by the mechanism. The “outside” probability is over the coins of the mechanism when the database is D (or, equivalently, D').

Proof. As in [2, 7] the proof relies on Azuma’s inequality. For this we need a bound $A \geq |\ln(\Pr[\mathcal{E}]/\Pr'[\mathcal{E}])|$ for any event \mathcal{E} , as well as an upper bound on the expectation $B \geq E[|\ln(\Pr[\mathcal{E}]/\Pr'[\mathcal{E}])|]$. Given bounds A and B , we can apply Azuma’s inequality

$$\Pr \left[\left| \sum_{i=1}^k \ln \frac{\Pr[\mathcal{E}_i]}{\Pr'[\mathcal{E}_i]} \right| > z\sqrt{k}(A + B) + kB \right] < e^{-z^2/2}.$$

The bound $A \leq \varepsilon$ is immediate from the fact that \mathcal{M} is ε -differentially private. \square

The theorem follows from the next lemma, which shows that $B < 2\varepsilon^2$.

Lemma 3.3. [8]. *If two distributions P and Q are ε -differentially-private w.r.t. each other for $0 \leq \varepsilon \leq 1$, then:*

$$0 \leq E_{s \sim P} [\ln (P[s]/Q[s])] \leq 2\varepsilon^2$$

Proof. We want to bound the relative entropy (or KL divergence):

$$D(P||Q) = \int_{-\infty}^{\infty} p[s] \ln(p[s]/q[s]) ds$$

(for convenience, we use here the natural logarithm). We know that for any P and Q it is the case that $D(P||Q) \geq 0$ (via the log-sum inequality), and so it suffices to bound $D(P||Q) + D(Q||P)$. We get:

$$\begin{aligned} D(P||Q) &\leq D(P||Q) + D(Q||P) \\ &= \int [p(s) \cdot (\ln(p(s)/q(s)) + \ln(q(s)/p(s))) \\ &\quad + (q(s) - p(s)) \cdot \ln(q(s)/p(s))] ds \\ &\leq \int [0 + |(q(s) - p(s)) \cdot \varepsilon] ds \\ &= \varepsilon \cdot \int [\max\{p(s), q(s)\} - \min\{p(s), q(s)\}] ds \\ &\leq \varepsilon \cdot \int [(e^\varepsilon - 1) \cdot \min\{p(s), q(s)\}] ds \\ &\leq \varepsilon \cdot (e^\varepsilon - 1) \end{aligned}$$

For $0 \leq \varepsilon \leq 1$, since $e^\varepsilon \leq 1 + 2\varepsilon$, we get that $D(P||Q) \leq 2\varepsilon^2$ as claimed. \square

To obtain Theorem 3.1, we choose z so that $e^{-z^2/2} < \delta$ and use noise generated according to $\text{Lap}(b)$ for each counting query, where $b = \rho/\varepsilon'$, and

$$z\sqrt{k}(\varepsilon' + (2\varepsilon')^2) + k(2(\varepsilon')^2) < \varepsilon.$$

3.2. Generalization Bounds. We have a large number \mathcal{Q} of queries to be approximated, and, using the tools described so far, we cannot get reasonably accurate answers by directly applying the addition of noise techniques. Following [6], we will find a way of obtaining good approximations to the answers to *most* queries by constructing an object that is designed to give good approximations to the answers of a *randomly selected* subset $S \subset \mathcal{Q}$.

Assume our n -row database DB consist of d -bit rows, $d < n$. Speaking informally, we will argue that, if we choose a random subset of na queries $S \subset \mathcal{Q}$, for sufficiently large a , and find *any* database Y that “agrees” with DB on the answers to all $q \in S$, then S “agrees” with DB on a substantial fraction of *all* $q \in \mathcal{Q}$.

We have put “agrees” in quotes because, for privacy purposes, we will add some noise to the answers $q(DB)$ for $q \in S$, and search for a (synthetic) database Y such that $|q(Y) - [q(DB) + \text{noise}]|$ is small for all $q \in S$. Letting r_q be the response (answer plus noise) for query $q(DB)$, there may not exist a database Y such that $q(Y) = r_q$ for all $q \in S$, since even the original database DB will not satisfy this condition. We therefore allow some error, requiring only that for all $q \in S$ we have $|r_q - q(Y)| < \lambda$ for some appropriately chosen λ .

Formally, we say that a database Y λ -fits a set of responses to the queries in S if

$$\max_{q \in S} |q(Y) - r_q| \leq \lambda,$$

where r_q is the noisy response on query q . We will choose λ to be a small multiplicative factor larger than the expected difference $|q(DB) - r_q|$. To be specific, letting b denote the parameter used in generating the noisy response r_q , so that $r_q = q(DB) + \text{Lap}(b)$, we will set $\lambda = \kappa b$, where κ is a security parameter. In this way, if the number k of queries is not too large, then the probability that there exists *no* such Y is negligible in κ .

Also, if a query is selected more than once, we do not add fresh noise after the first time; we simply use the response obtained the first time.

Theorem 3.4. [6] *Let \mathcal{D} be an arbitrary distribution on \mathcal{Q} . For all $\beta \in (0, 1)$ and $\eta \in [0, 1/2)$, if $a > \frac{2(\log(1/\beta) + nd)}{n(1 - 2\eta)}$, and if we λ -fit a database of size n to a set $S \sim \mathcal{D}^{na}$, then with probability at least $1 - \beta$ (over choice of S) we have $|q(Y) - q(DB)| \leq \lambda$ for at least a $(1/2 + \eta)$ fraction of \mathcal{D} . In other words, with probability at least $1 - \beta$ over choice of S , the probability that for a randomly selected query $q \sim \mathcal{D}$ we have $|q(Y) - q(DB)| \leq \lambda$ is at least $(1/2 + \eta)$.*

The proof of this theorem is just the proof of a related statement in [6], generalized to arbitrary distributions.

Proof. Fix a randomly chosen set of queries $S \subset \mathcal{Q}$ chosen according to \mathcal{D}^{na} and responses $R = \{r_q = q(DB) + \text{Lap}(b) | q \in S\}$. Examine a potential n -row database Y that might λ -fit R . Note that Y is described by an nd -bit string. Let us say Y is *bad* if $|q(Y) - r_q| > \lambda$ for at least a $(\log(1/\beta) + nd)/(na)$ fraction of \mathcal{D} , meaning that $\Pr_{q \sim \mathcal{D}}[|q(Y) - r_q| > \lambda] \geq (\log(1/\beta) + nd)/na$.

In other words, Y is bad if there exists a set $Q_Y \subset \mathcal{Q}$ of fractional weight at least $(\log(1/\beta) + nd)/na$ such that $|q(Y) - r_q| > \lambda$ for $q \in Q_Y$. For such a Y , what is the probability that Y gives λ -accurate answers for *every* $q \in S$? This is exactly the probability that none of the queries in S is in Q_Y , or

$$(1 - (\log(1/\beta) + nd)/na)^{na} \leq e^{-(\log(1/\beta) + nd)} \leq \beta \cdot 2^{-nd}.$$

Taking a union bound over all 2^{nd} possible choices for Y , the probability that there exists an nd -bit string Y that is accurate on all the queries in S but inaccurate on a set of fractional weight $(\log(1/\beta) + nd)/na$ is at most β .

For $0 < \beta < 1$, to ensure that, with probability $1 - \beta$, there exists a synthetic database Y that answers well on a $(1/2 + \eta)$ fraction of \mathcal{Q} , it is sufficient to have $(\log(1/\beta) + nd)/na < 1/2 - \eta$, or

$$a > \frac{2(\log(1/\beta) + nd)}{n(1 - 2\eta)}. \quad \square$$

3.3. Boosting for Queries. The third element in our method for privately answering a large number of arbitrary low-sensitivity queries is a type of *boosting*, a powerful technique from the learning literature due to Schapire [11]. Roughly speaking, boosting is a general method for improving the accuracy of any given learning algorithm. This phenomenally successful approach is of intense practical interest, as well as great theoretical beauty; see [12] for an excellent overview.

A classical example of boosting is in learning to recognize e-mail spam. We assume the existence of a *base learner* that, given labeled examples of spam (labeled 1) and non-spam (labeled -1), outputs a heuristic; the goal is to run the base learner many times and then to (somehow) combine the heuristics to obtain a highly accurate *classifier* for distinguishing spam from non-spam.

To be a little more specific, let us assume a universe \mathcal{X} of items (for example, e-mail messages) together with a distribution \mathcal{D} on \mathcal{X} . In practice, \mathcal{X} may be a set of training examples and \mathcal{D} is, initially, the uniform distribution on \mathcal{X} . Boosting is an iterative approach, in which each iteration will have the following form.

1. **Sample $S \sim \mathcal{D}^k$.** A set of elements S is selected according to \mathcal{D}^k (the choice of k will be determined later).
2. **Feed S to Base Learner.** The set S is presented to the base learner. The formal requirement for the base learner is that, for some $\eta > 0$, the base learner produces an algorithm A that correctly classifies at least a $(1/2 + \eta)$ fraction of the mass of \mathcal{D} . We will say, informally, that “ A does well on at least $(1/2 + \eta)$ of \mathcal{D} ”.
3. **Termination Test.** Letting $\{A_1, A_2, \dots\}$ be the collection of algorithms produced in all the iterations so far, *combine* all these algorithms to create an algorithm \mathcal{A} and decide whether or not \mathcal{A} is sufficiently accurate. If so, then terminate, with output \mathcal{A} . Typically, the termination test consists of computing the error rate of \mathcal{A} on the training set and comparing this to a pre-determined threshold. Alternatively, the algorithm may run for a fixed number of rounds, in which case the termination test is trivial.

4. **Update \mathcal{D} .** If the termination condition is not satisfied, then *update* the distribution \mathcal{D} , increasing the weight of wrongly labeled points and decreasing the weight of correctly labeled points.

To specify a concrete boosting algorithm we must define the combining operation and the method for updating \mathcal{D} . To *build* a boosting algorithm, we must also construct a base learner.

To our knowledge, in all applications prior to the work of Dwork, Rothblum, and Vadhan [8], the universe \mathcal{X} has been of data items. Instead, Dwork *et al.* apply boosting to the set of *queries*, and the result of the boosting is a data structure that permits relatively accurate answers to each query, while ensuring (ϵ, δ) -differential privacy. Naturally, this is called *boosting for queries*.

Following the framework for boosting outlined above, we now take \mathcal{D} to be a distribution on \mathcal{Q} , initially the uniform distribution. Assume the existence of a base learner that, given a number k of queries, produces an object A (say, a data structure) from which it is possible to obtain relatively accurate answers for at least a $(1/2 + \eta)$ fraction of the mass of \mathcal{D} . The objects A are combined by taking the median: given A_1, \dots, A_T , the quantity $q(DB)$ is estimated by computing the approximate values for $q(DB)$ yielded by each of the A_i and selecting the median. The algorithm will run for a fixed number T of rounds (roughly $T \in O(\log(|Q|))$).

There are “standard” methods for updating \mathcal{D} , for example, increasing the weight of poorly handled elements (in our case, queries) by a factor of e and decreasing the weight of well handled elements by the same factor. However, we need to protect the privacy of the database rows, and a single database row can simultaneously affect whether or not many queries are well or poorly approximated. We therefore need to mitigate the effect of any database row. This is done by attenuating the re-weighting procedure. Instead of always using a fixed ratio either for increasing the weight (when the answer is “accurate”) or decreasing it (when it is not), we set separate thresholds for “accuracy” and “inaccuracy”. Queries for which the error is below or above these thresholds have their weight decreased or increased (respectively) by a fixed amount. For queries whose error lies between these two thresholds, we scale the weight increase or decrease to the distance from the midpoint of the “accurate” and “inaccurate” thresholds. The attenuated scaling reduces the effect of any individual on a the reweighting of any query. This is because any individual can only affect the true answer to a query, and thus also the accuracy of the base learner’s output, by a small amount.

The larger the gap between the “accurate” and “inaccurate” thresholds, the smaller the effect of each individual on a query’s weight can be. This means that larger gaps are better for privacy. For accuracy, however, large gaps are bad. If the inaccuracy threshold is large, we can only guarantee that queries for which the base sanitizer is very inaccurate have their weight increased during reweighting. This degrades the accuracy guarantee of the boosted sanitizer. The

accuracy guarantee of the boosted sanitizer is roughly equal to the “inaccuracy” threshold.

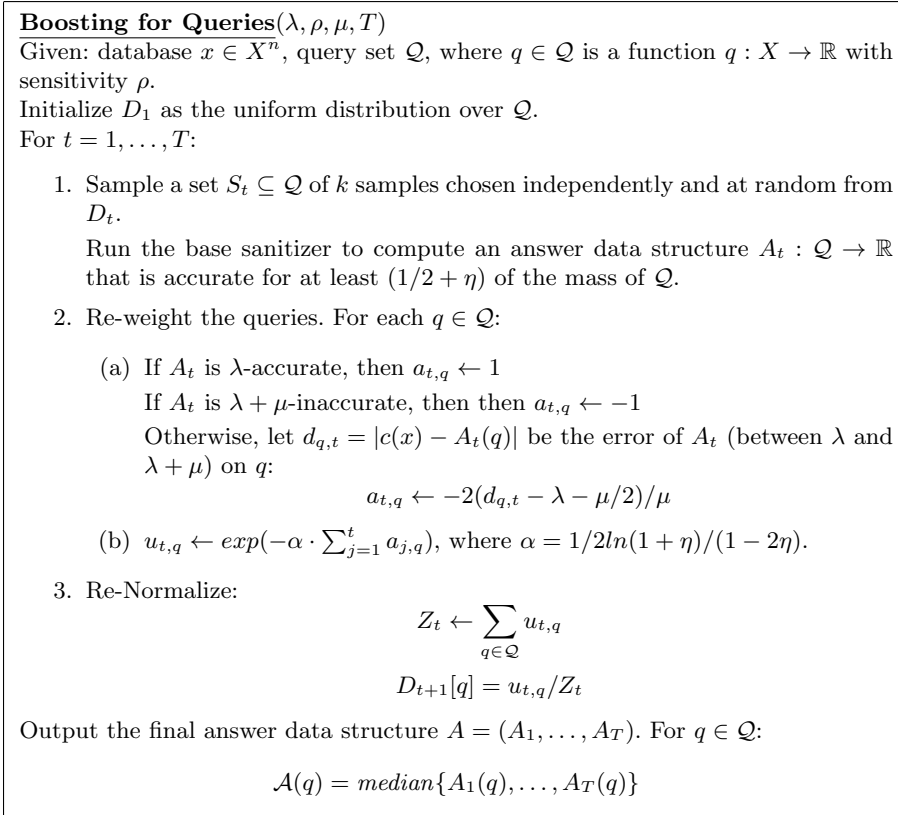


Figure 1. **Boosting for Queries** [8] (a variant of AdaBoost [13])

Theorem 3.5. [8] *Let \mathcal{Q} be a query family with sensitivity ρ . The algorithm of Figure 1 is a query-boosting algorithm. Assume that the base sanitizer, for any distribution on queries from \mathcal{Q} , on input k queries sampled from the distribution, outputs a data structure that gives λ -accurate answers at least a $1/2 + \eta$ fraction of the distribution’s mass (with all but $\exp(-\kappa)$ probability). Moreover, suppose the base sanitizer is $(\varepsilon_{\text{base}}, \delta_{\text{base}})$ -differentially private.*

The Boosting for Queries algorithm runs for $T = O(\log |\mathcal{Q}|/\eta^2)$ rounds. Its output is $((\varepsilon + T \cdot \varepsilon_{\text{base}}), T \cdot (\exp(-\kappa) + \delta_{\text{base}}))$ -differentially private. With all but $T \cdot \exp(-\kappa)$ probability, it gives $(\lambda + \mu)$ -accurate answers to all the queries in \mathcal{C} , where:

$$\mu = O\left(\log^2 |\mathcal{Q}| \cdot \rho \cdot \sqrt{k} \cdot \kappa / (\varepsilon \cdot \eta^5)\right)$$

4. Putting the Pieces Together

The idea for answering many low-sensitivity queries is now quite simple: we construct a base learner from Theorem 3.4 of Section 3.2 and then run the privacy-preserving Boosting for Queries algorithm.

In a little more detail, we set $k = na$, where a is as in the statement of Theorem 3.4. We will run the boosting algorithm for a fixed number $T = \log |Q|/\eta^2$ rounds.

In each round the base learner will choose a set S of k queries and compute noisy responses to these queries. By Theorem 3.1, in order to achieve $(\varepsilon_{base}, \delta_{base})$ -differential privacy for the base learner it is sufficient to use noise distributed according to $\text{Lap}(\rho/\varepsilon')$ to each answer, where z satisfies $e^{-z^2/2} < \delta_{base}$ and $z\sqrt{k}(\varepsilon' + (2\varepsilon')^2) + k(2(\varepsilon')^2) < \varepsilon_{base}$.

5. Conclusions and Future Work

These results are relevant to queries that may not be linear but nonetheless will have answers of sufficient size that the distortion added for privacy will not be too destructive. Is there a better base learner, one that will not lead to such large distortion?

At this point, no lower bounds on distortion are known for large numbers of counting queries, except the strong negative results of Dinur and Nissim showing that, to resist exponentially many counting queries, avoiding blatant non-privacy requires linear distortion for each query [2]. An exciting new direction for lower bounds was initiated by Hardt and Talwar [9]. It would be extremely interesting to extend their results to learn something about required distortion for the numbers and kinds (non-linear) of queries discussed in this paper.

A recent result of Roth and Roughgarden shows how to carry out the synthetic dataset generation of Blum, Liggett, and Roth in an on-line fashion, so that at all times there is a “current” synthetic database that (relatively) accurately answers the questions posed so far [10]. Can their techniques be applied to our data structures?

One of the major open areas in privacy-preserving data analysis is to better understand how differentially private mechanisms interact. The definition of differential privacy allows us to separate database utility from non-privacy. The database may teach that smoking causes cancer, and Smoker S is harmed because his insurance premiums have risen. But learning that smoking causes cancer is the whole point of a medical research database (Smoker S joins a smoking cessation program). Differential privacy resolves this “paradox” by ensuring that the risk of harm does not substantially increase *as a result of joining (or leaving) the database*. Thus, speaking informally, differential privacy bounds *incremental harm*. When the databases are created for a public good, the low

incremental harm guarantee provided by differential privacy may encourage participation.

When an individual participates in many differentially private databases, these increments can accumulate. Standard composition theorems bound the cumulative harm of participating in multiple differentially private databases. The evolution of confidence argument in Theorem 3.2 suggests that the composition of any k independent ε -differentially private mechanisms is “likely” to be “roughly” ($\sqrt{k}\varepsilon$, “unlikely”)-differentially private. This is in contrast to the known worst-case result that the composition of k ε -differentially private mechanisms is $k\varepsilon$ -differentially private. This should be made rigorous and the question of what it says about the “likely” effects of composition of (ε, δ) -differentially private should be examined.

Acknowledgement. The work described in this paper was done in collaboration with Guy Rothblum and Salil Vadhan. I am indebted to Guy Rothblum and Adam Smith for helpful discussions during the preparation of this manuscript.

References

- [1] A. Blum, K. Ligett, and A. Roth. A learning theory approach to non-interactive database privacy. In *Proceedings of the 40th ACM SIGACT Symposium on Theory of Computing*, 2008.
- [2] I. Dinur and K. Nissim. Revealing information while preserving privacy. In *Proceedings of the Twenty-Second ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*, pages 202–210, 2003.
- [3] C. Dwork. Differential privacy. In *Proceedings of the 33rd International Colloquium on Automata, Languages and Programming (ICALP)(2)*, pages 1–12, 2006.
- [4] C. Dwork, K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor. Our data, ourselves: privacy via distributed noise generation. In *Advances in Cryptology: Proceedings of EUROCRYPT*, pages 486–503, 2006.
- [5] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Theory of Cryptography Conference*, pages 265–284, 2006.
- [6] C. Dwork, M. Naor, O. Reingold, G. Rothblum, and S. Vadhan. When and how can privacy-preserving data release be done efficiently? In *Proceedings of the 2009 International ACM Symposium on Theory of Computing (STOC)*, 2009.
- [7] C. Dwork and K. Nissim. Privacy-preserving datamining on vertically partitioned databases. In *Proceedings of CRYPTO 2004*, volume 3152, pages 528–544, 2004.
- [8] C. Dwork, G. Rothblum, and S. Vadhan. Differential privacy and boosting, 2009. Manuscript.

-
- [9] M. Hardt and K. Talwar. On the geometry of differential privacy. In *to appear in STOC*, 2010.
 - [10] A. Roth and T. Roughgarden. The median mechanism: Interactive and efficient privacy with multiple queries. In *to appear in STOC*, 2010.
 - [11] R. Schapire. The strength of weak learnability. In *Proceedings of the 30th Annual Symposium on Foundations of Computer Science*, pages 28–33, 1989.
 - [12] R. Schapire. The boosting approach to machine learning: An overview. In *D. D. Denison, M. H. Hansen, C. Holmes, B. Mallick, B. Yu, editors, Nonlinear Estimation and Classification*. Springer, 2003.
 - [13] R. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. *Machine Learning*, 39:297–336, 1999.

Bridging Shannon and Hamming: List Error-correction with Optimal Rate

Venkatesan Guruswami*

Abstract

Error-correcting codes tackle the fundamental problem of recovering from errors during data communication and storage. A basic issue in coding theory concerns the modeling of the channel noise. Shannon's theory models the channel as a stochastic process with a known probability law. Hamming suggested a combinatorial approach where the channel causes worst-case errors subject only to a limit on the number of errors. These two approaches share a lot of common tools, however in terms of quantitative results, the classical results for worst-case errors were much weaker.

We survey recent progress on list decoding, highlighting its power and generality as an avenue to construct codes resilient to worst-case errors with information rates similar to what is possible against probabilistic errors. In particular, we discuss recent explicit constructions of list-decodable codes with information-theoretically optimal redundancy that is arbitrarily close to the fraction of symbols that can be corrupted by worst-case errors.

Mathematics Subject Classification (2010). Primary 11T71; Secondary 94B35.

Keywords. Error-correction algorithms; Explicit constructions; Reed-Solomon codes; Algebraic-geometric codes; Shannon capacity; List decoding; Polynomial reconstruction.

1. Introduction

Error-correcting codes enable reliable storage and transmission of data by providing a way to detect and rectify the errors caused by intervening noise. Mathematically, a code can be specified by an *encoding* function $E : \mathcal{M} \rightarrow \Sigma^n$ that

*Supported in part by a David and Lucile Packard Fellowship and NSF CCF-0953155.
Computer Science Department, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
E-mail: guruswami@cmu.edu.

maps a *message* $m \in \mathcal{M}$ into a redundant string $E(m)$ (called *codeword*) over some alphabet Σ . The set of all codewords is a subset $C \subset \Sigma^n$ called the *code* (or sometimes codebook). Only a small fraction of all possible strings in Σ^n are valid codewords, and the redundancy built into codewords is judiciously chosen in order to enable *decoding* the message m even from a somewhat distorted version of the codeword $E(m)$.

At a high level, the goal is to correct many errors without introducing too much redundancy — these conflicting goals impose a fundamental trade-off, and the basic goals of coding theory are to understand this trade-off, and construct explicit codes and efficient decoding algorithms for operating close to this trade-off. A convenient normalized measure of the redundancy of a code is given by its *rate*, which is defined as $\frac{\log |\mathcal{M}|}{n \log |\Sigma|}$ (unless specified logarithms are to the base 2). The rate of a code measures the proportion of non-redundant information conveyed per bit of the codeword. The larger the rate, the less redundant the encoding. Often the message set \mathcal{M} is identified with strings over the code's alphabet Σ (say with Σ^k) in which the case the rate is simple k/n . An important special case is when Σ is a finite field \mathbb{F} and the encoding map is a linear transformation $\mathbb{F}^k \rightarrow \mathbb{F}^n$. Such codes are called *linear codes*.

1.1. Modeling errors: Shannon vs. Hamming. An important issue in coding theory is the modeling of errors caused by the intervening noisy “channel.” There are two broad approaches to this. Shannon’s approach, in his pioneering work [44] that led to the birth of information theory, was to model the channel as a stochastic process with a precisely defined probability law. A simple example over alphabet $\Sigma = \{0, 1\}$ is the binary symmetric channel where the channel flips each transmitted bit with probability ρ , *independently* of other bits. Shannon precisely characterized the largest rate (called “capacity”) at which reliable communication is possible on such channels. In one of the early uses of the probabilistic method, he showed the existence of codes with rates approaching capacity that (together with maximum likelihood decoding) achieved exponentially small probability of miscommunication. Shannon’s work did not give a method to explicitly construct good codes or design efficient error-correction algorithms. Further there was no crisply abstracted (or at least easy to reason about) criterion for when a code was good in this model.

A different approach, implicit in the roughly concurrent work of Hamming [32], is to model the channel by a *worst-case* or *adversarial* process that can corrupt the codeword arbitrarily, subject only to a limit on the total number of errors caused. Both the locations of the corrupted symbols and the actual errors are assumed to be worst-case. The Hamming approach is more combinatorial, and notions such as the minimum distance between codewords, and connections to sphere packing (of Hamming balls), emerge as criteria for construction of good codes. Techniques from many areas of mathematics including algebra, graph theory, combinatorics, number theory, and the theory of algebraic

function fields and curves, have been very successfully brought to bear on the code construction problem.

To contrast the Shannon and Hamming models, in the former the channel behavior is oblivious to the overall message or codeword being transmitted, with the channel action on the i 'th symbol only dependent on that symbol, and perhaps a fixed amount of state information held by the channel. In the Hamming model, the channel behavior can depend arbitrarily on the codeword, causing a maliciously chosen error pattern. While this might be a pessimistic viewpoint, even if the channel is not an adversary, assuming independent errors governed by a precise channel law might be too strong. Codes for the Hamming model obviate the need for a very precise noise model, and are robust against a wide range of channel behaviors.

Unfortunately, requiring that the code be able to correct *every* pattern of up to a certain fraction, say ρ , of errors (instead of *typical* patterns) poses stringent limits. In particular, in order to ensure that a codeword $c \in \Sigma^n$ will not be confused for a codeword $c' \neq c$ even after up to ρn errors distort c , c must have Hamming distance more than $2\rho n$ from every other codeword. Thus every pair of distinct codewords must differ in at least a fraction 2ρ of positions. This distance requirement limits the number of codewords one can pack, and thus the rate of the code. For example, for codes over $\Sigma = \{0, 1\}$, when the fraction ρ of worst-case errors exceeds $1/4$, the rate of communication must approach zero, whereas even close to a fraction $1/2$ of random errors can be corrected with positive rate of communication in Shannon's model. Thus the generality of the worst-case noise model apparently comes at the price of a significant loss in rate.

1.2. List decoding. A simple relaxation of the requirement on the decoder, however, enables bridging this gap between worst-case and random errors. In this model, called *list decoding*, the decoder, given as input a noisy received word $y \in \Sigma^n$, must output a list of all codewords that are within Hamming distance ρn from y (where ρ is a bound on the fraction of worst-case errors). Thus, the decoder outputs all possible codewords that could have been corrupted into y if at most ρn errors occurred. Of course for this to be useful the codewords must be "sparsely distributed" so that no ball of radius ρn contains too many codewords. Surprisingly, allowing a small list suffices to approach the Shannon limit while correcting *all* error patterns. In particular, when transmitting bits, there *exist* codes list-decodable up to a ρ fraction of errors with rate approaching $1 - h(\rho)$ (where $h(x) = -x \log x - (1 - x) \log(1 - x)$ is the binary entropy function), which is the capacity of the binary symmetric channel that flips bits randomly and independently with probability ρ . In particular, the rate is positive even for $\rho \rightarrow 1/2$.

Let us address the obvious question that might immediately arise in the reader's mind: Why is list decoding a meaningful model for recovery from worst-case errors? There are several compelling reasons, of which we mention a few.

It is true that there are worst-case patterns of ρ fraction of errors that can lead to received words that cannot be unambiguously decoded as soon as ρ equals half the minimum fractional distance between codewords. However, for many codes, including random codes and all codes over large alphabets (such as Reed-Solomon codes) [36, 42], it can be shown that for most patterns of almost *twice* as many errors, there will be a unique close-by codeword. Thus, the list decoder outputting more than one codeword is a rare event. Therefore, it effectively decodes unambiguously, but we have obviated the need for a precise probabilistic channel model, and the results are robust to a wide variety of noise processes, including complicated combinations of i.i.d and bursty errors!

Further, with list decoding, instead of citing a pathological error pattern as an excuse to not decode other correctable patterns of errors, the decoder *always* corrects more errors, outputting a small list in the worst-case. Even for the worst-case errors which cannot be unambiguously decoded, returning a small list of possibilities is certainly no worse than declaring a decoding failure, and possibly quite useful if there is some context related information that can be used to identify the correct codeword from the list. In this vein, we should mention that in the last decade or two, list decoding has found many surprising applications beyond the coding theory, in algorithms, computational complexity theory, and cryptography. In these applications, there is either an application-specific method for breaking ties amongst the codewords in the list, or having a small list is not an issue at all. Finally, the primitive of list decoding seems rather versatile in implying solutions in some other models for bridging between worst-case and random errors also; we briefly touch upon this aspect at the end of the survey in Section 7.

The notion of list decoding dates back to work in the late 1950s by Elias [8] and Wozencraft [54]. Existential results indicating the potential of list decoding to correct many more errors have also been around for a while [55]. However, list decoding was revived with an algorithmic focus only in works motivated by complexity theory, beginning with the works of Goldreich and Levin [14], Ar *et al.* [2], and Sudan [48]. Ultimately, the usefulness of list decoding comes from the fact that, after a long hiatus, efficient list decoding algorithms for some important codes have been discovered in the last 10-15 years. These have turned some of the existential results on list decoding into constructive ones, and even led to the explicit construction of efficiently list-decodable codes with rates approaching the optimal information-theoretic limit.

In this survey, we discuss some of the background on error-correction, including some of the existential results on the potential of list decoding, and then focus on recent progress on list decoding algorithms for algebraic codes. We also mention some of the key open questions in the subject.

Related surveys: Surveys covering related material include a longer survey on algorithmic results for list decoding [19] and a “Research Highlight” written for

a broad computer science audience [26]. For reasons of space, we do not discuss the several compelling applications of list decoding beyond coding theory in this survey. Some of these are surveyed in [49, 50] or [17, Chap. 12], but these do not include some recent applications such as [43, 31]. Various “pseudorandom” objects can be described and studied in a unified way via appropriate variants of list decoding, and these are surveyed in [52].

1.3. Organization. We warm-up in Section 2 with a discussion of a particularly simple noise model, namely erasures. Over large alphabets, Reed-Solomon codes, which play an important role in list decoding, give a simple and optimal solution to the erasure recovery problem. Over the binary alphabet, we will mention how list decoding enables communication even when the fraction of erased bits approaches the obvious information-theoretic limit, though such codes are not known explicitly.

We then turn in Section 3 to the more challenging problem of decoding from errors. We state the existence theorems (established by the probabilistic method) which show that via list decoding it is possible to achieve rate similar to the Shannon model.

We next turn to constructive results, which aim to realize the potential of list decoding with explicit codes and efficient algorithms. In Section 4, we discuss list decoding of Reed-Solomon (RS) codes, including the “method of multiplicities” which leads to a “soft-decision” decoding algorithm of practical importance. Next, in Section 5, we discuss recently discovered variants of Reed-Solomon codes, called *folded* RS codes, and how they can be list decoded up to the information-theoretic limit, achieving the optimal trade-off between rate and fraction of errors decoded. Using a powerful “list recovery” property of these codes, one can reduce the alphabet size to a constant, and also construct good binary list-decodable codes, which we will briefly touch upon in Section 6. Finally, in Section 7, we will mention some alternate approaches to list decoding for bridging between the worst-case and random error models, and indicate the versatility of list decoding in devising solutions in these models as well.

2. Decoding From Erasures

We begin the technical discussion by considering the rather benign *erasure channel*. For $\alpha \in (0, 1)$ and an alphabet Σ , the erasure channel $\text{Erase}_\alpha(\Sigma)$ erases an *arbitrary* subset of up to a fraction α of the symbols transmitted, leaving the rest unaltered. In other words, it distorts a codeword $c \in \Sigma^n$ to $y \in (\Sigma \cup \{?\})^n$ with $y_i \in \{c_i, ?\}$ for $1 \leq i \leq n$ and $y_i = \{?\}$ for at most αn locations i (these correspond to the erased symbols). Note the locations of the erasures are known, so the decoding problem is just to “interpolate” the erased

symbols.¹ Erasures are relevant to modeling packet loss on the internet, where the identity of the missing packets can be determined using header information.

2.1. Optimal erasure recovery over large alphabets. Over large alphabets Σ , an important and well-studied family of codes called Reed-Solomon (RS) codes, give an optimal solution to the erasure recovery problem.

Definition 1 (Reed-Solomon codes). For a finite field \mathbb{F} with $|\mathbb{F}| \geq n$, an n -tuple $S = (a_1, a_2, \dots, a_n)$ of n distinct elements of \mathbb{F} , and an integer $1 \leq k \leq n$, the Reed-Solomon code $\text{RS}_{\mathbb{F},S}[n, k]$ is the k -dimensional subspace of \mathbb{F}^n defined as

$$\text{RS}_{\mathbb{F},S}[n, k] = \{(p(a_1), p(a_2), \dots, p(a_n)) \mid p \in \mathbb{F}[X] \text{ is a polynomial of degree } < k\}.$$

In other words, the message $(m_0, \dots, m_{k-1}) \in \mathbb{F}^k$ is viewed as a polynomial $m_0 + m_1X + \dots + m_{k-1}X^{k-1} \in \mathbb{F}[X]$ and it is encoded by its evaluation at a_1, a_2, \dots, a_n .

Since two distinct polynomials of degree $k-1$ over \mathbb{F} can have equal evaluations on at most $k-1$ points, any two distinct codewords of the above RS code differ in at least $n-k+1$ locations. Equivalently, given the evaluations of an unknown message polynomial p at any subset of k locations a_i , the polynomial p is uniquely determined, and in fact can be efficiently recovered by polynomial interpolation. Thus, the RS code enables recovery from up to $n-k$ erasures, or a fraction $1-R$ of erasures where $R = k/n$ is the rate of the code.

This trade-off between number of erasures and rate of the code is optimal, since it is information-theoretically impossible to recover the k symbols of the message from a string with less than k non-erased symbols. Thus Reed-Solomon codes give a simple and optimal solution for recovering from erasures. However, the alphabet size of these codes is at least as large as the block length n .

2.2. Algebraic-geometric codes. A generalization of RS codes called algebraic-geometric (AG) codes [16] can approach the above optimal trade-off over a large but fixed alphabet size that is independent of the block length. An AG code over a finite field \mathbb{F} is based on algebraic function field K over \mathbb{F} (a function field is a finite field extension of $\mathbb{F}(X)$). The message space of such an AG code is a linear space $\mathcal{L} \subset K$ of functions with a bounded number, say $< \ell$, of poles, all confined to a single point P_∞ of the algebraic curve corresponding to K . A function $f \in \mathcal{L}$ is encoded by its evaluation at a set S of n \mathbb{F} -rational points of K different from P_∞ . Since a function with $< \ell$ poles has $< \ell$ zeroes, a function $f \in \mathcal{L}$ is uniquely determined by its evaluations at ℓ points. The AG code thus enables recovery from up to $n-\ell$ erasures. (Given the evaluations of

¹In a harsher noise model called the *deletion channel* some symbols are deleted and the location of the deleted symbols is not known; we will not discuss this channel here.

a basis for \mathcal{L} at points in S , interpolating a function $f \in \mathcal{L}$ from its evaluation at ℓ points can be efficiently done by solving a linear system.) An excellent treatment of AG codes can be found in Stichtenoth's book [47].

By the Riemann-Roch theorem, the dimension of \mathcal{L} is at least $\ell - g$ where $g = g(K)$ is the genus of K , and so the code has rate at least $(\ell - g)/n$. When $|\mathbb{F}| = q = r^2$ is a square, there are known constructions of function fields K which have many rational points $N(K)$ in comparison to their genus $g(K)$, attaining the so-called *Drinfeld-Vladut* bound, with $N(K)/g(K) \rightarrow \sqrt{q} - 1$ [51, 12]. Using these function fields, one can prove the following (see [45] for an efficient construction algorithm):

Theorem 1. *For $0 < R < 1$ and $\varepsilon > 0$, one can explicitly construct a family of linear codes of rate R over an alphabet of size $O(1/\varepsilon^2)$ that enable polynomial time recovery from a fraction $1 - R - \varepsilon$ erasures.*

A lower bound of $\Omega(1/\varepsilon)$ is known on the alphabet size necessary for such codes. The following is a fascinating and longstanding open question.

Open Problem 1. *Close (or reduce) the gap between the $\Omega(1/\varepsilon)$ lower bound and $O(1/\varepsilon^2)$ upper bound on the size of alphabet needed for codes of rate R that enable recovery from a fraction $(1 - R - \varepsilon)$ of erasures.*

There are also combinatorial approaches to proving the above theorem, based on certain expander graphs [1, 23], which additionally achieve *linear* complexity encoding/decoding algorithms, but these require an alphabet size of $\exp((1/\varepsilon)^{O(1)})$. Interestingly, for random linear codes over \mathbb{F}_q to have the erasure recovery property stated in Theorem 1, one needs $q \geq \exp(\Omega(1/\varepsilon))$. AG codes thus beat the bounds achieved by the probabilistic method, a pretty rare phenomenon in combinatorial constructions!

2.3. Binary codes and list decoding from erasures. We now turn to erasure recovery with codes over a fixed small alphabet, and specifically binary codes (with alphabet $\{0, 1\}$) for definiteness. This will be the first foray into the theme of this survey on the distinction between worst-case and random noise.

A simple argument shows that a non-trivial binary code (with rate bounded away from 0 for large block lengths n) must have two codewords that differ in at most $n/2$ codeword positions. This implies that there are patterns of $n/2$ erasures from which unambiguous recovery of the erased symbols is not possible. In other words, for erasure fractions $\alpha \geq 1/2$, the rate of communication must approach 0. Nevertheless, for a *random* noise model BEC_α (the binary erasure channel) where each bit is erased independently with probability $\alpha \in (0, 1)$, it is well known [7] that there are codes of rate R approaching "capacity" $1 - \alpha$ that enable erasure recovery with high probability (over

the random noise caused by BEC_α).² In fact, a random linear code of such rate has this property with high probability. Explicit codes achieving capacity of BEC_α with polynomial time encoding/decoding algorithms are also known based on Forney's idea of code concatenation [11] (see [18, Sec. 3] for a description) and ultra-efficient algorithms are known based on low-density parity check codes [40, Chap. 3].

A binary linear code C is given by a linear transformation that maps a column vector $x \in \mathbb{F}_2^k$ to $Mx \in \mathbb{F}_2^n$ for some matrix $M \in \mathbb{F}_2^{n \times k}$ whose columns span C . For $T \subseteq \{1, 2, \dots, n\}$, the message x can be uniquely recovered from $(Mx)|_T$ (i.e., the bits of Mx corresponding to locations in T) iff $M|_T$, the submatrix of M with rows indexed by T , has rank k . Thus the capacity theorem equivalently states that for each $R \in (0, 1)$ and $\varepsilon > 0$, there is a matrix $M \in \mathbb{F}_2^{n \times Rn}$ such that a $1 - o_n(1)$ fraction of its $(R + \varepsilon)n \times Rn$ submatrices have rank Rn . (Moreover, this is true for a random matrix w.h.p.)

It turns out that it is possible to ensure that the small fraction of exceptional submatrices also have rank *close* to k . The proof is by the probabilistic method.

Theorem 2. *For each $R \in (0, 1)$ and $\varepsilon > 0$ and all large enough n , there is a matrix $M \in \mathbb{F}_2^{n \times Rn}$ such that every $(R + \varepsilon)n \times Rn$ submatrix of M has rank at least $Rn - C/\varepsilon$ for some absolute constant $C < \infty$. Moreover, this is true for a random matrix with probability $1 - e^{-\Omega_\varepsilon R(n)}$.*

The linear code generated by the columns of such a matrix can be *list decoded* from every pattern of $(1 - R - \varepsilon)$ erasures, obtaining a list of at most $2^{O(1/\varepsilon)}$ codewords consistent with the non-erased bits. Note that the *list size* is independent of n , and allowing for such a list enables recovering from *worst-case* erasures, without a loss in rate compared to what is possible in the probabilistic BEC_α model.

One can show that a list size of $\Omega(1/\varepsilon)$ (or equivalently rank deficiency of $\log(1/\varepsilon)$) is necessary.

Open Problem 2. *Close this exponential gap between the lower and upper bounds on list size for erasure list-decodable linear codes as a function of the distance ε to the optimal trade-off.*

If one does not insist on linearity, binary codes of rate $(1 - \alpha - \varepsilon)$ attaining a list size of $O(1/\varepsilon)$ for recovering from a worst-case fraction α of erasures are known, see [17, Chap. 10] and the references therein. The following is another significant open question.

Open Problem 3. *Construct a matrix with properties guaranteed by Theorem 2 explicitly (deterministically in time polynomial in n). Finding such an*

²Again, this rate is optimal, since w.h.p. there will be $\approx \alpha n$ erasures, and one must be able recover all the Rn message bits from the remaining $\approx (1 - \alpha)n$ bits.

explicit matrix, even with a relaxed rank bound of $Rn - g(1/\varepsilon)$ for an arbitrary function $g(\cdot)$ (or even $Rn - g(1/\varepsilon) \log n$), is also open.

Such a construction would be very interesting as it would give explicit list-decodable codes for bridging between probabilistic and worst-case erasures.

3. List Decoding from Errors: Existential Results

We turn to the problem of recovering not just from missing information (as in the erasures model) but from erroneous information. For concreteness we focus on binary codes in the initial discussion, and we will mention related bounds for larger alphabets towards the end of this section.

3.1. Random errors. Perhaps the simplest model of random bit errors is the *binary symmetric channel* BSC_ρ which is a memoryless probabilistic process that flips each bit independently with probability ρ , where $0 < \rho < 1/2$. Shannon's famous noisy coding theorem identifies the *capacity* of this channel to be $1 - h(\rho)$, where $h : [0, 1] \rightarrow [0, 1]$ is the binary entropy function $h(x) = -x \log_2 x - (1 - x) \log_2 (1 - x)$. Specifically, for any $\varepsilon > 0$, there are codes $C \subseteq \{0, 1\}^n$ of rate $1 - h(\rho) - \varepsilon$ for which there is a deterministic decoding function that recovers c from $c + e$ with probability $1 - e^{-\Omega_{\varepsilon, \rho}(n)}$ over the choice of $e \in \{0, 1\}^n$ from the binomial distribution with mean ρn . This result can be viewed as a packing of "mostly-disjoint" Hamming spheres of radius $\approx \rho n$, each with volume (number of points) $\approx 2^{h(\rho)n}$, centered around the $2^{(1-h(\rho)-\varepsilon)n}$ codewords of C , such that most of the points in each of these spheres do not belong to any other sphere.

This view also immediately suggests the converse theorem, provable by a volume packing argument, that a rate exceeding $1 - h(\rho)$ is not possible for communication on BSC_ρ . For each $c \in C$ and an error vector e added by BSC_ρ , w.h.p. $c + e$ belongs to a region consisting of $\approx 2^{h(\rho)n}$ strings. For reliable recovery of c , the decoding function must map most of the strings in this region to c . This implies one cannot pack more than $\approx 2^{(1-h(\rho))n}$ codewords.

Shannon's theorem is proved via the probabilistic method and only guarantees the existence of codes of rate close to $1 - h(\rho)$ for reliable communication on BSC_ρ . Subsequent work on algebraic error-correction and concatenated codes gave a polynomial time construction of such codes together with a decoding algorithm [10], though the complexity bounds are of the form $n^{O(1)} \exp(O(1/\varepsilon))$ for achieving rate $1 - h(\rho) - \varepsilon$ and thus scale poorly with the gap to capacity (see also [18, Sec. 3]). Nevertheless, for each fixed $\varepsilon > 0$, one could say that (in theory) good codes of rate within ε of optimal have been constructed for the BSC_ρ .

3.2. Worst-case errors. The case for worst-case or adversarial errors is more challenging. Let us consider a worst-case channel, denoted ADV_ρ , which corrupts an adversarially chosen subset of up to ρn codeword bits (where n is the block length). Unambiguous recovery of the codeword c from $c + e$ for an arbitrary error vector e of Hamming weight at most ρn clearly requires that the Hamming ball of radius ρn around c is disjoint from similar Hamming balls centered at every other codeword. As discussed in the introduction, this stringent combinatorial packing requirement places strong upper bounds on the code's rate. In particular, for $\rho > 1/4$, the rate must approach 0, whereas for BSC_ρ , communication at positive rate was possible for every $\rho < 1/2$. Also, for $\rho < 1/4$, the rate for communication on ADV_ρ must be bounded away from the capacity $1 - h(\rho)$ of BSC_ρ .

While such a perfectly disjoint packing of nearly $2^{(1-h(\rho))n}$ Hamming balls of radius ρn in $\{0,1\}^n$ does not exist, it turns out that it is possible to pack $2^{(1-h(\rho)-\varepsilon)n}$ such Hamming balls such that no $O(1/\varepsilon)$ of them intersect at a point, for any $\varepsilon > 0$. In fact a random packing has such a property with high probability. This shows that in the model of *list decoding*, there is hope to construct codes of rate approaching $1 - h(\rho)$ to correct errors caused by ADV_ρ if the decoder is allowed output a small list of codewords in the worst-case. The formal statement follows.

Definition 2. For $0 < \rho < 1$ and an integer $\ell \geq 1$, a code $C \subseteq \Sigma^N$ is said to be (ρ, ℓ) -*list decodable* if for every $y \in \Sigma^n$, there are at most ℓ codewords $c \in C$ such that the Hamming distance between y and c is at most ρn . Equivalently, Hamming balls of radius ρn around the codewords cover each point in Σ^n at most ℓ times.

The following theorem states that even with a modest list size ℓ , one can get close to a rate of $1 - h(\rho)$ for (ρ, ℓ) -list decodable binary codes. The bound $1 - h(\rho)$ on rate is best possible, since the expected number of codewords of C in a ball of radius ρn around a random $y \in \{0,1\}^n$ is at least $\frac{|C|}{2^n} \cdot 2^{(h(\rho)-o(1))n}$ which grows super-polynomially in n if the rate of C exceeds $1 - h(\rho) + o(1)$.

Theorem 3. For $\rho \in (0, 1/2)$ and an integer $\ell \geq 1$, for all large enough n there is a (ρ, ℓ) -list decodable binary code of block length n and rate at least $1 - h(\rho) - 1/\ell$.

Proof. The proof is an application of the probabilistic method [55, 9], and we sketch the simple argument. Let $R = 1 - h(\rho) - 1/\ell$. Pick a code $C \subseteq \{0,1\}^n$ by uniformly and independently picking $M = 2^{Rn}$ codewords. Fix a center y and a subset S of $(\ell + 1)$ codewords of C . Since these codewords are independent, the probability that all of them land in the ball of radius ρn around y (which has volume at most $2^{h(\rho)n}$) is at most $\left(\frac{2^{h(\rho)n}}{2^n}\right)^{\ell+1}$. A union bound over all 2^n choices of y and at most $M^{\ell+1}$ choices of S shows that if $R = 1 - h(\rho) - 1/\ell$, the code fails to be (ρ, ℓ) -list-decodable with probability at most $2^{-\Omega(n)}$. \square

The above argument only guarantees an arbitrary, not necessarily linear, code. The existence of (ρ, ℓ) -list decodable binary *linear* codes with a similar rate is more complicated to show, because codewords in a linear code are only pairwise (and not ℓ -wise) independent. Such linear codes were shown to exist via an application of the semi-random method in [22] (this method only worked for binary codes and did not yield a high probability result). Very recently, it was shown that a random linear code meets the bound of Theorem 3 with high probability [21] (and the proof also worked over all finite fields).

The above theorem implies that it is in principle possible to communicate at rate close to $1 - h(\rho)$ even against adversarial errors, provided that in the worst-case we allow the decoder to output a small list which includes the correct codeword. Unfortunately, no explicit codes with the list-decodability property guaranteed by Theorem 3 are known (see Open Problem 5).

We now state the analogous bounds for codes over an alphabet Σ with q elements. For $\rho \in [0, 1 - 1/q]$, the capacity of a probabilistic q -ary symmetric channel that flips $a \in \Sigma$ to each $a' \neq a \in \Sigma$ with probability $\rho/(q - 1)$ equals $1 - h_q(\rho)$, where $h_q : [0, 1] \rightarrow [0, 1]$ is the q -ary entropy function $h_q(x) = x \log_q(q - 1) - x \log_q x - (1 - x) \log_q(1 - x) = x \log_q(q - 1) + \frac{h(x)}{\log_2 q}$. For worst-case errors that arbitrarily corrupt up to a ρ fraction of symbols (where $0 < \rho < 1 - 1/q$), a similar argument to Theorem 3 guarantees the existence of q -ary (ρ, ℓ) -list decodable codes of rate at least $1 - h_q(\rho) - \frac{1}{\ell}$, showing that using list decoding it is possible to match the rate achievable for the q -ary symmetric channel also for worst-case errors.

3.3. Large alphabets. The function h_q satisfies $h_q(\rho) \leq \rho + \frac{h(\rho)}{\log_2 q}$, so keeping $\rho \in (0, 1)$ fixed and letting q increase, one sees that for every $\varepsilon > 0$, $1 - h_q(\rho) \geq 1 - \rho - \varepsilon$ when $q \geq 2^{1/\varepsilon}$. Thus we have the following existential result for list decoding over large alphabets:

Theorem 4. *For every R , $0 < R < 1$, and ε , $0 < \varepsilon < 1 - R$, every $q \geq 2^{2/\varepsilon}$ and all large enough integers n , there is a q -ary code of block length n that is $(1 - R - \varepsilon, \frac{2}{\varepsilon})$ -list decodable.*

With a code of rate R , the largest error fraction ρ that one can hope to tolerate is $1 - R$. To see this, consider a benign channel that simply erases (or replaces with some fixed string) the last ρn symbols of the codeword. Clearly the only useful information available to the decoder is the first $(1 - \rho)n$ symbols of the codeword from which it must be able to recover the Rn message symbols. Thus we must have $1 - \rho \geq R$, or $\rho \leq 1 - R$.

The above theorem therefore says that with list decoding one can approach this simple information-theoretic limit, and error correct as long as the error fraction is even slightly below the fraction of redundant symbols built into the codewords. The challenge, of course, is to construct an *explicit* list-decodable

code with such guarantees, along with an efficient list error-correction algorithm. Such a construction was recently obtained by the author and Rudra [25] (see also [26]), albeit with a list size and decoding complexity of $n^{\Omega(1/\varepsilon)}$. The codes achieving this optimal trade-off between rate and fraction of errors corrected are a variant of Reed-Solomon codes called “folded Reed-Solomon codes.” To describe these codes and the ideas underlying their list decoding, we begin with a discussion of list decoding algorithms for RS codes. These algorithms are important for their own sake due to the ubiquity and practical applications of RS codes.

4. List Decoding of Reed-Solomon Codes

Consider the Reed-Solomon code $\text{RS}_{\mathbb{F},S}[n, k]$ from Definition 1. As discussed in Section 2.1, one can decode this code (uniquely) from $n - k$ erasures, which is optimal. We now turn to correcting errors in a codeword of this RS code.

4.1. Unique decoding RS codes. We begin with the case when the number of errors is small enough so that unambiguous recovery of the correct message is possible. Specifically, let $p \in \mathbb{F}[X]$ be a degree $k - 1$ polynomial, and let $y \in \mathbb{F}^n$ be such that $y_i = p(a_i)$ for all but τ values of $i \in \{1, 2, \dots, n\}$. (Thus, y is a corrupted version of the RS codeword encoding p , with at most τ errors.) Since distinct codewords of $\text{RS}_{\mathbb{F},S}[n, k]$ differ in at least $n - k + 1$ positions, when $\tau \leq (n - k)/2$, then p is the unique polynomial whose evaluations differ from y in at most τ positions. However, it is not obvious how to recover p *efficiently* given y . If we knew the locations of errors, i.e., the set $E = \{i \mid p(a_i) \neq y_i\}$, then we could recover p by polynomial interpolation using its values at a_i , $i \notin E$. There are too many (exponential in n) possibilities for the set E to simply try them all.

Back in 1960, even before polynomial running time was formalized as the notion underlying efficient algorithms, Peterson [39] described a polynomial time algorithm to solve the above problem! We now describe the idea behind a different algorithm, due to Welch and Berlekamp [53], following the elegant description by Gemmell and Sudan [13]. Since for each $i \notin E$, $y_i = p(a_i)$, the bivariate polynomial

$$P(X, Y) = (Y - p(X)) \prod_{i \in E} (X - a_i)$$

satisfies $P(a_i, y_i) = 0$ for $i = 1, 2, \dots, n$. If we could somehow compute $P(X, Y)$, then one can efficiently find its linear (in Y) factor $Y - p(X)$ and recover p . Note that $P(X, Y) = E_1(X)Y - N_1(X)$ for some polynomials $E_1, N_1 \in \mathbb{F}[X]$ with degrees at most τ and $\tau + k - 1$ respectively. Therefore, in an attempt

to find $P(X, Y)$, we interpolate a *nonzero* bivariate polynomial $Q(X, Y) = E_2(X)Y - N_2(X)$ satisfying:

1. $\deg(E_2) \leq \tau$ and $\deg(N_2) \leq \tau + k - 1$.
2. $Q(a_i, y_i) = 0$ for $i = 1, 2, \dots, n$.

This can be done by setting up a system of linear equations over \mathbb{F} with unknowns being the coefficients of $E_2(X)$ and $N_2(X)$, and n homogeneous linear constraints $Q(a_i, y_i) = 0$ in these unknowns. Solving this linear system can certainly be done in polynomial time, and also admits fast, practical methods.

There may be many polynomials $Q \in \mathbb{F}[X, Y]$ satisfying the above constraints, but one can prove that all of them must have $Y - p(X)$ as a factor. To prove this, consider the polynomial $R(X) = Q(X, p(X))$. Whenever $p(a_i) = y_i$, we have $R(a_i) = Q(a_i, p(a_i)) = Q(a_i, y_i) = 0$. So R has at least $(n - \tau)$ distinct roots. The degree of R is at most $\max\{\deg(N_2), \deg(E_2) + k - 1\} \leq \tau + k - 1$. Thus if $n - \tau > \tau + k - 1$, i.e., if $\tau \leq \frac{n-k}{2}$, R must be the zero polynomial. Recalling that $R(X) = Q(X, p(X))$, this means that $Y - p(X)$ is a factor of $Q(X, Y)$. Note that $p(X)$ can be efficiently computed as the ratio $\frac{N_2(X)}{E_2(X)}$.

This gives an efficient algorithm to correct up to $\frac{n-k}{2}$ errors, or a fraction $\frac{1-R}{2}$ of errors as a function of the rate R . This trade-off is the best possible if we insist on unique recovery of the correct codeword, as it is easy to see that any code of rate R must have two distinct codewords that differ in at most a fraction $1 - R$ of locations. Recall that with list decoding, Theorem 4 says that there exist codes for which it is possible to correct a factor two more errors.

4.2. Reed-Solomon list decoding. We now turn to list decoding Reed-Solomon codes beyond the fraction $(1 - R)/2$ of errors. Before turning to the algorithmic aspects, we pause to comment on a combinatorial fact: using the fact that any two distinct codewords of $\text{RS}_{\mathbb{F}, S}[n, k]$ differ in more than $n - k$ positions, it can be shown (via the so-called “Johnson bound,” see for instance [17, Chap. 3]) that for a number of errors $\tau \leq n - \sqrt{nk} = (1 - \sqrt{R})n$, the size of the list that needs to be output by the decoder is guaranteed to be small (at most $O(n^2)$). Whether one can prove a polynomial list size bound for RS codes for even larger τ or whether the list size must necessarily grow super-polynomially beyond the Johnson bound remains an interesting open question. Some partial results establishing combinatorial limitations of list decoding Reed-Solomon codes appear in [24, 3].

One of the key results in algorithmic list decoding is that RS codes can be list decoded up to the $1 - \sqrt{R}$ bound *efficiently* [48, 29]. By the AM-GM inequality, $1 - \sqrt{R} > (1 - R)/2$ for $R \in (0, 1)$, so this gives an improvement over the traditional bounds for every rate. Further, for $R \rightarrow 0$, one can decode when the fraction of errors approaches 100%, thus enabling meaningful recovery even when noise overwhelms the correct information. This qualitative aspect is at

the root of the several influential applications of list decoding in complexity theory and cryptography [49], [17, Chap. 12].

4.2.1. Sudan's algorithm. We begin with Sudan's elegant algorithm for list decoding up to the bound of $\approx 1 - \sqrt{2R}$ [48]. The approach is via bivariate polynomial interpolation, based on finding a nonzero polynomial $Q \in \mathbb{F}[X, Y]$ such that every degree $k - 1$ polynomial $p \in \mathbb{F}[X]$ that must be output will be among the monic linear factors $Y - p(X)$ of $Q(X, Y)$. The idea is that $Y - p(X)$ passes through at least $n - \tau$ points (a_i, y_i) with relatively low-degree. So if a relatively low-degree Q is interpolated through *all* the points, i.e., $Q(a_i, y_i) = 0$ for every $i \in \{1, 2, \dots, n\}$, then one might hope that $Y - p(X)$ will emerge as a factor. Of course, this would be impossible if there are too many such target polynomials $p(X)$, but we know that if τ is not too large, there can be at most a few such polynomials. (The argument that the algorithm works does *not* need this bound, and gives an algorithmic proof of the combinatorial bound.)

Lemma 5. *Suppose $Q \in \mathbb{F}[X, Y]$ has degree in X, Y at most d_X, d_Y and satisfies $Q(a_i, y_i) = 0$ for $i \in \{1, 2, \dots, n\}$. Let p be a degree $k - 1$ polynomial p such that $p(a_i) = y_i$ for at least $n - \tau$ values of $i \in \{1, 2, \dots, n\}$. If $\tau < n - d_X - (k - 1)d_Y$, then $Y - p(X)$ must be a factor of $Q(X, Y)$.*

Proof. Consider $R(X) = Q(X, p(X))$. It has degree at most $d_X + (k - 1)d_Y$. On the other hand, $R(a_i) = 0$ whenever $p(a_i) = y_i$, and thus it has $n - \tau$ distinct roots. If $n - \tau > d_X + (k - 1)d_Y$, we must have $R(X) = 0$, or equivalently $(Y - p(X)) \mid Q(X, Y)$. \square

How small a degree suffices for such a non-zero Q to exist? Note that once we guarantee the existence of Q , such a bivariate polynomial can be efficiently found by solving a system of n homogeneous linear equations in the coefficients of its monomials, with one such linear constraint per interpolation condition $Q(a_i, y_i) = 0$. For a non-zero Q to exist, it suffices if the number of monomials of Q exceeds n . With degree bounds d_Y and d_X for the degree of Q in Y and X , the number of monomials is $(d_X + 1)(d_Y + 1)$. If we take $d_Y = \ell$ to be the target list size, $d_X = \lfloor \frac{n}{\ell} \rfloor$ ensures $(d_X + 1)(d_Y + 1) > n$, and therefore the existence of the desired Q . The above lemma then implies that up to τ errors can be list decoded with lists of size ℓ , if $\tau \leq n - \left(\frac{n}{\ell} + k\ell\right)$. Taking $\ell = \sqrt{n/k}$ to maximize the upper bound on τ , we get an algorithm to list decode up to $n - 2\sqrt{kn}$ errors, or a fraction $1 - 2\sqrt{R}$ of errors.

What is relevant to Lemma 5 is the “ $(1, k - 1)$ -weighted total degree” $d_X + (k - 1)d_Y$ rather than the individual degrees d_X, d_Y . Optimizing for the number of monomials with bounded $(1, k - 1)$ -weighted total degree, one can improve the fraction of errors list decoded to $\approx 1 - \sqrt{2R}$.

One aspect we did not address is finding the factors $Y - p(X)$ of a bivariate polynomial $Q \in \mathbb{F}[X, Y]$. The task of bivariate polynomial factorization admits

polynomial time algorithms [33]. Since the actual task here is easier, it can also be solved by root finding over a suitable extension field, or via a simple randomized reduction to Berlekamp's univariate polynomial factorization algorithm (see [19, Sec. 4.5] for a discussion, or [41] for a more efficient method tailored to finding linear bivariate factors).

4.2.2. The method of multiplicities. We now discuss the work of the author and Sudan [29], which introduced the powerful, if somewhat mysterious, idea of using multiplicities during interpolation to obtain an improved algorithm to list decode RS codes of rate R up to a fraction $1 - \sqrt{R}$ of errors. Note that this matches the combinatorial Johnson bound mentioned earlier, and beyond this radius we do not know if the number of codewords is guaranteed to be polynomially bounded, which is an a priori requirement for efficient list decoding.

We do not have space to develop the rationale of using multiplicities via some illustrative examples, and we point the reader to [19, Chap. 4] or [17, Chap. 6] for such a description. Here we sketch the technical aspects of the algorithm and its analysis. Recall that the above algorithm involves two steps: an interpolation step that finds a nonzero polynomial $Q \in \mathbb{F}[X, Y]$, and a factorization/root-finding step where all polynomials $p \in \mathbb{F}[X]$ that must be output are found amongst the factors $Y - p(X)$ of $Q(X, Y)$. The second step will remain the same in the improved algorithm. In the first step, we will demand more from the polynomial Q , namely that it has a zero of multiplicity w at each (a_i, y_i) , where $w \geq 1$ is the "multiplicity parameter" that governs the performance of the algorithm. (The choice $w = 1$ corresponds to Sudan's algorithm.) This will require an increase in the degree of Q , since for each (a_i, y_i) , we will require that all Hasse derivatives of Q at (a_i, y_i) of order up to w vanish, which leads to $\binom{w+1}{2}$ homogeneous linear constraints per point. To ensure that $(d_X + 1)(d_Y + 1) > n \binom{w+1}{2}$, we can scale d_X, d_Y up by a factor of $\approx w/\sqrt{2}$.

However, this increase is more than compensated in the second step. In particular, the analog of Lemma 5 can conclude that $Q(X, p(X)) = 0$ assuming $w(n - \tau) > d_X + (k - 1)d_Y$, thanks to a factor w gain in number of roots at each a_i for which $p(a_i) = y_i$. Optimizing the choice of d_X, d_Y as before, one can obtain an algorithm for list decoding $\approx 1 - \sqrt{2R}$ errors. Further, optimizing the $(1, k - 1)$ -weighted total degree instead of the individual degrees d_X, d_Y , the fraction of errors list decoded improves to $1 - \sqrt{R}$. We record the main claim about Reed-Solomon list decoding below.

Theorem 6 ([29]). *For every Reed-Solomon code $\text{RS}_{\mathbb{F}, S}[n, k]$, there is a list decoding algorithm to correct up to $n - \sqrt{kn}$ errors that runs in time polynomial in $n, |\mathbb{F}|$ and outputs a list of at most $O(n^2)$ polynomials.³*

³In most common instantiations of RS codes, $|\mathbb{F}|$ grows polynomially in n , in fact $|\mathbb{F}| = n$ or $n + 1$. But over large fields, one can also have a randomized list decoding algorithm running in time polynomial in $n, \log |\mathbb{F}|$.

Remark 1. The above algebraic list decoding algorithms, including the one using multiplicities in the interpolation phase, can also be generalized to the family of algebraic-geometric codes (which were briefly described in Section 2.2) [29], [17, Chap. 6]. The algorithm runs in polynomial time given access to a polynomial amount of pre-processed information about the AG code and underlying function field. Further, the algorithm can be generalized to work in an abstract algebraic framework and decode “redundant residue codes” where the messages comprise elements of a ring of bounded size (as per some measure), and are encoded by their residue modulo a collection of ideals [27], [17, Chap. 7]. A number-theoretic example of such codes are Chinese Remainder codes, where an integer $m \in [0, B)$ is encoded by its residue modulo primes p_1, p_2, \dots, p_n for $B \ll \prod_{i=1}^n p_i$ [15].

Remark 2. The polynomial interpolation method and the method of multiplicities have been recently used with spectacular success to obtain near-tight results on the size of Kakeya sets over finite fields [5, 6].

4.2.3. Soft-decision decoding. The multiplicities based decoding has another benefit, which has received widespread attention from a practical standpoint. This is the ability to exploit “soft information” concerning the reliability of various symbols. When a field element, say the i 'th symbol $y_i \in \mathbb{F}$ of a RS codeword, is transmitted on a physical channel, it is “modulated” into some real signal $\Lambda_i \in \mathbb{R}^b$. The channel noise distorts the signal into Λ'_i . The receiver must then “demodulate” this signal into the most likely field element(s), giving a (typically small) set of field elements $\{\alpha_{i,1}, \dots, \alpha_{i,a}\} \subset \mathbb{F}$ each with an associated weight corresponding to a reliability/confidence estimate. (These confidence estimates are called the “soft” information.)

Multiplicities provide a way to encode this soft information during the interpolation, with larger multiplicities for pairs $(a_i, \alpha_{i,j})$ with a higher weight. By using multiplicities in proportion to the weights, one can prove the following general result about list decoding codewords with large “weighted agreement.”

Theorem 7 ([29]). *For the RS code $\text{RS}_{\mathbb{F},S}[n, k]$ and any $\varepsilon > 0$, there is a decoding algorithm with runtime polynomial in $n, |\mathbb{F}|, 1/\varepsilon$ with the following guarantee. Given as input a collection of non-negative rational weights $W_{i,\alpha}$ for $1 \leq i \leq n$ and $\alpha \in \mathbb{F}$, the algorithm finds a list of all codewords $c \in \text{RS}_{\mathbb{F},S}[n, k]$ satisfying*

$$\sum_{i=1}^n W_{i,c_i} \geq \sqrt{(k-1) \sum_{i,\alpha} W_{i,\alpha}^2} + \varepsilon \max_{i,\alpha} W_{i,\alpha}. \quad (1)$$

In the case when for each i , $W_{i,\alpha} = 1$ for $\alpha = y_i$ and 0 otherwise, the above gives the decoding guarantee of Theorem 6 for list decoding $y = (y_1, y_2, \dots, y_n) \in \mathbb{F}^n$. To put the above result to good use, we need a good

way to assign the weights $W_{i,\alpha}$ that capture the likelihood that the i 'th symbol y_i equals α . Koetter and Vardy [34] developed a “front end” that chooses weights that are optimal in a certain sense, based on channel observations and transition probabilities, and also made important complexity optimizations in the interpolation step. This yields a soft-decision decoding algorithm for RS codes that has led to good coding gains in practice. It is worth noting that even though list decoding was motivated by worst-case errors, the ideas have also led to important advances for decoding under probabilistic channels. Theorem 7 is also useful in decoding concatenated codes, where the symbols of a RS codeword are further encoded by a binary inner code. Here the weights are obtained by decoding the inner code and are carefully picked to have bounded ℓ_2 norm (for plugging into (1)); see [30], [17, Chap. 8] for further details on such uses of soft decoding of RS codes.

5. List Decoding with Optimal Rate: Folded RS Codes

For several years after the publication of [29], there was no improvement to the $1 - \sqrt{R}$ trade-off between fraction of list decoded errors and rate R . Also, as mentioned earlier we still do not know if this bound can be improved upon for decoding RS codes. However, some recent work has shown that for some variants of RS codes, one can in fact do better. We now sketch the key ideas in this line of work which culminated with a list-decodable code construction achieving the optimal $1 - R - \varepsilon$ trade-off for any desired $\varepsilon > 0$.

5.1. Encoding multiple polynomials. The $1/2$ in the exponent of the $1 - \sqrt{R}$ decoding radius for RS codes came as a result of bivariate interpolation: a polynomial $Q(X, Y)$ of $(1, k-1)$ -weighted degree D has $\approx \frac{D^2}{2k}$ monomials, which leads to a $D = O(\sqrt{kn}) = O(\sqrt{Rn})$ bound on the degree needed to interpolate through n points (and this in turn leads to a $(1 - O(\sqrt{R}))n$ bound for the number of errors corrected). If we had the flexibility to interpolate in 3 dimensions, then a polynomial $Q(X, Y, Z)$ with $(1, k-1, k-1)$ -weighted D has $\approx \frac{D^3}{6k^2}$ monomials, so $D = O(R^{2/3}n)$ suffices for the interpolating through n points $(a_i, y_i, z_i) \in \mathbb{F}^3$. One could perhaps wishfully hope that this can be the basis of an algorithm for list decoding a fraction $\approx 1 - R^{2/3}$ of errors, and more generally, by interpolating in $s+1$ dimensions, a fraction $\approx 1 - R^{s/(s+1)}$ of errors, which exceeds $1 - R - \varepsilon$ for s chosen large enough.

The RS codeword encoding a polynomial $p \in \mathbb{F}[X]$ could naturally be viewed as a subset $\{(a_i, p(a_i)) \mid i = 1, 2, \dots, n\} \subseteq \mathbb{F}^2$ on which a bivariate polynomial can be interpolated. To interpolate a trivariate polynomial, we need a set of points in \mathbb{F}^3 that correspond in some natural way to the (noisy) codeword. Consider the variant of Reed-Solomon codes with evaluation points

$S = \{a_1, \dots, a_n\} \subseteq \mathbb{F}$ where the message consists of an arbitrary *pair* of polynomials $p_1, p_2 \in \mathbb{F}[X]$, each of degree at most $k - 1$, which are encoded into a codeword $c \in (\mathbb{F}^2)^n$ (over alphabet $\mathbb{F} \times \mathbb{F}$) where $c_i = (p_1(a_i), p_2(a_i))$. Note that the rate R of this code is the same as that of the RS code, namely k/n . Now the received word to be decoded consists of $(y_i, z_i) \in \mathbb{F}^2$ for $1 \leq i \leq n$, and we can interpolate a nonzero $Q \in \mathbb{F}[X, Y, Z]$ of $(1, k - 1, k - 1)$ -weighted degree at most D (for a suitably chosen degree bound D) such that $Q(a_i, y_i, z_i) = 0$ for $i = 1, 2, \dots, n$.

By picking the parameter D appropriately and also enforcing multiplicities in the interpolation step, it is not hard to prove the following claim by proceeding along the lines of Lemma 5 and ensuing arguments in Section 4.2:

$$\begin{aligned} &\text{If } (p_1(a_i), p_2(a_i)) = (y_i, z_i) \text{ for at least } t \text{ values of } i, \text{ and} \\ &t \geq (1 + o(1)) \sqrt[3]{k^2 n} \approx R^{2/3} n, \text{ then } Q(X, p_1(X), p_2(X)) = 0 \end{aligned} \quad (2)$$

If we could efficiently determine all pairs of polynomials (f, g) of degree at most $k - 1$ that satisfy $Q(X, f(X), g(X)) = 0$ for a given low-degree $Q \in \mathbb{F}[X, Y, Z]$, we would be done with list decoding all the pairs (p_1, p_2) whose encoding agrees with the received word on $t \approx R^{2/3} n$ positions. Unfortunately, this algebraic task is impossible in general, as there can be exponentially many (at least $|\mathbb{F}|^k$) solutions (f, g) to $Q(X, f(X), g(X)) = 0$. (As a simple example, consider $Q(X, Y, Z) = Y - Z$; now (f, f) is a solution for every $f \in \mathbb{F}[X]$ of degree at most $k - 1$.)

5.2. Parvaresh-Vardy codes. In the above scheme, we could only obtain one algebraic relation between p_1, p_2 , whereas two algebraically independent relations are needed to pin down p_1, p_2 to a small number of possibilities. To circumvent this problem, Parvaresh and Vardy [38] put forth the ingenious idea of obtaining the extra algebraic relation essentially “as default,” by enforcing it as an *a priori* condition satisfied at the encoder. Specifically, instead of letting the two message polynomials (p_1, p_2) be independent and uncorrelated, they required them to satisfy an appropriate algebraic condition. For instance, as a concrete choice, one can insist that $p_2(X) = p_1(X)^h \pmod{E(X)}$ for some monic $E \in \mathbb{F}[X]$ of degree k that is irreducible over \mathbb{F} , and a suitable exponent h . In this case, the message of the new code is just a degree $k - 1$ polynomial p_1 (as with RS codes), but it is encoded by the evaluations of both p_1 , and $p_2 = p_1^h \pmod{E}$, at the points a_i .

Given a nonzero polynomial Q , one can now determine a list of all such messages p_1 for which $Q(X, p_1(X), p_2(X)) = 0$, by recovering the residue $\overline{p_1} = p_1(X) \pmod{E(X)}$ of p_1 in the extension field $K = \mathbb{F}[X]/(E(X))$. The key is that this residue $\overline{p_1}$ is a root of the low-degree polynomial $S \in K[T]$ defined by $S(T) = Q(X, T, T^h) \pmod{E(X)}$. (Some care is required to ensure that S is in fact nonzero; in particular the exponent h must be chosen to be larger than

the degree of Q in Y, Z , but these are just technicalities and ignored in this description.)

This transformation of the RS codes is, however, not for free, and costs heavily in terms of the rate. The rate of the Parvaresh-Vardy code is $\frac{k}{2n}$ — half that of the original RS code — since the encoding has twice as many symbols (evaluations of *both* p_1, p_2). Plugging into the Claim (2), Parvaresh and Vardy get codes of rate R list-decodable up to a fraction $1 - (2R)^{2/3}$ errors. For $R < 1/16$, this gives a small improvement over the $1 - \sqrt{R}$ bound for RS codes. The natural extension of this idea to encoding s polynomials and using $(s+1)$ -variate interpolation gives a decoding radius $1 - (sR)^{s/(s+1)}$, and (optimizing in s) a decoding radius $1 - O(R \log(1/R))$ for low rates $R \rightarrow 0$. Unfortunately, the improvement over RS codes is confined to low rates.

5.3. Folded RS codes. We now turn to the work of the author and Rudra on optimal rate list-decodable codes [25], obtained by a “folding” of the Reed-Solomon code. The big rate loss in the Parvaresh-Vardy code construction was due to the inclusion of the evaluations of a second polynomial p_2 along with those of the message polynomial p_1 . Note that the codes by construction can never have a rate exceeding $1/2$, whereas we would like list-decodable codes of rate R for any desired $R \in (0, 1)$.

5.3.1. Algebra behind folding. To motivate the approach, consider the code where instead of picking p_2 as above, we take $p_2(X) = p_1(-X)$ (let us assume the characteristic of \mathbb{F} is not 2; if the field has characteristic 2 we can take $p_2(X) = p_1(X + a)$ for some $a \in \mathbb{F}^*$). Also assume that $n = |\mathbb{F}^*|$ and the evaluation points are ordered so that $a_{2i} = -a_{2i-1}$ for $1 \leq i \leq n/2$. In the encoding of p_1 by the evaluations of (p_1, p_2) at $\{a_1, \dots, a_n\}$, note that $(p_1(a_{2i-1}), p_2(a_{2i-1})) = (p_1(a_{2i-1}), p_1(-a_{2i-1})) = (p_1(a_{2i-1}), p_1(a_{2i})) = (p_2(a_{2i}), p_1(a_{2i}))$. So the codeword symbols at locations $2i-1, 2i$ are the same up to the ordering within the pair. Therefore, we can compress the codeword with no loss in information, by deleting half of the evaluations $(p_1(a_{2i}), p_2(a_{2i}))$. This recovers the factor 2 lost in the Parvaresh-Vardy construction. In fact, note the resulting code is essentially just the RS code, but considered as a code over \mathbb{F}^2 of length $n/2$, by “bundling” together the evaluations at $\alpha, -\alpha$ together for each $\alpha \in \mathbb{F}$.

The argument described in Section 5.1 above shows that all polynomials p_1 differing from the received word on at most a fraction $1 - R^{2/3}$ of places must satisfy $Q(X, p_1(X), p_1(-X)) = 0$. So the question once again is how many such polynomials can exist, and whether one can find them all efficiently? Unfortunately, there could still be exponentially many such solutions. For instance, consider $Q(X, Y, Z) = Y - Z$. For every $f \in \mathbb{F}[X]$ of degree at most $(k-1)/2$, $p_1(X) = f(X^2)$ clearly satisfies $Q(X, p_1(X), p_1(-X)) = f(X^2) - f((-X)^2) = 0$.

The above example also shows that there can be $\approx |\mathbb{F}|^{k/r}$ polynomials $p_1 \in \mathbb{F}[X]$ of degree less than k satisfying $Q(X, p_1(X), p_1(\alpha X)) = 0$ for a nonzero α , if the order of α in the multiplicative group \mathbb{F}^* is r . Thus to get a reasonable (polynomial in $n, |\mathbb{F}|$) upper bound on the number of solutions, the multiplicative order of α must be large.

The following lemma shows that a form of converse also holds. It is the key algebraic fact underlying the list decoding algorithm for folded RS codes.

Lemma 8. *Let $Q \in \mathbb{F}_q[X, Y, Z]$ be a nonzero polynomial with degrees d_Y, d_Z in Y, Z less than q . Let γ be a primitive element in \mathbb{F}_q . Let $1 \leq k < q$. The number of degree $k - 1$ polynomials $f \in \mathbb{F}_q[X]$ such that $Q(X, f(X), f(\gamma X)) = 0$ is at most $d_Y + qd_Z < q^2$, and a list of all such f can be found in time polynomial in q .*

For the formal proof, see [25] or [19, Sec. 6.4]. The algebraic crux is the following identity for every polynomial $f \in \mathbb{F}_q[X]$:

$$f(\gamma X) \equiv f(X)^q \pmod{(X^{q-1} - \gamma)} . \tag{3}$$

Further, the polynomial $P(X) = X^{q-1} - \gamma$ is irreducible over \mathbb{F}_q . Therefore, the condition $Q(X, f(X), f(\gamma X)) = 0$ implies the equation $T(\bar{f}, f^q) = 0$ where $\bar{f} = f \pmod{P(X)}$ and $T(Y, Z) = Q(X, Y, Z) \pmod{P(X)}$ is a polynomial with coefficients from the extension field $\mathbb{F}_q[X]/(P(X))$. This implies that \bar{f} , and hence f if its degree is less than $q - 1$, can be found amongst the roots of the polynomial $T(Y, Y^q)$, which has degree at most $d_Y + qd_Z$, over the field $\mathbb{F}_q[X]/(P(X))$.

Analogously, the following generalization holds for higher order interpolation:

Lemma 9. *Let $Q \in \mathbb{F}_q[X, Y_1, Y_2, \dots, Y_s]$ be a nonzero polynomial with the degree in the Y_i 's less than q , and let γ be a primitive element in \mathbb{F}_q . Then there are at most q^s polynomials $f \in \mathbb{F}_q[X]$ of degree less than $(q - 1)$ satisfying*

$$Q(X, f(X), f(\gamma X), \dots, f(\gamma^{s-1} X)) = 0 .$$

Further, a list of all these polynomials can be found in $q^{O(s)}$ time.

5.3.2. Code description and main result. Unlike the code where we bundled together $f(\alpha), f(-\alpha)$ together, when γ is primitive (or has high order), one cannot bundle together the evaluations of f on an whole orbit of the action by multiplication by γ . The folded RS code proposed in [25] consists of bundling together consecutive m symbols of the RS codeword $(f(1), f(\gamma), \dots, f(\gamma^{n-1}))$ for some fixed integer constant m called the *folding parameter* (which we assume for convenience divides the block length $n = q - 1$). Formally,

Definition 3 (Folded Reed-Solomon Code). The m -folded version of the RS code $\text{RS}_{\mathbb{F}_q, \mathbb{F}_q^*}[n, k]$ is a code of block length $N = n/m$ over the alphabet \mathbb{F}_q^m .

The encoding of a polynomial $f \in \mathbb{F}_q[X]$ of degree at most $k-1$, has as its j 'th symbol, for $0 \leq j < n/m$, the m -tuple $(f(\gamma^{jm}), f(\gamma^{j(m+1)}), \dots, f(\gamma^{j(m+m-1)}))$.

The folded version of a RS code thus carries the same information, just “bundled” differently. It is a code of exactly the same rate as the original RS code, but is defined over a larger alphabet. At a high level, folding restricts the flexibility in the subset of evaluation points that an adversary can corrupt. We now state the main result concerning decoding these codes from [25].

Theorem 10. *For every positive integer m and integer s , $1 \leq s \leq m$, the folded RS code with the parameters q, n, N, k from Definition (3) can be list decoded up to a radius*

$$N - (1 + o(1)) \frac{(kn)^{1/(s+1)}}{m - s + 1}, \quad (4)$$

in time at most $q^{O(s)}$, and the list size output by the decoder will be at most q^s .

The parameter s corresponds to the number of dimensions in the interpolation, and the stated bound is obtained through $(s+1)$ -variate interpolation. Specifically, the decoding algorithm interpolates a low-degree nonzero polynomial $Q \in \mathbb{F}_q[X, Y_1, Y_2, \dots, Y_s]$ such that any message f whose folded RS encoding is within distance (4) from the received word must satisfy $Q(X, f(X), f(\gamma X), \dots, f(\gamma^{s-1} X))$. The list of all such degree $< k$ polynomials can be efficiently found by Lemma 9.

By picking m large enough compared to s , and noting that the rate $R = k/n$ and $n = Nm$, the fraction of decoded errors can be made larger than $1 - (1 + \zeta)R^{s/(s+1)}$ for any desired $\zeta > 0$. In the limit of large s (specifically, for $s = \Theta(\varepsilon^{-1} \log(1/R))$), the fraction of errors corrected approaches $1 - R$, leading to the main conclusion about optimal rate list-decodable codes.

Theorem 11. *For every $\varepsilon > 0$ and $0 < R < 1$, there is a family of folded Reed-Solomon codes which have rate at least R and which can be list decoded up to a fraction $1 - R - \varepsilon$ of errors in time $(N/\varepsilon^2)^{O(\varepsilon^{-1} \log(1/R))}$ where N is the block length of the code. The alphabet size of the code as a function of the block length N is $(N/\varepsilon^2)^{O(1/\varepsilon^2)}$.*

Remark 3. The large alphabet size and decoding complexity in the above result are a shortcoming. Fortunately, the alphabet size can be reduced to a constant depending only on ε , and in fact with a dependence that is not far from optimal. We will sketch this in the next section.

Open Problem 4. *Can one improve the decoding complexity (for list decoding up to a fraction $1 - R - \varepsilon$ of errors) to have a better dependence on $1/\varepsilon$,*

and/or guarantee that the list size will be bounded by a constant independent of n ?

Recall that by the existential result in Theorem 4, a list size of $O(1/\varepsilon)$ suffices.

Remark 4 (Connection to Artin-Frobenius automorphisms). The algebraic crux of the folded RS decoder, identity (3), was that the automorphism Γ of the function field $K = \mathbb{F}_q(X)$ induced by $X \mapsto \gamma X$ satisfied the identity $\Gamma(f) \equiv f^q \pmod{P(X)}$ for all polynomials f , where $P(X) = X^{q-1} - \gamma$. That is, the automorphism Γ induces a low-degree map w.r.t the evaluations of polynomials at the “place” corresponding to $P(X)$. In general, one can obtain such low-degree relations between residues via the Frobenius automorphisms of places in Galois extensions. Specifically, if a place P of a field K is inert in a finite Galois extension L/K with a place P' in L above it, then the Frobenius automorphism Frob_P satisfies $\text{Frob}_P(x) = x^{\|P\|} \pmod{P'}$ for every $x \in L$ that is regular at P' , where $\|P\|$ is the size of the residue field at P . Using this approach, we were able to extend the folded RS code construction to folded versions of certain algebraic-geometric codes based on cyclotomic function fields [20].

6. List-decodable Codes Over Smaller Alphabets

6.1. Binary codes. Let us now discuss results on constructing binary list-decodable codes. The existential result of Theorem 3 says that there exist $(\rho, O(1/\varepsilon))$ -list decodable codes of rate $1 - h(\rho) - \varepsilon$, for any desired error fraction $\rho \in (0, 1/2)$. However, unlike large alphabets, an explicit construction with rate approaching the optimal bound of $1 - h(\rho)$ is not known, and the following is a *major* open question.

Open Problem 5. Fix $\rho \in (0, 1/2)$. Can one give an explicit construction of binary codes of rate $1 - h(\rho) - \varepsilon$ for any desired constant $\varepsilon > 0$ that are (ρ, ℓ) -list decodable, even for a list-size ℓ that grows as a polynomially bounded function $\ell(n) = n^{O_\varepsilon(1)}$ of the block length n of the code?

Can one construct such codes together with an efficient list decoding algorithm that on input $y \in \{0, 1\}^n$ outputs the list of codewords within Hamming distance ρn of y in polynomial time.

Though the above challenge is wide open, one can construct binary list-decodable codes achieving a reasonable trade-off, the so-called Zyablov bound, between the error fraction ρ and rate R , for any $\rho \in (0, 1/2)$. The bound below appears somewhat complicated, and a useful regime to compare it with the

optimal rate of $1 - h(\rho)$ is when $\rho = 1/2 - \gamma$ for $\gamma \rightarrow 0$: in this case $1 - h(\rho) \approx \gamma^2$ whereas the Zyablov bound is only $\approx \gamma^3$.

Theorem 12 ([25]). *For every $\rho \in (0, 1/2)$ and $\varepsilon > 0$, there is a polynomial time constructible family of binary codes of rate at least*

$$\max_{\substack{\rho_1, \rho_2 = \rho \\ \rho_1 < 1, \rho_2 < 1/2}} (1 - \rho_1)(1 - h(\rho_2)) - \varepsilon$$

that can be list decoded in polynomial time up to a fraction ρ of errors.

The idea behind the above claim is *code concatenation*, where we first encode the message as per an “outer” code over a large alphabet, and then each symbol is further encoded by an “inner” binary code which has the optimal trade-off between rate and list-decodability. The inner code is of small enough length that one can find such a good code essentially by brute-force in time polynomial in the overall code length.

Theorem 12 is proved by concatenating the folded RS codes guaranteed by Theorem 11 of rate $\approx 1 - \rho_1$ with inner codes of rate $\approx 1 - h(\rho_2)$ that are $(\rho_2, \ell = O_\varepsilon(1))$ -list decodable (as guaranteed to exist by Theorem 3). The idea behind the decoding is to first list decode the various inner blocks up to fractional radius ρ_2 . This gives a list of size at most ℓ for the possible symbols at each position of the outer codeword. Every folded RS codeword that must be output can have at most a fraction ρ_1 of symbols which do not belong to the respective lists. Now while the folded RS code of rate $\approx 1 - \rho_1$ can recover from a fraction ρ_1 of errors, we now have a harder problem, as for each position we do not have a unique symbol but a list of ℓ possible symbols. Somewhat remarkably, folded RS codes can handle such bounded size lists with no loss in rate! (The alphabet size will depend on ℓ ; see [25] for exact details.) This extension to list decoding is called “list recovery” and is a powerful primitive that is useful in composing list-decodable codes together.

6.2. Optimal rate list-decodable codes over fixed alphabets. The powerful list recovery properties offered by folded RS codes, together with techniques based on expander graphs to redistribute symbols of codewords, can also be used to attain the optimal trade-off between error-correction radius and rate of Theorem 11 over an alphabet of fixed size depending only on ε . For details, see [25] or [19, Chap. 7].

Theorem 13. *For every $R \in (0, 1)$ and $\varepsilon > 0$, there is a polynomial time constructible family of codes over an alphabet of size $2^{O(\varepsilon^{-4} \log(1/\varepsilon))}$ that have rate at least R and can be list decoded up to a fraction $(1 - R - \varepsilon)$ of errors in polynomial time.*

Remark 5. $2^{\Omega(1/\varepsilon)}$ is a lower bound on the alphabet size needed for list decoding up to a fraction $1 - R - \varepsilon$ of errors with rate R , so the above alphabet size is in the right ballpark.

7. Alternate Bridges Between Worst-case and Random Errors

We conclude this survey with a brief discussion of some other models besides list decoding that can be used to handle worst-case errors with rates similar to random errors. One of these is to allow randomized coding strategies where the sender and receiver share secret randomness (hidden from the channel) that is used to pick a coding scheme at random from a family of codes. Using such strategies, one can achieve a rate approaching $1 - h(\rho)$ for communicating on the adversarial channel ADV_ρ (for example, by randomly permuting the symbols and adding a random offset to codes achieving capacity on BSC_ρ).

The amount of shared randomness in the above setting can be reduced if we make computational assumptions on the channel [35] — the encoder and decoder only need to share a private seed for a pseudorandom generator. One can also reduce the shared randomness to logarithmic amounts without computational assumptions by using list-decodable codes together with standard message authentication schemes [46]. It is also possible to eliminate the shared randomness and instead require a public key [37], and this solution also relies on list-decodable codes. If explicit list-decodable codes of rate approaching $1 - h(\rho)$ for correcting a fraction ρ of errors were constructed, these would imply optimal rate explicit codes in these models as well.

In the ADV_ρ model, the channel picks the error vector e after seeing the codeword. A weaker model is that of an oblivious additive channel that can pick any worst-case error vector e (in particular the e need not have any specific distribution such as the binomial distribution), but must do so before seeing the codeword. The following result is known in this model [4]: there *exist* binary codes with encoding function $E : \{0, 1\}^k \rightarrow \{0, 1\}^n$ of rate approaching $1 - h(\rho)$ and a decoding function D such that for *every* error vector e of Hamming weight at most ρn , for *most* messages $m \in \{0, 1\}^k$, $D(E(m) + e) = m$. Note the quantifiers are flipped compared to Shannon's result for BSC_ρ : instead of decoding most error vectors for every message, we can decode most messages for every error vector. It was recently shown how one can get codes with this property by combining list-decodable codes with certain algebraic manipulation detection codes [28]. Once again, this shows the versatility of the primitive of list decoding, and highlights the importance of constructing better explicit list-decodable codes.

We have mentioned several open questions in this survey, and we end by reiterating the central Open Problem 5 on the challenge of an explicit con-

struction of binary codes of rate approaching $1 - h(\rho)$ for list decoding up to a fraction ρ of errors.

References

- [1] N. Alon and M. Luby. A linear time erasure-resilient code with nearly optimal recovery. *IEEE Transactions on Information Theory*, 42(6):1732–1736, 1996.
- [2] S. Ar, R. Lipton, R. Rubinfeld, and M. Sudan. Reconstructing algebraic functions from mixed data. *SIAM Journal on Computing*, 28(2):488–511, 1999.
- [3] E. Ben-Sasson, S. Kopparty, and J. Radhakrishnan. Subspace polynomials and list decoding of Reed-Solomon codes. In *Proceedings of the 47th Annual Symposium on Foundations of Computer Science*, pages 207–216, 2006.
- [4] I. Csiszár and P. Narayan. The capacity of the arbitrarily varying channel revisited: Positivity, constraints. *IEEE Trans. on Info. Theory*, 34(2):181–193, 1988.
- [5] Z. Dvir. On the size of Kakeya sets in finite fields. *J. Amer. Math. Soc.*, 22:1093–1097, 2009.
- [6] Z. Dvir, S. Kopparty, S. Saraf, and M. Sudan. Extensions to the method of multiplicities, with applications to Kakeya sets and mergers. In *Proceedings of the 50th Symposium on Foundations of Computer Science*, pages 181–190, 2009.
- [7] P. Elias. Coding for two noisy channels. *Information Theory, Third London Symposium*, pages 61–76, September 1955.
- [8] P. Elias. List decoding for noisy channels. *Technical Report 335, Research Laboratory of Electronics, MIT*, 1957.
- [9] P. Elias. Error-correcting codes for list decoding. *IEEE Transactions on Information Theory*, 37:5–12, 1991.
- [10] G. D. Forney. *Concatenated Codes*. MIT Press, Cambridge, MA, 1966.
- [11] G. D. Forney. Generalized Minimum Distance decoding. *IEEE Transactions on Information Theory*, 12:125–131, 1966.
- [12] A. Garcia and H. Stichtenoth. A tower of Artin-Schreier extensions of function fields attaining the Drinfeld-Vlăduț bound. *Inventiones Math.*, 121:211–222, 1995.
- [13] P. Gemmell and M. Sudan. Highly resilient correctors for multivariate polynomials. *Information Processing Letters*, 43(4):169–174, 1992.
- [14] O. Goldreich and L. Levin. A hard-core predicate for all one-way functions. In *Proceedings of the 21st Annual ACM Symposium on Theory of Computing*, pages 25–32, May 1989.
- [15] O. Goldreich, D. Ron, and M. Sudan. Chinese remaindering with errors. *IEEE Transactions on Information Theory*, 46(5):1330–1338, 2000.
- [16] V. D. Goppa. Codes on algebraic curves. *Soviet Math. Doklady*, 24:170–172, 1981.

-
- [17] V. Guruswami. *List decoding of error-correcting codes*. Number 3282 in Lecture Notes in Computer Science. Springer, 2004.
- [18] V. Guruswami. Iterative Decoding of Low-Density Parity Check Codes. *Bulletin of the European Association for Theoretical Computer Science*, 90, 2006.
- [19] V. Guruswami. *Algorithmic Results in List Decoding*, volume 2 of *Foundations and Trends in Theoretical Computer Science*. NOW publishers, 2007.
- [20] V. Guruswami. Cyclotomic function fields, Artin-Frobenius automorphisms, and list error-correction with optimal rate. *Algebra and Number Theory*, 2010. Accepted for publication. Preliminary version in 41st ACM Symp. on Theory of Computing.
- [21] V. Guruswami, J. Håstad, and S. Kopparty. On the list-decodability of random linear codes. In *Proceedings of the 42th ACM Symposium on Theory of Computing*, 2010. To appear. Available at <http://arxiv.org/abs/1001.1386>.
- [22] V. Guruswami, J. Håstad, M. Sudan, and D. Zuckerman. Combinatorial bounds for list decoding. *IEEE Transactions on Information Theory*, 48(5):1021–1035, 2002.
- [23] V. Guruswami and P. Indyk. Linear-time encodable/decodable codes with near-optimal rate. *IEEE Trans. on Information Theory*, 51(10):3393–3400, 2005.
- [24] V. Guruswami and A. Rudra. Limits to list decoding Reed-Solomon codes. *IEEE Transactions on Information Theory*, 52(8):3642–3649, 2006.
- [25] V. Guruswami and A. Rudra. Explicit codes achieving list decoding capacity: Error-correction with optimal redundancy. *IEEE Transactions on Information Theory*, 54(1):135–150, 2008.
- [26] V. Guruswami and A. Rudra. Error-correction up to the information-theoretic limit. *Communications of the ACM*, 52(3):87–95, March 2009.
- [27] V. Guruswami, A. Sahai, and M. Sudan. Soft-decision decoding of Chinese Remainder codes. In *Proceedings of the 41st IEEE Symposium on Foundations of Computer Science*, pages 159–168, 2000.
- [28] V. Guruswami and A. Smith. Explicit capacity-achieving codes for worst-case additive errors. *CoRR*, abs/0912.0965, 2009.
- [29] V. Guruswami and M. Sudan. Improved decoding of Reed-Solomon and algebraic-geometric codes. *IEEE Transactions on Information Theory*, 45:1757–1767, 1999.
- [30] V. Guruswami and M. Sudan. List decoding algorithms for certain concatenated codes. In *32nd ACM Symposium on Theory of Computing*, pages 181–190, 2000.
- [31] V. Guruswami, C. Umans, and S. Vadhan. Unbalanced expanders and randomness extractors from Parvaresh-Vardy codes. *Journal of the ACM*, 56(4), 2009.
- [32] R. W. Hamming. Error Detecting and Error Correcting Codes. *Bell System Technical Journal*, 29:147–160, April 1950.
- [33] E. Kaltofen. Polynomial factorization 1987–1991. *Proceedings of LATIN '92, I. Simon (Ed.), Springer LNCS*, 583:294–313, 1992.
- [34] R. Koetter and A. Vardy. Algebraic soft-decision decoding of Reed-Solomon codes. *IEEE Transactions on Information Theory*, 49(11):2809–2825, 2003.

- [35] R. J. Lipton. A new approach to information theory. In *Proceedings of the 11th Annual Symposium on Theoretical Aspects of Computer Science*, pages 699–708, 1994.
- [36] R. J. McEliece and L. Swanson. On the decoder error probability for Reed-Solomon codes. *IEEE Transactions on Information Theory*, 32(5):701–703, 1986.
- [37] S. Micali, C. Peikert, M. Sudan, and D. A. Wilson. Optimal error correction against computationally bounded noise. In *Proceedings of the 2nd Theory of Cryptography Conference (TCC)*, pages 1–16, 2005.
- [38] F. Parvaresh and A. Vardy. Correcting errors beyond the Guruswami-Sudan radius in polynomial time. In *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 285–294, 2005.
- [39] W. W. Peterson. Encoding and error-correction procedures for Bose-Chaudhuri codes. *IEEE Transactions on Information Theory*, 6:459–470, 1960.
- [40] T. Richardson and R. Urbanke. *Modern Coding Theory*. Cambridge University Press, 2008.
- [41] R. Roth and G. Ruckenstein. Efficient decoding of Reed-Solomon codes beyond half the minimum distance. *IEEE Trans. on Info. Theory*, 46(1):246–257, 2000.
- [42] A. Rudra and S. Uurtamo. Two theorems in list decoding. *CoRR*, abs/1001.1781, 2010.
- [43] R. Shaltiel and C. Umans. Simple extractors for all min-entropies and a new pseudorandom generator. *Journal of the ACM*, 52(2):172–216, 2005.
- [44] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.
- [45] K. W. Shum, I. Aleshnikov, P. V. Kumar, H. Stichtenoth, and V. Deolalikar. A low-complexity algorithm for the construction of algebraic-geometric codes better than the Gilbert-Varshamov bound. *IEEE Trans. on Information Theory*, 47(6):2225–2241, 2001.
- [46] A. Smith. Scrambling adversarial errors using few random bits, optimal information reconciliation, and better private codes. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 395–404, 2007.
- [47] H. Stichtenoth. *Algebraic Function Fields and Codes*. Universitext, Springer-Verlag, Berlin, 1993.
- [48] M. Sudan. Decoding of Reed-Solomon codes beyond the error-correction bound. *Journal of Complexity*, 13(1):180–193, 1997.
- [49] M. Sudan. List decoding: Algorithms and applications. *SIGACT News*, 31:16–27, 2000.
- [50] L. Trevisan. Some applications of coding theory in computational complexity. *Quaderni di Matematica*, 13:347–424, 2004.
- [51] M. A. Tsfasman, S. G. Vlăduț, and T. Zink. Modular curves, Shimura curves, and codes better than the Varshamov-Gilbert bound. *Math. Nachrichten*, 109:21–28, 1982.
- [52] S. P. Vadhan. The unified theory of pseudorandomness. *SIGACT News*, 38(3):39–54, 2007.

-
- [53] L. R. Welch and E. R. Berlekamp. Error correction of algebraic block codes. *US Patent Number 4,633,470*, December 1986.
- [54] J. M. Wozencraft. List Decoding. *Quarterly Progress Report, Research Laboratory of Electronics, MIT*, 48:90–95, 1958.
- [55] V. V. Zyablov and M. S. Pinsker. List cascade decoding. *Problems of Information Transmission*, 17(4):29–34, 1981 (in Russian); pp. 236–240 (in English), 1982.

Inapproximability of NP-complete Problems, Discrete Fourier Analysis, and Geometry

Subhash Khot*

Abstract

This article gives a survey of recent results that connect three areas in computer science and mathematics: (1) (Hardness of) computing approximate solutions to NP-complete problems. (2) Fourier analysis of boolean functions on boolean hypercube. (3) Certain problems in geometry, especially related to isoperimetry and embeddings between metric spaces.

Mathematics Subject Classification (2010). Primary 68Q17.

Keywords. NP-completeness, Approximation algorithms, Inapproximability, Probabilistically Checkable Proofs, Discrete Fourier analysis.

1. Introduction

The well-known $P \neq NP$ hypothesis says that a large class of computational problems known as NP-complete problems do not have efficient algorithms. An algorithm is called efficient if it runs in time polynomial in the length of the input. A natural question is whether one can efficiently compute *approximate* solutions to NP-complete problems and how good an approximation one can achieve. We are interested in both upper and lower bounds: designing algorithms with a guarantee on the approximation (upper bounds) as well as results showing that no efficient algorithm exists that achieves an approximation guarantee beyond a certain threshold (lower bounds). It is the latter question, namely the

*Supported by NSF CAREER grant CCF-0833228, NSF Expeditions grant CCF-0832795, and BSF grant 2008059.

251 Mercer Street, Courant Institute of Mathematical Sciences, New York University, New York, NY-10012, USA. E-mail: khot@cs.nyu.edu.

lower bounds, that is the focus of this article. Such results are known as *inapproximability* or *hardness of approximation* results, proved under a standard hypothesis such as $P \neq NP$.

Let us consider the Max-3Lin problem as an illustration. We are given a system of linear equations over $GF(2)$ with three variables in each equation and the goal is to find an assignment that satisfies the maximum number of equations. This is known to be an NP-complete problem. There is a trivial approximation algorithm that achieves a multiplicative approximation guarantee of 2. The algorithm simply assigns a random value in $GF(2)$ to each variable and in expectation satisfies half of the equations. The optimal assignment may satisfy all (or nearly all) equations, and thus the assignment produced by the algorithm is within factor 2 of the optimal assignment. On the other hand, a famous result of Håstad [25] shows that such a trivial algorithm is the best one can hope for! Specifically, let $\varepsilon > 0$ be an arbitrarily small constant. Then given an instance of Max-3Lin that has an assignment satisfying $1 - \varepsilon$ fraction of the equations, no efficient algorithm can find an assignment that satisfies $\frac{1}{2} + \varepsilon$ fraction of the equations unless $P = NP$.

It turns out that such inapproximability results are closely related to Fourier analysis of boolean functions on a boolean hypercube and to certain problems in geometry, especially related to isoperimetry. This article aims to give a survey of these connections. We anticipate that the intended audience of this article is not necessarily familiar with the techniques in computer science. We therefore focus more on the Fourier analytic and geometric aspects and only give a brief overview of how such results are used in (and often arise from) the context of inapproximability. We describe an overall framework in Section 2 and then illustrate the framework through several examples in the succeeding sections.

2. Framework for Inapproximability Results

Approximation Algorithms and Reductions

Let \mathcal{I} denote an NP-complete problem. For an instance I of the problem with input size N , let $\text{OPT}(I)$ denote the value of the optimal solution. For a specific polynomial time approximation algorithm, let $\text{ALG}(I)$ denote the value of the solution that the algorithm finds (or its expected value if the algorithm is randomized). Let $C > 1$ be a parameter that could be a function of N .

Definition 2.1. *An algorithm is said to achieve an approximation factor of C if on every instance I ,*

$$\begin{array}{ll} \text{ALG}(I) \geq \text{OPT}(I)/C & \text{if } \mathcal{I} \text{ is a maximization problem,} \\ \text{ALG}(I) \leq C \cdot \text{OPT}(I) & \text{if } \mathcal{I} \text{ is a minimization problem} \end{array}$$

A maximization problem \mathcal{I} is proved to be inapproximable by giving a reduction from a canonical NP-complete problem such as 3SAT¹ to a *gap version* of \mathcal{I} . Specifically, suppose there is a polynomial time reduction that maps a 3SAT formula ϕ to an instance I of the problem \mathcal{I} , such that for constants $0 < s < c$, we have:

1. (Completeness): If ϕ has a satisfying assignment, then $\text{OPT}(I) \geq c$.
2. (Soundness): If ϕ has no satisfying assignment, then $\text{OPT}(I) \leq s$.

Such a reduction implies that if there were an algorithm with approximation factor strictly less than $\frac{c}{s}$ for the problem \mathcal{I} , then it would enable one to efficiently decide whether a 3SAT formula is satisfiable, and hence $P = NP$. Inapproximability results for minimization problems can be proved in a similar way.

The PCP Theorem

In practice, a reduction as described above is often a sequence of (potentially very involved) reductions. In fact, the first reduction in the sequence is the famous *PCP Theorem* [18, 4, 2] which can be phrased as a reduction from 3SAT to a gap version of 3SAT. For a 3SAT formula ϕ , let $\text{OPT}(\phi)$ denote the maximum fraction of clauses that can be satisfied by any assignment. Thus $\text{OPT}(\phi) = 1$ if and only if ϕ is satisfiable. The PCP Theorem states that there is a universal constant $\alpha < 1$ and a polynomial time reduction that maps a 3SAT instance ϕ to another 3SAT instance ψ such that:

1. (Completeness): If $\text{OPT}(\phi) = 1$, then $\text{OPT}(\psi) = 1$.
2. (Soundness): If $\text{OPT}(\phi) < 1$, then $\text{OPT}(\psi) \leq \alpha$.

We stated the PCP Theorem as a combinatorial reduction. There is an equivalent formulation of it in terms of *proof checking*. The theorem states that every NP statement has a polynomial size proof that can be checked by a probabilistic polynomial time verifier by reading only a constant number of bits in the proof! The verifier has the completeness and the soundness property: every correct statement has a proof that is accepted with probability 1 and every proof of an incorrect statement is accepted with only a small probability, say at most 1%. The equivalence between the two views, namely reduction versus proof checking, is simple but illuminating, and has influenced much of the work in this area.

¹A 3SAT formula ϕ is a logical AND of a set of clauses, where each clause is a logical OR of three boolean variables, possibly negated. The goal is to decide whether the formula has a satisfying boolean assignment.

Gadgets based on Hypercube

The core of a reduction often involves a combinatorial object called a *gadget* and the reduction itself consists of taking several copies of the gadget and then appropriately connecting them together. The class of gadgets that is relevant for this article is the class of hypercube based gadgets. A simple example is the hypercube $\{-1, 1\}^n$ itself thought of as a graph. The edges of the hypercube are all pairs of inputs that differ on exactly one co-ordinate. When the computational problem under consideration is the **Graph Partitioning** problem, we are interested in partitioning a graph into two equal parts so as to minimize the number of crossing edges. A cut in the hypercube is same as a function $f : \{-1, 1\}^n \mapsto \{-1, 1\}$. The number of edges cut divided by a normalizing factor of 2^n is known as *average sensitivity* of the function. It is well-known that the minimum average sensitivity of a balanced function is 1 and the minimizer is precisely the *dictatorship* function, i.e. the function $f(x) = x_{i_0}$ for some fixed co-ordinate $i_0 \in \{1, \dots, n\}$. Note that the dictatorship function depends only on a single co-ordinate. On the other hand, a theorem of Friedgut [19] shows that any function whose average sensitivity is at most k , is very *close* to a function that depends only on $2^{O(k)}$ co-ordinates. In the contrapositive, if a function depends on too many co-ordinates and thus is *far from being a dictatorship*, then its average sensitivity must be large. Such “dictatorship is good; any function that is far from being a dictatorship is bad” kind of results are precisely the properties that we need from the gadget.

In the following, we will sketch the overall framework for inapproximability results proved via hypercube based gadgets. We refrain from describing the components of a reduction other than the gadget itself, as these typically involve computer science techniques that the reader may not be familiar with. We then illustrate this framework through several examples.

The Framework

Let $\mathcal{F} := \{f \mid f : \{-1, 1\}^n \mapsto \{-1, 1\}, \mathbb{E}[f] = 0\}$ be the class of all balanced boolean functions on the hypercube. Let

$$\text{DICT} := \{f \mid f \in \mathcal{F}, \forall x \in \{-1, 1\}^n, f(x) = x_{i_0} \text{ for some } i_0 \in \{1, \dots, n\}\},$$

be the class of dictatorship functions. Note that a dictatorship function depends only on a single co-ordinate. We aim to define a class **FFD** of functions that are to be considered as functions far from being a dictatorship. This class should include functions such as **MAJORITY** $:= \text{sign}(\sum_{i=1}^n x_i)$, **PARITY** $:= \prod_{i=1}^n x_i$, and random functions; these functions depend on all the co-ordinates in a non-trivial manner. Towards this end, let the influence of the i^{th} co-ordinate on a function f be defined as:

$$\text{Infl}_i(f) := \Pr_x [f(x_1, \dots, x_i, \dots, x_n) \neq f(x_1, \dots, -x_i, \dots, x_n)].$$

For a dictatorship function, the relevant co-ordinate has influence 1 and all other influences are zero. Thus one may define **FFD** as the class of functions

all of whose influences are small. This includes MAJORITY (all influences are $O(\frac{1}{\sqrt{n}})$), but excludes PARITY (all influences are 1) and random functions (all influences are very close to $\frac{1}{2}$). We therefore give a more refined definition that also turns out to be the most useful for the applications.

It is well-known that any function $f : \{-1, 1\}^n \mapsto \mathbb{R}$ has a Fourier (or Fourier-Walsh) representation:

$$f(x) = \sum_{S \subseteq \{1, \dots, n\}} \widehat{f}(S) \prod_{i \in S} x_i,$$

where the $\widehat{f}(S) \in \mathbb{R}$ are the Fourier coefficients. When f is a boolean function, by Parseval's identity, $\sum_S \widehat{f}(S)^2 = \mathbb{E}[f^2] = 1$. It is easily proved that:

$$\text{Infl}_i(f) = \sum_{i \in S} \widehat{f}(S)^2.$$

For an integer d , we define the *degree d influence* as:

$$\text{Infl}_i^d(f) = \sum_{i \in S, |S| \leq d} \widehat{f}(S)^2.$$

Finally, for an integer d and a parameter $\eta > 0$, let

$$\text{FFD}_{d,\eta} := \{f \mid f \in \mathcal{F}, \forall i \in \{1, \dots, n\}, \text{Infl}_i^d(f) \leq \eta\}.$$

In words, $\text{FFD}_{d,\eta}$ is the class of all functions that are far from being a dictatorship, in the sense that all degree d -influences are at most η . We will think of d as a large and η as a small constant, and $n \rightarrow \infty$ as an independent parameter. Clearly, MAJORITY, PARITY, and random functions are in this class. For MAJORITY, the influences are $O(\frac{1}{\sqrt{n}})$, and so are the degree d influences. For PARITY, the only non-zero Fourier coefficient $\widehat{f}(S)$ is for $S = \{1, \dots, n\}$ and hence all degree d -influences are zero. For a random function, the Fourier mass is concentrated on sets $|S| = \Omega(n)$, and hence the degree d -influences are negligible. We are now ready to informally state the connection between inapproximability results and Fourier analytic results:

Theorem 2.2. (Informal) Suppose \mathcal{I} is a maximization problem and $\text{Val} : \mathcal{F} \mapsto \mathbb{R}^+$ is a valuation on balanced boolean functions. Suppose there are constants $0 < s < c$ such that,

1. (Completeness): $\forall f \in \text{DICT}, \text{Val}(f) \geq c$.
2. (Soundness): $\forall f \in \text{FFD}_{d,\eta}, \text{Val}(f) \leq s$.

Assume a certain complexity theoretic hypothesis. Then given an instance of the problem \mathcal{I} that has a solution with value at least c , no polynomial time algorithm can find a solution with value exceeding s . In particular, there is no polynomial time algorithm for the problem \mathcal{I} with approximation factor strictly less than c/s .

The theorem is stated in a very informal manner and calls for several comments: (1) The choice of the valuation $\text{Val}(\cdot)$ depends very much on the problem \mathcal{I} and different problems lead to different interesting valuations. (2) We will be interested in the limiting case when $d \rightarrow \infty, \eta \rightarrow 0$. Often we will have $s = s' + \delta$ where s' is a specific constant and $\delta \rightarrow 0$ as $d \rightarrow \infty, \eta \rightarrow 0$. (3) The complexity theoretic hypothesis should ideally be $\text{P} \neq \text{NP}$, but often it will be the *Unique Games Conjecture* (see below). (4) An analogous theorem holds for minimization problems as well.

We apply the framework of Theorem 2.2 to several computational problems in the rest of the article. For each problem, we state the problem definition, the valuation $\text{Val}(\cdot)$ that is used, how the soundness property follows from a Fourier analytic result, related geometric results, and then the inapproximability result that can be proved. Before we begin, we state several properties of the dictatorship functions that will be useful and state the Unique Games Conjecture for the sake of completeness.

The valuation $\text{Val}(\cdot)$ is supposed to capture a certain property of dictatorship functions. Let us observe a few such properties:

1. Dictatorships are linear, i.e. $\forall x, y \in \{-1, 1\}^n, f(xy) = f(x)f(y)$, where xy denotes the string that is bitwise product of strings x and y .
2. Dictatorships are stable under noise, i.e. if input $x \in \{-1, 1\}^n$ is chosen uniformly at random, and $y \in \{-1, 1\}^n$ is obtained by flipping every bit of x with probability ε , then the probability that $f(x) \neq f(y)$ is ε . In contrast, MAJORITY is less stable and the probability is $\theta(\sqrt{\varepsilon})$, whereas PARITY is very unstable and the probability is very close to $\frac{1}{2}$.
3. If $C \subseteq \{-1, 1\}^n$ is a random sub-cube with dimension εn , then with probability $1 - \varepsilon$, a dictatorship function is constant on C . A sub-cube of dimension k is the set of all inputs that agree on a specific setting of input bits outside of T for some subset of co-ordinates $T \subseteq \{1, \dots, n\}$, $|T| = k$.
4. The Fourier mass of a dictatorship function is concentrated at the first level, i.e. on sets $|S| = 1$. In contrast, the Fourier mass of MAJORITY at the first level is very close to $\frac{2}{\pi}$ and that of the PARITY function is zero.

The Unique Games Conjecture

Most of the inapproximability results presented in this article rely on the Unique Games Conjecture [28] stating that a certain computational problem called the Unique Game is very hard to approximate. We do state the conjecture here, but since we are focussing only on a certain component of a reduction, we will not have an occasion to use the statement. It is easier to understand the conjecture in terms of a special case: an instance of the Linear Unique Game is a system

of linear equations over \mathbb{Z}_n where every equation is of the form $x_i - x_j = c_{ij}$, $\{x_1, \dots, x_N\}$ are variables, and $c_{ij} \in \mathbb{Z}_n$ are constants. The goal is to find an assignment to the variables that satisfies a *good* fraction of the equations.

The Unique Games Conjecture states that for every constant $\varepsilon > 0$, there is a large enough constant $n = n(\varepsilon)$, such that given an instance of **Linear Unique Game** over \mathbb{Z}_n that has an assignment satisfying $1 - \varepsilon$ fraction of the equations, no polynomial time algorithm can find an assignment that satisfies (even) an ε fraction of the equations.²

A comment about the term “Unique Game”. The term “game” refers to the context of 2-Prover-1-Round games where the problem was studied initially. Given an instance of the **Linear Unique Game**, consider the following game between two provers and a verifier: the verifier picks an equation $x_i - x_j = c_{ij}$ at random, sends the variable x_i to prover P_1 and the variable x_j to prover P_2 . Each prover is supposed to answer with a value in \mathbb{Z}_n , and the verifier accepts if and only if $a_1 - a_2 = c_{ij}$ where a_1 and a_2 are the answers of the two provers respectively. The strategies of the provers correspond to assignments $\sigma_1, \sigma_2 : \{x_1, \dots, x_N\} \mapsto \mathbb{Z}_n$. The *value* of the game is the maximum over all prover strategies, the probability that the verifier accepts. It can be easily seen that this value is between β and $\max\{1, 4\beta\}$ where β is the maximum fraction of equations that can be satisfied by any assignment. The term “unique” refers to the property of the equations $x_i - x_j = c_{ij}$ that for every value to one variable, there is a unique value to the other variable so that the equation is satisfied.

3. Max-3Lin and Linearity Test with Perturbation

Max-3Lin Problem: Given a system of linear equations over $GF(2)$ with each equation containing three variables. The goal is to find an assignment that satisfies a maximum fraction of equations.

Valuation: We define $\text{Val}(f)$ as the probability that f passes the *linearity test* along with a small perturbation. Specifically, pick two inputs $x, y \in \{-1, 1\}^n$ uniformly at random and let $w := xy$. Let z be a string obtained by flipping each bit of w with probability ε independently. Note that the correlation of every bit in z with the corresponding bit in w is $1 - 2\varepsilon$ and let $z \sim_{1-2\varepsilon} w$ denote this. Define

$$\text{Val}(f) := \Pr_{x,y,z \sim_{1-2\varepsilon} w} [f(z) = f(x)f(y)].$$

²The original conjecture is stated in terms of a more general problem, but it is shown in [29] that the conjecture is equivalent to the statement here in terms of linear unique games. Also, the “hardness” is conjectured to be NP-hardness rather than just saying that there is no polynomial time algorithm.

The optimization problem concerns linear equations with three variables, and the valuation is defined in terms of a test that depends linearly on the values of f at three random (but correlated) inputs.

Completeness: If $f \in \text{DICT}$, then it is easily seen that $\text{Val}(f) = 1 - \varepsilon$. Indeed, for some fixed co-ordinate $i_0 \in \{1, \dots, n\}$, $f(x) = x_{i_0}, f(y) = y_{i_0}, f(z) = z_{i_0}$, and z_{i_0} is obtained by flipping the value of $x_{i_0}y_{i_0}$ with probability ε . Hence we have $f(z) = f(x)f(y)$ with probability $1 - \varepsilon$.

Soundness: We will sketch a proof showing that if $f \in \text{FFD}_{d,\eta}$, then $\text{Val}(f) \leq \frac{1}{2} + \delta$ where $\delta \rightarrow 0$ as $d \rightarrow \infty, \eta \rightarrow 0$. The key observation is that the probability of acceptance of the test can be written in terms of Fourier coefficients of f . It is a rather straightforward exercise (that we skip) to show that:

$$\begin{aligned} \text{Val}(f) &= \frac{1}{2} + \frac{1}{2} \sum_{S \subseteq \{1, \dots, n\}} \widehat{f}(S)^3 (1 - 2\varepsilon)^{|S|} \\ &= \frac{1}{2} + \frac{1}{2} \sum_{S \subseteq \{1, \dots, n\}} \widehat{f}(S)^2 \left(\widehat{f}(S) \cdot (1 - 2\varepsilon)^{|S|} \right). \end{aligned}$$

Note that $\sum_S \widehat{f}(S)^2 = 1$ and since the function is balanced $\widehat{f}(\emptyset) = 0$. Thus it suffices to show that for every $S \neq \emptyset$, $\left| \widehat{f}(S)(1 - 2\varepsilon)^{|S|} \right| \leq \delta$. Since the degree d -influence of each co-ordinate is at most η , it must be that for every set $S \neq \emptyset$, either $|S| > d$ or $\widehat{f}(S)^2 \leq \eta$, as otherwise any co-ordinate in S will have degree d -influence at least η . Thus setting $\delta = \max\{(1 - 2\varepsilon)^d, \sqrt{\eta}\}$ proves the claim.

Inapproximability Result: Applying Theorem 2.2, gives the following inapproximability result proved by Håstad [25].

Theorem 3.1. *Assume $P \neq NP$ and let $\varepsilon, \delta > 0$ be arbitrarily small constants. Given an instance of the Max-3Lin problem that has an assignment satisfying $1 - \varepsilon$ fraction of the equations, no polynomial time algorithm can find an assignment that satisfies $\frac{1}{2} + \delta$ fraction of the equations. In particular, there is no polynomial time algorithm for the Max-3Lin problem with approximation factor strictly less than 2.*

4. Max- k CSP and Gowers Uniformity

Max- k CSP Problem: Given a set of N boolean variables, and a system of constraints such that each constraint depends on k variables, find an assignment to the variables that satisfies a maximum fraction of constraints. For the ease of presentation, we assume that $k = 2^q - 1$ is a large constant.

Valuation: We define $\text{Val}(f)$ to be the probability that f passes the hypergraph linearity test with perturbation. The test is a generalized and iterated version of the linearity test with perturbation in Section 3. Specifically, pick q inputs

$x^1, \dots, x^q \in \{-1, 1\}^n$ at random. For every set $S \subseteq \{1, \dots, q\}, |S| \geq 2$, let $w^S := \prod_{i \in S} x^i$ and z^S be obtained by flipping each bit of w^S with probability ε independently, i.e. $z^S \sim_{1-2\varepsilon} w^S$. The test passes if for every $S, f(z^S) = \prod_{i \in S} f(x^i)$, i.e.

$$\text{Val}(f) := \Pr_{x^1, \dots, x^q, z^S \sim_{1-2\varepsilon} w^S} \left[\forall |S| \geq 2, f(z^S) = \prod_{i \in S} f(x^i) \right].$$

Completeness: If $f \in \text{DICT}$, it is easily seen that $\text{Val}(f) \geq 1 - \varepsilon \cdot 2^q$, as there are $2^q - q - 1$ sets $|S| \geq 2$, and the test for each S could fail with probability ε due to the ε -noise/perturbation.

Soundness: It can be shown that if $f \in \text{FFD}_{d,\eta}$, then $\text{Val}(f) \leq \frac{1}{2^{2^q - q - 1}} + \delta$ where $\delta \rightarrow 0$ as $d \rightarrow \infty$ and $\eta \rightarrow 0$. Note that there are $2^q - q - 1$ sub-tests, one for each $|S| \geq 2$. If f has all influences small, then these tests behave as if they were independent tests, each tests accepts with probability essentially $\frac{1}{2}$, and hence the probability that all tests accept simultaneously is essentially $\frac{1}{2^{2^q - q - 1}}$.

Samorodnitsky and Trevisan [43] relate the acceptance probability of the test to the Gowers Uniformity norms [22] of a function, and then show that for a function with all influences small, the Gowers Uniformity norm is small as well.

Definition 4.1. Gowers Uniformity: Let $f : \{-1, 1\}^n \mapsto \{-1, 1\}$ be a function, and $\ell \geq 1$ be an integer. The dimension- ℓ uniformity of f is defined as:

$$U^\ell(f) := \mathbb{E}_{x, x^1, \dots, x^\ell} \left[\prod_{S \subseteq \{1, \dots, \ell\}} f \left(x \cdot \prod_{i \in S} x^i \right) \right].$$

Theorem 4.2. ([43]) If f is a balanced function such that $\forall i \in \{1, \dots, n\}, \text{Infl}_i(f) \leq \eta$, then $U^\ell(f) \leq \sqrt{\eta} \cdot 2^{O(\ell)}$.

Inapproximability Result:

Theorem 4.3. ([43]) Assume the Unique Games Conjecture and let $\varepsilon, \delta > 0$ be arbitrarily small constants. Then given an instance of Max- k CSP problem, $k = 2^q - 1$, that has an assignment satisfying $1 - \varepsilon \cdot 2^q$ fraction of the constraints, no polynomial time algorithm can find an assignment that satisfies at least $\frac{1}{2^{2^q - q - 1}} + \delta$ fraction of the constraints. In particular, there is no polynomial time algorithm for the Max- k CSP problem with approximation factor strictly less than $2^{2^q - q - 1} = \theta(2^k/k)$.

We note that an algorithm with approximation factor of $O(2^k/k)$ is known [8] and therefore the inapproximability result is nearly optimal.

5. Graph Partitioning and Bourgain's Noise Sensitivity Theorem

Graph Partitioning Problem: Given a graph $G(V, E)$, find a partition of the graph into two equal (or roughly equal) parts so as to minimize the fraction of edges cut. Note that this is a minimization problem.

Valuation: We define $\text{Val}(f) = \text{NS}_\varepsilon(f)$, the ε -noise sensitivity of f , i.e. the probability that f passes the perturbation test with ε -noise. Specifically, pick input $x \in \{-1, 1\}^n$ at random, and let y be a string obtained by flipping each bit of the string x with probability ε , i.e. $x \sim_{1-2\varepsilon} y$. Define

$$\text{Val}(f) := \text{NS}_\varepsilon(f) := \text{Prob}_{x \sim_{1-2\varepsilon} y} [f(x) \neq f(y)].$$

The optimization problem concerns balanced cuts in graphs. Consider a complete graph with vertices $\{-1, 1\}^n$ and non-negative weights on edges where the weight of an edge (x, y) is exactly the probability that the pair (x, y) is picked by the perturbation test. View a balanced function $f : \{-1, 1\}^n \mapsto \{-1, 1\}$ as a cut in the graph. Thus $\text{Val}(f)$ is exactly the total weight of edges cut by f .

Completeness: If $f \in \text{DICT}$, then it is easily seen that $\text{Val}(f) = \varepsilon$.

Soundness: It turns out that for large enough d and small enough η (depending on ε), if $f \in \text{FFD}_{d, \eta}$, then $\text{Val}(f) \geq \Omega(\sqrt{\varepsilon})$. This follows either from the Majority Is Stablest Theorem [38] that we will describe in the next section or essentially from the Bourgain's Theorem stated below. Bourgain's Theorem only gives a lower bound of $\Omega(\varepsilon^c)$ for any constant $c > \frac{1}{2}$, but its conclusion is stronger in the following sense: if the noise sensitivity of a balanced boolean function is $O(\varepsilon^c)$, then not only that f has a variable with significant influence, in fact f is close to a function that depends only on a bounded number of co-ordinates. The precise statement is:

Theorem 5.1. (Bourgain [7]) *Let $c > \frac{1}{2}$ be fixed. Then for all sufficiently small $\varepsilon > 0$, if f is a balanced function with ε -noise sensitivity $O(\varepsilon^c)$, then there is a boolean function g that agrees with f on 99% of the inputs and g depends only on $2^{O(1/\varepsilon^2)}$ co-ordinates.*

We would like to point out that Bourgain's Theorem came as an answer to a question posed by Håstad who was interested in such a theorem towards application to inapproximability.

Inapproximability Result: Applying Theorem 2.2 gives the following inapproximability result proved by Khot and Vishnoi [34]. The result applies to a generalization of the Graph Partitioning problem: one has so-called *demands*, i.e. a collection of pairs of vertices and we are interested in cuts that are balanced w.r.t. the demands, i.e. cuts that separate at least a constant fraction of the demands. The Graph Partitioning problem is a special case when all $\binom{|V|}{2}$ vertex pairs occur as demands.

Theorem 5.2. ([34]) *Assume the Unique Games Conjecture. Given a graph $G(V, E)$ along with demands that has a balanced partition that cuts at most ε fraction of the edges, no polynomial time algorithm can find a balanced partition that cuts at most $o(\sqrt{\varepsilon})$ fraction of the edges. In particular, there is no polynomial time algorithm for the Graph Partitioning problem with an approximation factor that is a universal constant.*

Connection to Metric Embeddings

The Graph Partitioning problem has a close connection to the theory of metric embeddings. We refer to Naor's article [39] for a detailed treatment of this connection and give a brief overview here. Theorem 5.2 rules out a constant factor approximation algorithm for the Graph Partitioning problem with demands; however the result is conditional on the Unique Games Conjecture. It is also interesting to have unconditional results that rule out a specific class of algorithms such as those based on *Semi-definite Programming relaxation*. It turns out that the performance of an SDP algorithm for the Graph Partitioning problem is closely related to the question of embedding the *negative type* metrics into the class of ℓ_1 metrics. An N -point finite metric $d(\cdot, \cdot)$ is said to be of *negative type* if the metric \sqrt{d} is isometrically embeddable in ℓ_2 . Let $c_1(\text{NEG}, N)$ be the least number such that every N -point negative type metric embeds into the class of ℓ_1 metrics with *distortion* $c_1(\text{NEG}, N)$, i.e. preserving all distances up to a factor of $c_1(\text{NEG}, N)$. It is known that $c_1(\text{NEG}, N)$ is same up to a constant factor, the performance of the SDP algorithm for the Graph Partitioning problem on N -vertex graphs. Given an N -vertex graph that has a balanced partition that cuts ε fraction of the edges, the SDP algorithm finds a balanced partition that cuts $O(\varepsilon \cdot c_1(\text{NEG}, N))$ fraction of the edges. Goemans and Linial [21, 37] conjectured that $c_1(\text{NEG}, N)$ is a universal constant independent of N ; this would be contrary to the statement of Theorem 5.2 since the theorem rules out every polynomial time algorithm that might achieve a constant factor approximation, and in particular an SDP-based one. In fact, using the techniques used to prove Theorem 5.2, Khot and Vishnoi [34] were able to disprove the Goemans and Linial conjecture:

Theorem 5.3. ([34]) $c_1(\text{NEG}, N) \geq \Omega((\log \log N)^c)$ for some constant $c > 0$.

An interesting aspect of this theorem is that the construction of the negative type metric is inspired by the Unique Games Conjecture and the PCP reduction used to prove Theorem 5.2, but the construction itself is explicit and the lower bound unconditional. Regarding the upper bounds on $c_1(\text{NEG}, N)$, in a breakthrough work, Arora, Rao, and Vazirani [3] showed that the SDP algorithm gives $O(\sqrt{\log N})$ approximation to the Graph Partitioning problem (without demands). This was extended to the demands version of the problem by Arora, Lee, and Naor [1], albeit with a slight loss in the approximation factor. As discussed, the latter result is equivalent to an upper bound on $c_1(\text{NEG}, N)$.

Theorem 5.4. ([1]) $c_1(\text{NEG}, N) \leq O(\sqrt{\log N} \cdot \log \log N)$.³

Using an alternate construction based on the geometry of Heisenberg group, a sequence of works by Lee and Naor [36], Cheeger and Kleiner [11, 12], Cheeger, Kleiner, and Naor [13, 14] obtained a stronger lower bound than Theorem 5.3:

Theorem 5.5. ([36, 11, 12, 13, 14]) $c_1(\text{NEG}, N) \geq \Omega((\log N)^c)$ for some constant $c > 0$.

The lower bound of Theorem 5.3 is also strengthened in a different direction by Raghavendra and Steurer [40] (also by Khot and Saket [33] with quantitatively weaker result):

Theorem 5.6. ([40, 33]) *There is an N -point negative type metric such that its submetric on any subset of t points is isometrically ℓ_1 -embeddable, but the whole metric incurs distortion of at least t to embed into ℓ_1 , and $t = (\log \log N)^c$ for some constant $c > 0$.*

The KKL Theorem

A result of Kahn, Kalai, and Linial [27] was used by Chawla *et al* [10] to prove a theorem analogous to Theorem 5.2, and also by Krauthagamer and Rabani [35] and Devanur *et al* [15] to improve the lower bound in Theorem 5.3 to $\Omega(\log \log N)$. The KKL result has many other applications and we state it below:

Theorem 5.7. ([27]) *Every balanced boolean function $f : \{-1, 1\}^n \mapsto \{-1, 1\}$ has a variable whose influence is $\Omega\left(\frac{\log n}{n}\right)$.*

6. Majority Is Stablest and Borell's Theorem

In the last section, we studied the ε -noise sensitivity of balanced boolean functions. Bourgain's Theorem gives a lower bound of $\Omega(\varepsilon^c)$ on the noise sensitivity of a balanced function whose all influences are small and $c > \frac{1}{2}$. We also mentioned that the Majority Is Stablest Theorem gives a lower bound of $\Omega(\sqrt{\varepsilon})$. In fact it gives an exact lower bound, namely $\frac{1}{\pi} \arccos(1 - 2\varepsilon)$, which turns out to be useful for an inapproximability result for the Max-Cut problem presented in the next section. Indeed, the Majority Is Stablest Theorem was invented for this application!

³Arora, Lee, and Naor [1] in fact give an embedding of an N -point negative type metric into ℓ_2 (which is isometrically embeddable into ℓ_1) with distortion $O(\sqrt{\log N} \cdot \log \log N)$. Since ℓ_1 metrics are of negative type, this gives an embedding of an N -point ℓ_1 metric into ℓ_2 with the same distortion. The result essentially matches a decades long lower bound of Enflo [17] who showed that embedding N -point ℓ_1 metric into ℓ_2 incurs distortion $\Omega(\sqrt{\log N})$.

Theorem 6.1. (Mossel, O’Donnell, Oleszkiewicz [38]) *Let $0 < \varepsilon < \frac{1}{2}$. If $f \in \text{FFD}_{d,\eta}$, then*

$$\text{NS}_\varepsilon(f) \geq \frac{1}{\pi} \arccos(1 - 2\varepsilon) - \delta$$

and $\delta \rightarrow 0$ as $d \rightarrow \infty, \eta \rightarrow 0$.

We present a sketch of the proof as it demonstrates the connection to an isoperimetric problem in geometry and its solution by Borell [6]. The proof involves an application of the *invariance principle* that has also been studied by Rotar [42] and Chatterjee [9]. Here is a rough statement of the invariance principle:

Invariance Principle [38, 42, 9]: *Suppose f is a low degree multi-linear polynomial in n variables and all its variables have small influence. Then the distribution of the values of f is nearly identical when the input is a uniform random point from $\{-1, 1\}^n$ or a random point from \mathbb{R}^n with standard Gaussian measure.*

The invariance principle allows us to translate the noise sensitivity problem on boolean hypercube to a similar problem in the Gaussian space and the latter problem has already been solved by Borell! Towards this end, let $f \in \text{FFD}_{d,\eta}$ be a boolean function on n -dimensional hypercube. We intend to lower bound its ε -noise sensitivity. We know that f has a representation as a multi-linear polynomial, namely its Fourier expansion:

$$f(x) = \sum_S \widehat{f}(S) \prod_{i \in S} x_i \quad \forall x \in \{-1, 1\}^n.$$

Let $f^* : \mathbb{R}^n \mapsto \mathbb{R}$ be a function that has the same representation as a multi-linear polynomial as f :

$$f^*(x^*) = \sum_S \widehat{f}(S) \prod_{i \in S} x_i^* \quad \forall x^* \in \mathbb{R}^n. \quad (1)$$

Since $f \in \text{FFD}_{d,\eta}$, all its influences are small. Assume for the moment that f is also of *low* degree. By the invariance principle, the distributions of $f(x)$ and $f^*(x^*)$ are nearly identical, and let us assume them to be identical for the sake of simplicity. This implies that $\mathbb{E}[f^*] = \mathbb{E}[f] = 0$ and since f is boolean, so is f^* . In other words, f^* is a partition of \mathbb{R}^n (with Gaussian measure) into two sets of equal measure. The next observation is that the ε -noise sensitivity of f is same as the ε -“Gaussian noise sensitivity” of $f^* : \mathbb{R}^n \mapsto \{-1, 1\}$. To be precise, let (x^*, y^*) be a pair of $(1 - 2\varepsilon)$ -correlated n -dimensional Gaussians, i.e. for every co-ordinate i , (x_i^*, y_i^*) are $(1 - 2\varepsilon)$ -correlated standard Gaussians. One way to generate such a pair is to pick two independent standard n -dimensional Gaussians x^* and z^* , and let $y^* = (1 - 2\varepsilon)x^* + \sqrt{1 - (1 - 2\varepsilon)^2}z^*$, and thus one can think of y^* as a small perturbation of x^* . Let the ε -noise sensitivity of a function $f^* : \mathbb{R}^n \mapsto \{-1, 1\}$ be defined as:

$$\text{NS}_\varepsilon(f^*) := \Pr_{x^* \sim_{1-2\varepsilon} y^*} [f^*(x^*) \neq f^*(y^*)].$$

When f^* is a multi-linear polynomial as in (1), it is easily observed that

$$\text{NS}_\varepsilon(f^*) = \frac{1}{2} - \frac{1}{2} \sum_S \widehat{f}(S)^2 (1 - 2\varepsilon)^{|S|}.$$

But this expression is same as the ε -noise sensitivity of the boolean function f and thus $\text{NS}_\varepsilon(f) = \text{NS}_\varepsilon(f^*)$ and Theorem 6.1 follows from Borell’s result that lower bounds $\text{NS}_\varepsilon(f^*)$.

Theorem 6.2. (Borell [6]) *If $g^* : \mathbb{R}^n \mapsto \{-1, 1\}$ is a measurable function with $\mathbb{E}[g^*] = 0$, then*

$$\text{NS}_\varepsilon(g^*) \geq \text{NS}_\varepsilon(\text{HALF SPACE}) = \frac{1}{\pi} \arccos(1 - 2\varepsilon),$$

where HALF-SPACE is the partition of \mathbb{R}^n by a hyperplane through origin.

We note that the parameter δ in the statement of Theorem 6.1 accounts for additive errors involved at multiple places during the argument: firstly, the distributions $f(x)$ and $f^*(x^*)$ are only nearly identical. Secondly, even though $f \in \text{FFD}_{d,\eta}$, f is not necessarily of bounded degree, and the invariance principle is not directly applicable. One gets around this issue by *smoothing* f that *kills* the high order Fourier coefficients (which are then discarded) and only slightly affecting the noise sensitivity. The *truncated* version of f has bounded degree and the invariance principle can be applied. We also note that the statement of Borell’s Theorem holds for g^* that is $[-1, 1]$ -valued when the noise sensitivity is defined as $\frac{1}{2} - \frac{1}{2} \langle g^*, T_{1-2\varepsilon} g^* \rangle$ and $T_{1-2\varepsilon}$ is the Ornstein-Uhlenbeck operator.

7. Max-Cut Problem

Max-Cut Problem: Given a graph $G(V, E)$, find a partition that maximizes the number of edges cut.

Valuation: We define $\text{Val}(f)$ as the ε -noise sensitivity of f for an appropriately chosen constant $\varepsilon > \frac{1}{2}$. Specifically, pick input $x \in \{-1, 1\}^n$ at random, and let y be a string obtained by flipping each bit of the string x with probability ε , i.e. $x \sim_{1-2\varepsilon} y$. Define

$$\text{Val}(f) := \text{Prob}_{x \sim_{1-2\varepsilon} y} [f(x) \neq f(y)].$$

The optimization problem concerns cuts in graphs. As in Section 5, we consider the complete graph with vertices $\{-1, 1\}^n$ and non-negative weights on edges representing the probability that a pair (x, y) is picked, and $f : \{-1, 1\}^n \mapsto \{-1, 1\}$ as a cut in the graph. An important thing to note here is that for the Graph Partitioning problem, the goal is to minimize the noise sensitivity for a balanced cut, and $\varepsilon > 0$ is a small constant. On the other hand, for the Max-Cut problem, the goal is to maximize the noise sensitivity, and $\varepsilon > \frac{1}{2}$.

Completeness: If $f \in \text{DICT}$, then it is easily seen that $\text{Val}(f) = \varepsilon$.

Soundness: The Majority Is Stablest Theorem states that MAJORITY is the most *stable* function among the class of low influence balanced boolean functions. It is implicit in this statement that the noise rate is strictly less than $\frac{1}{2}$. It turns out, essentially from the same theorem, that when the noise rate is above $\frac{1}{2}$, MAJORITY is the most *unstable* function among the class of low influence boolean functions (even including the unbalanced ones). This allows us to show that if $f \in \text{FFD}_{d,\eta}$,

$$\text{Val}(f) \leq \frac{1}{\pi} \arccos(1 - 2\varepsilon) + \delta, \quad \text{where } \delta \rightarrow 0 \text{ as } d \rightarrow \infty, \eta \rightarrow 0.$$

Inapproximability Result: Khot *et al* [29] proved the following inapproximability result for the Max-Cut problem and the Majority Is Stablest Theorem was conjectured therein.

Theorem 7.1. ([29]) *Assume the Unique Games Conjecture and let $\varepsilon > \frac{1}{2}$. Let $\delta > 0$ be an arbitrarily small constant. Given a graph $G(V, E)$ that has a partition that cuts at least ε fraction of the edges, no polynomial time algorithm can find a partition that cuts at least $\frac{1}{\pi} \arccos(1 - 2\varepsilon) + \delta$ fraction of the edges. In particular, there is no polynomial time algorithm for the Max-Cut problem with an approximation factor that is strictly less than $\frac{\varepsilon}{\frac{1}{\pi} \arccos(1 - 2\varepsilon)}$.*

In the above theorem, one can choose $\varepsilon > \frac{1}{2}$ so as to maximize the inapproximability factor. Let $\alpha_{GW} := \max_{\varepsilon \in [\frac{1}{2}, 1]} \frac{\varepsilon}{\frac{1}{\pi} \arccos(1 - 2\varepsilon)} \approx 1.13$. The theorem rules out an efficient algorithm with approximation factor strictly less than α_{GW} . On the other hand, the well-known SDP-based algorithm of Goemans and Williamson [20] achieves an approximation factor of exactly α_{GW} and thus is the optimal algorithm (modulo the Unique Games Conjecture).

8. Independent Set and the It Ain't Over Till It's Over Theorem

Independent Set Problem: Given a graph $G(V, E)$, find the largest independent set. A set $I \subseteq V$ is called independent if no edge of the graph has both endpoints in I . It is known from a result of Håstad [24], that given an N -vertex graph that has an independent set of size $N^{1-\varepsilon}$, no polynomial time algorithm can find an independent set of size N^ε unless $\text{P} = \text{NP}$. In this section, we are interested in the case when the graph is almost 2-colorable, i.e. has two disjoint independent sets of size $(\frac{1}{2} - \varepsilon)N$ each.

Valuation: We define $\text{Val}(f)$ as probability that f is constant on a random εn dimensional sub-cube. For a set of co-ordinates $S \subseteq \{1, \dots, n\}$ and a string

$x \in \{-1, 1\}^n$, a sub-cube $C_{S,x}$ corresponds to the set of all inputs that agree with x *outside* of S , i.e.

$$C_{S,x} := \{z \mid z \in \{-1, 1\}^n, \forall i \notin S, z_i = x_i\}.$$

A random sub-cube $C_{S,x}$ of dimension εn is picked by selecting a random set $S \subseteq \{1, \dots, n\}$, $|S| = \varepsilon n$ and a random string x . Define:

$$\text{Val}(f) := \Pr_{|S|=\varepsilon n, x} [f \text{ is constant on } C_{S,x}].$$

The connection between this test and the **Independent Set** problem is rather subtle. One constructs a graph whose vertices are all pairs (C, b) where C is an εn -dimensional sub-cube and $b \in \{-1, 1\}$ is a bit. The intended purpose of this vertex is to capture the possibility that $f|_C \equiv b$. If two sub-cubes C, C' have non-empty intersection and $b \neq b'$, then we cannot have both $f_C = b$ and $f_{C'} = b'$, and we introduce an edge between vertices (C, b) and (C', b') to denote this conflict. This construction is known as the FGLSS construction, invented in [18]. It is not difficult to see that an independent set in this graph corresponds to a boolean function and the size of the independent set is proportional to the probability that the function passes the random sub-cube test.

Completeness: If $f \in \text{DICT}$, then $f(x) = x_{i_0}$ for some fixed co-ordinate i_0 . It is easily seen that for a random sub-cube $C_{S,x}$, unless $i_0 \in S$, f is constant on the sub-cube. Since $|S| = \varepsilon n$, we have $\text{Val}(f) = 1 - \varepsilon$.

Soundness: If $f \in \text{FFD}_{d,\eta}$, then it can be showed that $\text{Val}(f) \leq \delta$ where $\delta \rightarrow 0$ as $d \rightarrow \infty, \eta \rightarrow 0$. It follows from the **It Ain't Over Till It's Over** Theorem of Mossel *et al* [38] which in fact says something stronger: if f has all influences small, then for almost all sub-cubes C , not only that f is non-constant on C , but f takes both the values $\{-1, 1\}$ on a constant fraction of points in C . A formal statement appears below:

Theorem 8.1. *For every $\varepsilon, \delta > 0$, there exist $\gamma, \eta > 0$ and integer d such that if $f \in \text{FFD}_{d,\eta}$, and C is a random εn -dimensional sub-cube, then*

$$\Pr_C \left[\left| \mathbb{E}[f(x)|x \in C] \right| \geq 1 - \gamma \right] \leq \delta.$$

The theorem is proved using the invariance principle. Bansal and Khot [5] gave an alternate simple proof without using the invariance principle (the random sub-cube test is proposed therein), but the conclusion is only that f is non-constant on almost every sub-cube (which suffices for their application to Independent Set problem).

Inapproximability Result:

Theorem 8.2. ([5]) *Assume the Unique Games Conjecture and let $\varepsilon, \delta > 0$ be arbitrarily small constants. Then given an N -vertex graph $G(V, E)$ that is almost 2-colorable, i.e. has two disjoint independent sets of size $(\frac{1}{2} - \varepsilon)N$ each, no polynomial time algorithm can find an independent set of size δN .*

Friedgut's Theorem

Khot and Regev [32] proved a weaker result than Theorem 8.2: assuming the Unique Games Conjecture, given an N -vertex graph $G(V, E)$ that has an independent set of size $(\frac{1}{2} - \varepsilon)N$, no polynomial time algorithm can find an independent set of size δN . This gives $2 - \varepsilon$ inapproximability factor for the Vertex Cover problem.⁴ The result is optimal since an algorithm that finds a maximal matching and takes all endpoints of the edges in the matching gives a 2-approximation for the Vertex Cover problem. Khot and Regev's paper (and its precursor Dinur and Safra [16]) use the following theorem of Friedgut [19]:

Theorem 8.3. ([19]) *Let $f : \{-1, 1\}^n \mapsto \{-1, 1\}$ be a function such that the average sensitivity (i.e. sum of all influences) is at most k . Then there exists a function g that agrees with f on $1 - \beta$ fraction of inputs and depends only on $2^{3k/\beta}$ co-ordinates.*

9. Kernel Clustering and the Propeller Problem

Kernel Clustering Problem: Given an $N \times N$ (symmetric) positive semidefinite matrix $A = (a_{ij})$ with $\sum_{i,j=1}^N a_{ij} = 0$, partition the index set $\{1, \dots, N\}$ into k sets T_1, \dots, T_k so as to maximize $\sum_{\ell=1}^k \sum_{i,j \in T_\ell} a_{ij}$. In words, we seek to partition the matrix into $k \times k$ block diagonal form and then maximize the sum of entries of all diagonal blocks. Since the matrix is PSD, this sum is necessarily non-negative. The problem is actually a special case of the Kernel Clustering problem studied in [30, 31] and we don't state the more general problem here. We think of $k \geq 2$ as a small constant.

Valuation: We define $\text{Val}(f)$ as the Fourier mass of f at the first level. We need to consider k -ary functions on k -ary hypercube, i.e. functions $f : \{1, \dots, k\}^n \mapsto \{1, \dots, k\}$. There is a natural generalization for the notions of dictatorship functions, Fourier representation, influences, and functions that are far from dictatorship. We don't formally state these notions here and directly state the definition of $\text{Val}(f)$:

$$\text{Val}(f) := \sum_{S \in \{0,1,\dots,k-1\}^n, |S|=1} \widehat{f}(S)^2,$$

where $\widehat{f}(S)$ is the Fourier coefficient corresponding to a *multi-index* $S \in \{0, 1, \dots, k-1\}^n$ and $|S|$ denotes the number of its non-zero co-ordinates. The connection between the Kernel Clustering problem and the specific valuation is that the (squared) Fourier mass is a PSD function of the values of f .

⁴A vertex cover in a graph is complement of an independent set. The Vertex Cover problem seeks to find a vertex cover of minimum size.

Completeness:⁵ If $f \in \text{DICT}$, then $\text{Val}(f) = 1 - \frac{1}{k}$.

Soundness: If $f \in \text{FFD}_{d,\eta}$, then $\text{Val}(f) \leq C(k) + \delta$ where $\delta \rightarrow 0$ as $d \rightarrow \infty$ and $\eta \rightarrow 0$. We would like to know functions that maximize the Fourier mass at the first level among the class of functions that are far from dictatorships. Since f has all its influences small, one can apply the invariance principle, and reduce this question to a certain geometric question, and the constant $C(k)$ is the solution to this geometric question. We state the geometric question below:

Definition 9.1. Let A_1, \dots, A_k be a partition of \mathbb{R}^{k-1} into k measurable sets and for $1 \leq \ell \leq k$, let z_ℓ be the Gaussian moment vector over A_ℓ , i.e.

$$z_\ell := \int_{A_\ell} x \, d\gamma \quad \text{where } \gamma \text{ is standard Gaussian measure on } \mathbb{R}^{k-1}.$$

Then $C(k)$ is the supremum (it is achieved) of the sum of squared lengths of z_ℓ 's over all possible partitions, i.e.

$$C(k) := \sup_{\mathbb{R}^{k-1} = A_1 \cup \dots \cup A_k} \sum_{\ell=1}^k \|z_\ell\|^2. \tag{2}$$

It seems challenging to characterize an optimal partition for $k \geq 4$. For $k = 2$, the optimal partition of \mathbb{R} into two sets is the partition into positive and negative real line, and $C(2) = \frac{1}{\pi}$. For $k = 3$, the optimal partition of \mathbb{R}^2 into three sets is the ‘‘propeller’’, i.e. partition into three cones with angle 120° each, and $C(3) = \frac{9}{8\pi}$. One would expect that for $k = 4$, the optimal partition of \mathbb{R}^3 into four sets is the partition into four cones given by a regular tetrahedron. This turns out to be false as numerical computation shows that the value of this partition is worse than $C(3) = \frac{9}{8\pi}$ that can be achieved by letting $\mathbb{R}^3 = \mathbb{R}^2 \times \mathbb{R}$ and then partitioning \mathbb{R}^2 as a propeller. In fact Khot and Naor [30] conjecture that the propeller partition is the optimal one for any $k \geq 3$:

Conjecture 9.2. Propeller Conjecture: For every $k \geq 3$, $C(k) = C(3)$. In words, the optimal partition of \mathbb{R}^{k-1} into k sets in the sense of (2) is achieved by letting $\mathbb{R}^{k-1} = \mathbb{R}^2 \times \mathbb{R}^{k-3}$ and partitioning \mathbb{R}^2 as a propeller.

Inapproximability Result:

Theorem 9.3. ([30, 31]) Assume the Unique Games Conjecture and let $\varepsilon, \delta > 0$ be arbitrarily small constants. Then given an instance $A = (a_{ij})$ with value $1 - \frac{1}{k} - \varepsilon$, no polynomial time algorithm can find a solution with value at least $C(k) + \delta$. In particular, there is no polynomial time algorithm for the Kernel Clustering problem with approximation factor strictly less than $\frac{1-1/k}{C(k)}$.

⁵When $k = 2$, we have boolean functions on boolean hypercube, and one would expect that for a dictatorship function, the Fourier mass at the first level equals 1. We instead get $\frac{1}{2}$ due to a slightly different (but equivalent) representation of functions.

10. Conclusion

We have presented several examples to demonstrate the connections between inapproximability, discrete Fourier analysis, and geometry. There are many more examples and we conclude with pointing out a few:

- Plurality is Stablest Conjecture:** In Section 6 and 7, we presented the connections between the Max-Cut problem, the Majority Is Stablest Theorem, and Borell's Theorem stating that a halfspace through origin is the most noise-stable balanced partition of \mathbb{R}^n . The Max-Cut problem can be generalized to the Max- k Cut problem where one seeks to partition a graph into $k \geq 3$ sets so as to maximize the number of edges cut. An optimal inapproximability result for this problem is implied by the Plurality Is Stablest Conjecture stating that the Plurality function from $\{1, \dots, k\}^n$ to $\{1, \dots, k\}$ is the most stable under noise among the class of functions that are balanced and whose all influences are small. This conjecture in turn is implied by the Standard Simplex Conjecture stating that the standard k -simplex partition is the most noise-stable balanced partition of \mathbb{R}^n with $n \geq k - 1$ (see [26]).
- Sub-cube Test:** Consider a variant of the test discussed in Section 8: Assume that f is balanced, and one tests whether f is constant -1 on a random sub-cube of linear dimension. We know that if a function f passes the test with constant probability, say α , then it must have an influential variable. However f need not be close to a junta (i.e. a function depending on a bounded number of co-ordinates). Is it necessarily true that there is a function g that is close to a junta, monotonically above f , and passes the test with probability close to α ? We say that g is monotonically above f if $\forall x, f(x) = 1 \implies g(x) = 1$. Such a result, though interesting on its own, might be useful towards inapproximability of graph coloring problem.
- Lasserre Gaps:** Theorem 5.6 states that there is an N -point negative type metric that is locally ℓ_1 -embeddable, but not globally ℓ_1 -embeddable. In computer science, this result can be thought of as an *integrality gap* result for the so-called Sherali-Adams linear programming relaxation. An integrality gap result is an explicit construction showing that there is a gap between the true optimum and the optimum of the linear or semidefinite programming relaxation. Such results are taken as evidence that LP/SDP relaxation would not lead to a good approximation algorithm. There is a SDP relaxation known as Lasserre relaxation that is at least as powerful as the Sherali-Adams relaxation. It is a challenging open problem to prove integrality gap results for the Lasserre relaxation (for any problem of interest such as Max-Cut, Vertex Cover, or Unique Game). This could lead to interesting questions in Fourier analysis and/or geometry.
- Small Set Expansion Problem:** Raghavendra and Steurer [41] give a connection between the small set expansion problem and the Unique

Games Conjecture. Given an N -vertex graph, the goal is to find a set of vertices S of size δN that is nearly non-expanding, i.e. only a tiny fraction of edges incident on S leave S . One could conjecture that finding such sets is computationally intractable. Such a conjecture (see [41] for a formal statement) implies the Unique Games Conjecture as shown in [41].

- **Bounded Spectral Norm:** A result of Green and Sanders [23] states that every function $f : GF(2)^n \mapsto \{0, 1\}$ that has bounded spectral norm (defined as the sum of absolute values of its Fourier coefficients) can be expressed as a sum of a bounded number of functions each of which is an indicator function of an affine subspace of $GF(2)^n$. This result has the same flavor as “dictatorships are good; functions far from dictatorships are bad”, except that now indicators of affine subspaces are considered as the “good” functions. Since there is such a close connection between such theorems and inapproximability results, it would be interesting to find an application to inapproximability, if there is one.

References

- [1] S. Arora, J. Lee, and A. Naor. Euclidean distortion and the sparsest cut. In *Proc. 37th ACM Symposium on Theory of Computing*, pages 553–562, 2005.
- [2] S. Arora, C. Lund, R. Motawani, M. Sudan, and M. Szegedy. Proof verification and the hardness of approximation problems. *Journal of the ACM*, 45(3):501–555, 1998.
- [3] S. Arora, S. Rao, and U. Vazirani. Expander flows, geometric embeddings and graph partitioning. In *Proc. 36th ACM Symposium on Theory of Computing*, pages 222–231, 2004.
- [4] S. Arora and S. Safra. Probabilistic checking of proofs : A new characterization of NP. *Journal of the ACM*, 45(1):70–122, 1998.
- [5] N. Bansal and S. Khot. Optimal long code test with one free bit. In *Proc. 50th IEEE Symposium on Foundations of Computer Science*, 2009.
- [6] C. Borell. Geometric bounds on the Ornstein-Uhlenbeck velocity process. *Z. Wahrsch. Verw. Gebiete*, 70(1):1–13, 1985.
- [7] J. Bourgain. On the distribution of the Fourier spectrum of boolean functions. *Israel J. of Math.*, (131):269–276, 2002.
- [8] M. Charikar, K. Makarychev, and Y. Makarychev. Near-optimal algorithms for maximum constraint satisfaction problems. In *SODA*, 2007.
- [9] S. Chatterjee. A simple invariance theorem. *arXiv:math/0508213v1*, 2005.
- [10] S. Chawla, R. Krauthgamer, R. Kumar, Y. Rabani, and D. Sivakumar. On the hardness of approximating multicut and sparsest-cut. In *Proc. 20th IEEE Conference on Computational Complexity*, pages 144–153, 2005.

- [11] J. Cheeger and B. Kleiner. Differentiating maps into L^1 and the geometry of BV functions. *To appear in Ann. Math., preprint available at <http://arxiv.org/abs/math/0611954>*, 2006.
- [12] J. Cheeger and B. Kleiner. On the differentiation of Lipschitz maps from metric measure spaces to Banach spaces. *Inspired by S.S. Chern, Volume 11 of Nankai Tracts. Math.*, pages 129–152, 2006.
- [13] J. Cheeger, B. Kleiner, and A. Naor. Compression bounds for Lipschitz maps from the Heisenberg group to l_1 . *Preprint*, 2009.
- [14] J. Cheeger, B. Kleiner, and A. Naor. A $(\log n)^{\Omega(1)}$ integrality gap for the sparsest cut sdp. In *Proc. 50th IEEE Symposium on Foundations of Computer Science*, 2009.
- [15] N. Devanur, S. Khot, R. Saket, and N. Vishnoi. Integrality gaps for sparsest cut and minimum linear arrangement problems. In *Proc. 38th ACM Symposium on Theory of Computing*, 2006.
- [16] I. Dinur and S. Safra. The importance of being biased. In *Proc. 34th Annual ACM Symposium on Theory of Computing*, 2002.
- [17] P. Enflo. On the nonexistence of uniform homeomorphisms between lp-spaces. *Ark. Mat.*, 8:103–105, 1969.
- [18] U. Feige, S. Goldwasser, L. Lovász, S. Safra, and M. Szegedy. Interactive proofs and the hardness of approximating cliques. *Journal of the ACM*, 43(2):268–292, 1996.
- [19] E. Friedgut. Boolean functions with low average sensitivity depend on few coordinates. *Combinatorica*, 18(1):27–35, 1998.
- [20] M. Goemans and D. Williamson. 0.878 approximation algorithms for MAX-CUT and MAX-2SAT. In *Proc. 26th ACM Symposium on Theory of Computing*, pages 422–431, 1994.
- [21] M. X. Goemans. Semidefinite programming in combinatorial optimization. *Math. Program.*, 79:143–161, 1997.
- [22] T. Gowers. A new proof of Szemerédi’s theorem for progressions of length four. *Geometric and Functional Analysis*, 8(3):529–551, 1998.
- [23] B. Green and T. Sanders. Boolean functions with small spectral norm. *Geom. Funct. Anal.*, 18(1):144–162, 2008.
- [24] J. Hastad. Clique is hard to approximate within $n^{1-\epsilon}$. *Acta Mathematica*, 182:105–142, 1999.
- [25] J. Hastad. Some optimal inapproximability results. *Journal of ACM*, 48:798–859, 2001.
- [26] M. Isaksson and E. Mossel. New maximally stable gaussian partitions with discrete applications. *arXiv:0903.3362v3*, 2010.
- [27] J. Kahn, G. Kalai, and N. Linial. The influence of variables on boolean functions. In *Proc. 29th Symposium on the Foundations of Computer Science*, pages 68–80, 1988.
- [28] S. Khot. On the power of unique 2-prover 1-round games. In *Proc. 34th ACM Symposium on Theory of Computing*, 2002.

- [29] S. Khot, G. Kindler, E. Mossel, and R. O’Donnell. Optimal inapproximability results for max-cut and other 2-variable csps? In *Proc. 45th IEEE Symposium on Foundations of Computer Science*, pages 146–154, 2004.
- [30] S. Khot and A. Naor. Approximate kernel clustering. In *Proc. 49th IEEE Symposium on Foundations of Computer Science*, 2008.
- [31] S. Khot and A. Naor. Sharp kernel clustering algorithms and their associated Grothendieck inequalities. In *SODA*, 2010.
- [32] S. Khot and O. Regev. Vertex cover might be hard to approximate to within $2 - \epsilon$. In *Proc. 18th IEEE Conference on Computational Complexity*, 2003.
- [33] S. Khot and R. Saket. Sdp integrality gaps with local ℓ_1 -embeddability. In *Proc. 50th IEEE Symposium on Foundations of Computer Science*, pages 565–574, 2009.
- [34] S. Khot and N. Vishnoi. The unique games conjecture, integrality gap for cut problems and embeddability of negative type metrics into ℓ_1 . In *Proc. 46th IEEE Symposium on Foundations of Computer Science*, 2005.
- [35] R. Krauthgamer and Y. Rabani. Improved lower bounds for embeddings into l_1 . In *Proc. ACM-SIAM Symposium on Discrete Algorithms*, 2006.
- [36] J. R. Lee and A. Naor. l_p metrics on the Heisenberg group and the Goemans-Linial conjecture. In *Proc. 47th IEEE Symposium on Foundations of Computer Science*, pages 99–108, 2006.
- [37] N. Linial. Finite metric spaces-combinatorics, geometry and algorithms. In *Proc. International Congress of Mathematicians*, volume 3, pages 573–586, 2002.
- [38] E. Mossel, R. O’Donnell, and K. Oleszkiewicz. Noise stability of functions with low influences: invariance and optimality. In *Proc. 46th IEEE Symposium on Foundations of Computer Science*, pages 21–30, 2005.
- [39] A. Naor. L_1 embeddings of the Heisenberg group and fast estimation of graph isoperimetry. In *Proc. the International Congress of Mathematicians*, 2010.
- [40] P. Raghavendra and D. Steurer. Integrality gaps for strong sdp relaxations of unique games. In *Proc. 50th IEEE Symposium on Foundations of Computer Science*, pages 575–585, 2009.
- [41] P. Raghavendra and D. Steurer. Graph expansion and the unique games conjecture. In *Proc. 42nd ACM Symposium on Theory of Computing*, 2010.
- [42] V. Rotar’. Limit theorems for polylinear forms. *J. Multivariate Anal.*, 9(4):511–530, 1979.
- [43] A. Samorodnitsky and L. Trevisan. Gowers uniformity, influence of variables, and PCPs. In *Proc. 38th ACM Symposium on Theory of Computing*, 2006.

Algorithms, Graph Theory, and Linear Equations in Laplacian Matrices

Daniel A. Spielman*

Abstract

The Laplacian matrices of graphs are fundamental. In addition to facilitating the application of linear algebra to graph theory, they arise in many practical problems.

In this talk we survey recent progress on the design of provably fast algorithms for solving linear equations in the Laplacian matrices of graphs. These algorithms motivate and rely upon fascinating primitives in graph theory, including low-stretch spanning trees, graph sparsifiers, ultra-sparsifiers, and local graph clustering. These are all connected by a definition of what it means for one graph to approximate another. While this definition is dictated by Numerical Linear Algebra, it proves useful and natural from a graph theoretic perspective.

Mathematics Subject Classification (2010). Primary 68Q25; Secondary 65F08.

Keywords. Preconditioning, Laplacian Matrices, Spectral Graph Theory, Sparsification.

1. Introduction

We all learn one way of solving linear equations when we first encounter linear algebra: Gaussian Elimination. In this survey, I will tell the story of some remarkable connections between algorithms, spectral graph theory, functional analysis and numerical linear algebra that arise in the search for asymptotically faster algorithms. I will only consider the problem of solving systems of linear equations in the Laplacian matrices of graphs. This is a very special case, but it is also a very interesting case. I begin by introducing the main characters in the story.

*This material is based upon work supported by the National Science Foundation under Grant Nos. 0634957 and 0915487. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

Department of Computer Science, Yale University, New Haven, CT 06520-8285.
E-mail: daniel.spielman@yale.edu.

1. **Laplacian Matrices and Graphs.** We will consider weighted, undirected, simple graphs G given by a triple (V, E, w) , where V is a set of vertices, E is a set of edges, and w is a weight function that assigns a positive weight to every edge. The Laplacian matrix L of a graph is most naturally defined by the quadratic form it induces. For a vector $\mathbf{x} \in \mathbb{R}^V$, the Laplacian quadratic form of G is

$$\mathbf{x}^T L \mathbf{x} = \sum_{(u,v) \in E} w_{u,v} (\mathbf{x}(u) - \mathbf{x}(v))^2.$$

Thus, L provides a measure of the smoothness of \mathbf{x} over the edges in G . The more \mathbf{x} jumps over an edge, the larger the quadratic form becomes.

The Laplacian L also has a simple description as a matrix. Define the weighted degree of a vertex u by

$$d(u) = \sum_{v \in V} w_{u,v}.$$

Define D to be the diagonal matrix whose diagonal contains d , and define the weighted adjacency matrix of G by

$$A(u, v) = \begin{cases} w_{u,v} & \text{if } (u, v) \in E \\ 0 & \text{otherwise.} \end{cases}$$

We have

$$L = D - A.$$

It is often convenient to consider the *normalized Laplacian* of a graph instead of the Laplacian. It is given by $D^{-1/2} L D^{-1/2}$, and is more closely related to the behavior of random walks.

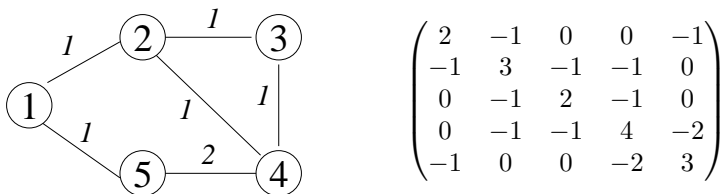


Figure 1. A Graph on five vertices and its Laplacian matrix. The weights of edges are indicated by the numbers next to them. All edges have weight 1, except for the edge between vertices 4 and 5 which has weight 2.

2. **Cuts in Graphs.** A large amount of algorithmic research is devoted to finding algorithms for partitioning the vertices and edges of graphs (see [LR99, ARV09, GW95, Kar00]). Given a set of vertices $S \subset V$, we

define the boundary of S , written $\partial(S)$ to be the set of edges of G with exactly one vertex in S .

For a subset of vertices S , let $\chi_S \in \mathbb{R}^V$ denote the characteristic vector of S (one on S and zero outside). If all edge weights are 1, then $\chi_S^T L \chi_S$ equals the number of edges in $\partial(S)$. When the edges are weighted, it measures the sum of their weights.

Computer Scientists are often interested in finding the sets of vertices S that minimize or maximize the size of the boundary of S . In this survey, we will be interested in the sets of vertices that minimize the size of $\partial(S)$ divided by a measure of the size of S . When we measure the number of vertices in S , we obtain the *isoperimetric number* of S ,

$$i(S) \stackrel{\text{def}}{=} \frac{|\partial(S)|}{\min(|S|, |V-S|)}.$$

If we instead measure the S by the weight of its edges, we obtain the *conductance* of S , which is given by

$$\phi(S) \stackrel{\text{def}}{=} \frac{w(\partial(S))}{\min(d(S), d(V-S))},$$

where $d(S)$ is the sum of the weighted degrees of vertices in the set S and $w(\partial(S))$ is the sum of the weights of the edges on the boundary of S . The isoperimetric number of a graph and the conductance of a graph are defined to be the minima of these quantities over subsets of vertices:

$$i_G \stackrel{\text{def}}{=} \min_{S \subset V} i(S) \quad \text{and} \quad \phi_G \stackrel{\text{def}}{=} \min_{S \subset V} \phi(S).$$

It is often useful to divide the vertices of a graph into two pieces by finding a set S of low isoperimetric number or conductance, and then partitioning the vertices according to whether or not they are in S .

3. **Expander Graphs.** Expander graphs are the regular, unweighted graphs having high isoperimetric number and conductance. Formally, a sequence of graphs is said to be a sequence of expander graphs if all of the graphs in the sequence are regular of the same degree and there exists a constant $\alpha > 0$ such that $\phi_G > \alpha$ for all graphs G in the family. The higher α , the better.

Expander graphs pop up all over Theoretical Computer Science (see [HLW06]), and are examples one should consider whenever thinking about graphs.

4. **Cheeger's Inequality.** The discrete versions of Cheeger's inequality [Che70] relate quantities like the isoperimetric number and the conductance of a graph to the eigenvalues of the Laplacian and the normalized

Laplacian. The smallest eigenvalue of the Laplacian and the normalized Laplacian is always zero, and it has multiplicity 1 for a connected graph. The discrete versions of Cheeger’s inequality (there are many, see [LS88, AM85, Alo86, Dod84, Var85, SJ89]) concern the smallest non-zero eigenvalue, which we denote λ_2 . For example, we will exploit the tight connection between conductance and the smallest non-zero eigenvalue of the normalized Laplacian:

$$2\phi_G \geq \lambda_2(D^{-1/2}LD^{-1/2}) \geq \phi_G^2/2.$$

The time required for a random walk on a graph to mix is essentially the reciprocal of $\lambda_2(D^{-1/2}LD^{-1/2})$. Sets of vertices of small conductance are obvious obstacles to rapid mixing. Cheeger’s inequality tells us that they are the main obstacle. It also tells us that all non-zero eigenvalues of expander graphs are bounded away from zero. Indeed, expander graphs are often characterized by the gap between their Laplacian eigenvalues and zero.

5. **The Condition Number of a Matrix.** The *condition number* of a symmetric matrix, written $\kappa(A)$, is given by

$$\kappa(A) \stackrel{\text{def}}{=} \lambda_{\max}(A)/\lambda_{\min}(A),$$

where $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the largest and smallest eigenvalues of A (for general matrices, we measure the singular values instead of the eigenvalues). For singular matrices, such as Laplacians, we instead measure the *finite condition number*, $\kappa_f(A)$, which is the ratio between the largest and smallest non-zero eigenvalues.

The condition number is a fundamental object of study in Numerical Linear Algebra. It tells us how much the solution to a system of equations in A can change when one perturbs A , and it may be used to bound the rate of convergence of iterative algorithms for solving linear equations in A . From Cheeger’s inequality, we see that expander graphs are exactly the graphs whose Laplacian matrices have low condition number. Formally, families of expanders may be defined by the condition that there is an absolute constant c such that $\kappa_f(G) \leq c$ for all graphs in the family.

Spectrally speaking, the best expander graphs are the Ramanujan Graphs [LPS88, Mar88], which are d -regular graphs for which

$$\kappa_f(G) \leq \frac{d + 2\sqrt{d-1}}{d - 2\sqrt{d-1}}.$$

As d grows large, this bound quickly approaches 1.

6. **Random Matrix Theory.** Researchers in random matrix theory are particularly concerned with the singular values and eigenvalues of random

matrices. Researchers in Computer Science often exploit results from this field, and study random matrices that are obtained by down-sampling other matrices [AM07, FK99]. We will be interested in the Laplacian matrices of randomly chosen subgraphs of a given graph.

7. **Spanning Trees.** A tree is a connected graph with no cycles. As trees are simple and easy to understand, it often proves useful to approximate a more complex graph by a tree (see [Bar96, Bar98, FRT04, ACF⁺04]). A spanning tree T of a graph G is a tree that connects all the vertices of G and whose edges are a subset of the edges of G . Many varieties of spanning trees are studied in Computer Science, including maximum-weight spanning trees, random spanning trees, shortest path trees, and low-stretch spanning trees. I find it amazing that spanning trees should have anything to do with solving systems of linear equations.

This survey begins with an explanation of where Laplacian matrices come from, and gives some reasons they appear in systems of linear equations. We then briefly explore some of the popular approaches to solving systems of linear equations, quickly jumping to preconditioned iterative methods. These methods solve linear equations in a matrix A by multiplying vectors by A and solving linear equations in another matrix, called a *preconditioner*. These methods work well when the preconditioner is a good approximation for A and when linear equations in the preconditioner can be solved quickly. We will precondition Laplacian matrices of graphs by Laplacian matrices of other graphs (usually subgraphs), and will use tools from graph theory to reason about the quality of the approximations and the speed of the resulting linear equation solvers. In the end, we will see that linear equations in any Laplacian matrix can be solved to accuracy ϵ in time

$$O((m + n \log n (\log \log n)^2) \log \epsilon^{-1}),$$

if one allows polynomial time to precompute the preconditioners. Here n is the dimension and m is the number of non-zeros in the matrix. When m is much less than n^2 , this is less time than would be required to even read the inverse of a general n -by- n matrix.

The best balance we presently know between the complexity of computing the preconditioners and solving the linear equations yields an algorithm of complexity

$$O(m \log^c n \log 1/\epsilon),$$

for some large constant c . We hope this becomes a small constant, say 1 or 2, in the near future (In fact, it just did [KMP10]).

Highlights of this story include a definition of what it means to approximate one graph by another, a proof that every graph can be approximated by a sparse graph, an examination of which trees best approximate a given graph, and local algorithms for finding clusters of vertices in graphs.

2. Laplacian Matrices

Laplacian matrices of graphs are symmetric, have zero row-sums, and have non-positive off-diagonal entries. We call any matrix that satisfies these properties a Laplacian matrix, as there always exists some graph for which it is the Laplacian.

We now briefly list some applications in which the Laplacian matrices of graphs arise.

1. **Regression on Graphs.** Imagine that you have been told the value of a function \mathbf{f} on a subset W of the vertices of G , and wish to estimate the values of \mathbf{f} at the remaining vertices. Of course, this is not possible unless \mathbf{f} respects the graph structure in some way. One reasonable assumption is that the quadratic form in the Laplacian is small, in which case one may estimate \mathbf{f} by solving for the function $\mathbf{f} : V \rightarrow \mathbb{R}$ minimizing $\mathbf{f}^T L \mathbf{f}$ subject to \mathbf{f} taking the given values on W (see [ZGL03]). Alternatively, one could assume that the value of \mathbf{f} at every vertex v is the weighted average of \mathbf{f} at the neighbors of v , with the weights being proportional to the edge weights. In this case, one should minimize

$$\|D^{-1}L\mathbf{f}\|$$

subject to \mathbf{f} taking the given values on W . These problems inspire many uses of graph Laplacians in Machine Learning.

2. **Spectral Graph Theory.** In Spectral Graph Theory, one studies graphs by examining the eigenvalues and eigenvectors of matrices related to these graphs. Fiedler [Fie73] was the first to identify the importance of the eigenvalues and eigenvectors of the Laplacian matrix of a graph. The book of Chung [Chu97] is devoted to the Laplacian matrix and its normalized version.
3. **Solving Maximum Flow by Interior Point Algorithms.** The Maximum Flow and Minimum Cost Flow problems are specific linear programming problems that arise in the study of network flow. If one solves these linear programs by interior point algorithms, then the interior point algorithms will spend most of their time solving systems of linear equations that can be reduced to restricted Laplacian systems. We refer the reader who would like to learn more about these reductions to one of [DS08, FG07].
4. **Resistor Networks.** The Laplacian matrices of graphs arise when one models electrical flow in networks of resistors. The vertices of a graph correspond to points at which we may inject or remove current and at which we will measure potentials. The edges correspond to resistors, with the weight of an edge being the reciprocal of its resistance. If $\mathbf{p} \in \mathbb{R}^V$

denotes the vector of potentials and $\mathbf{i}_{ext} \in \mathbb{R}^V$ the vectors of currents entering and leaving vertices, then these satisfy the relation

$$L\mathbf{p} = \mathbf{i}_{ext}.$$

We exploit this formula to compute the effective resistance between pairs of vertices. The effective resistance between vertices u and v is the difference in potential one must impose between u and v to flow one unit of current from u to v . To measure this, we compute the vector \mathbf{p} for which $L\mathbf{p} = \mathbf{i}_{ext}$, where

$$\mathbf{i}_{ext}(x) = \begin{cases} 1 & \text{for } x = u, \\ -1 & \text{for } x = v, \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

We then measure the difference between $\mathbf{p}(u)$ and $\mathbf{p}(v)$.

5. **Partial Differential Equations.** Laplacian matrices often arise when one discretizes partial differential equations. For example, the Laplacian matrices of path graphs naturally arise when one studies the modes of vibrations of a string. Another natural example appears when one applies the finite element method to solve Laplace's equation in the plane using a triangulation with no obtuse angles (see [Str86, Section 5.4]). Boman, Hendrickson and Vavasis [BHV08] have shown that the problem of solving general elliptic partial differential equations by the finite element method can be reduced to the problem of solving linear equations in restricted Laplacian matrices.

Many of these applications require the solution of linear equations in Laplacian matrices, or their restrictions. If the values at some vertices are restricted, then the problem in the remaining vertices becomes one of solving a linear equation in a diagonally dominant symmetric M -matrix. Such a matrix is called a Stieltjes matrix, and may be expressed as a Laplacian plus a non-negative diagonal matrix. A Laplacian is always positive semi-definite, and if one adds a non-negative non-zero diagonal matrix to the Laplacian of a connected graph, the result will always be positive definite. The problem of computing the smallest eigenvalues and corresponding eigenvectors of a Laplacian matrix is often solved by the repeated solution of linear equations in that matrix.

3. Solving Linear Equations in Laplacian Matrices

There are two major approaches to solving linear equations in Laplacian matrices. The first are direct methods. These are essentially variants of Gaussian

elimination, and lead to exact solutions. The second are the iterative (indirect) methods. These provide successively better approximations to a system of linear equations, typically requiring a number of iterations proportional to $\log \epsilon^{-1}$ to achieve accuracy ϵ .

3.1. Direct Methods. When one applies Gaussian Elimination to a matrix A , one produces a factorization of A in the form LU where U is an upper-triangular matrix and L is a lower-triangular matrix with 1s on the diagonal. Such a factorization allows one to easily solve linear equations in a matrix A , as one can solve a linear equation in an upper- or lower-triangular matrix in time proportional to its number of non-zero entries. When solving equations in symmetric positive-definite matrices, one uses the more compact Cholesky factorization which has the form LL^T , where L is a lower-triangular matrix. If you are familiar with Gaussian elimination, then you can understand Cholesky factorization as doing the obvious elimination to preserve symmetry: every row-elimination is followed by the corresponding column-elimination. While Laplacian matrices are not positive-definite, one can use essentially the same algorithm if one stops when the remaining matrix has dimension 2.

When applying Cholesky factorization to positive definite matrices one does not have to permute rows or columns to avoid having pivots that are zero [GL81]. However, the choice of which row and column to eliminate can have a big impact on the running time of the algorithm. Formally speaking, the choice of an elimination ordering corresponds to the choice of a permutation matrix P for which we factor $PAP^T = LL^T$. By choosing an elimination ordering carefully, one can sometimes find a factorization of the form LL^T in which L is very sparse and can be computed quickly. For the Laplacian matrices of graphs, this process has a very clean graph theoretic interpretation. The rows and columns correspond to vertices. When one eliminates the row and column corresponding to a vertex, the resulting matrix is the Laplacian of a graph in which that vertex has been removed, but in which all of its neighbors have been connected. The weights with which they are connected naturally depend upon the weights with which they were connected to the eliminated vertex. Thus, we see that the number of entries in L depends linearly on the sum of the degrees of vertices when they are eliminated, and the time to compute L depends upon the sum of the squares of the degrees of eliminated vertices.

For example, if G is a path graph then its Laplacian will be tri-diagonal. A vertex at the end of the path has degree 1, and its elimination results in a path that is shorter by one. Thus, one may produce a Cholesky factorization of a path graph with at most $2n$ non-zero entries in time $O(n)$. One may do the same if G is a tree: a tree always has a vertex of degree 1, and its elimination results in a smaller tree. Even when dealing with a graph that is not a tree, similar ideas may be applied. Many practitioners use the Minimum Degree Ordering [TW67] or the Approximate Minimum Degree Ordering [ADD96] in

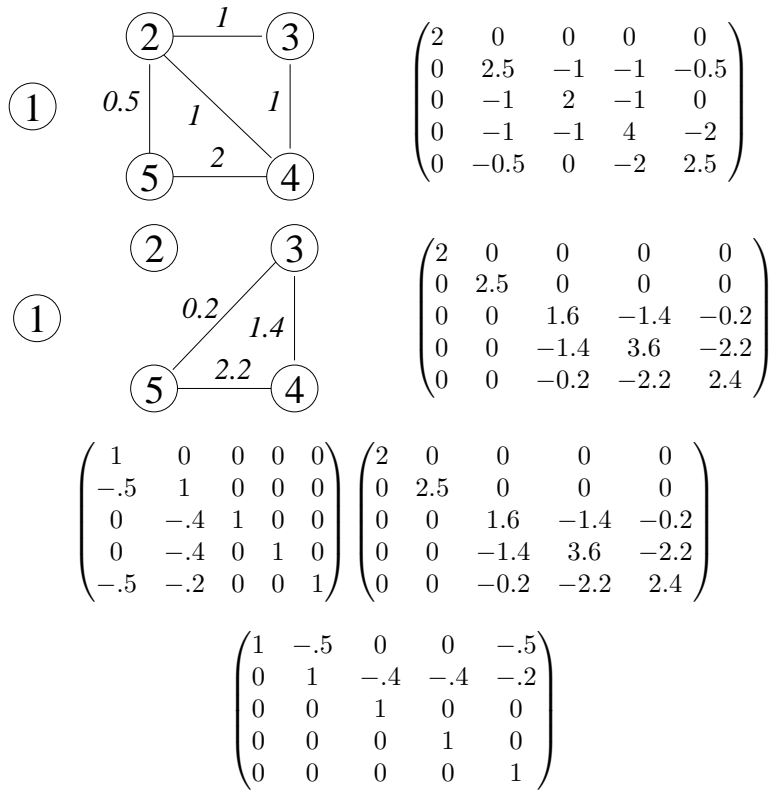


Figure 2. The first line depicts the result of eliminating vertex 1 from the graph in Figure 1. The second line depicts the result of also eliminating vertex 2. The third line presents the factorization of the Laplacian produced so far.

an attempt to minimize the number of non-zero entries in L and the time required to compute it.

For graphs that can be disconnected into pieces of approximately the same size without removing too many vertices, one can find orderings that result in lower-triangular factors that are sparse. For example, George’s Nested Dissection Algorithm [Geo73] can take as input the Laplacian of a weighted \sqrt{n} -by- \sqrt{n} grid graph and output a lower-triangular factorization with $O(n \log n)$ non-zero entries in time $O(n^{3/2})$. This algorithm was generalized by Lipton, Rose and Tarjan to apply to any planar graph [LRT79]. They also proved the more general result that if G is a graph such that all subgraphs of G having k vertices can be divided into pieces of size at most αk (for some constant $\alpha < 1$) by the removal of at most $O(k^\sigma)$ vertices (for $\sigma > 1/2$), then the Laplacian of G has a lower-triangular factorization with at most $O(n^{2\sigma})$ non-zero entries that can be computed in time $O(n^{3\sigma})$. For example, this would hold if k -vertex subgraph of G has isoperimetric number at most $O(k^\sigma)$.

Of course, one can also use Fast Matrix Inversion [CW82] to compute the inverse of a matrix in time approximately $O(n^{2.376})$. This approach can also be used to accelerate the computation of LL^T factorizations in the algorithms of Lipton, Rose and Tarjan (see [LRT79] and [BH74]).

3.2. Iterative Methods. Iterative algorithms for solving systems of linear equations produce successively better approximate solutions. The most fundamental of these is the Conjugate Gradient algorithm. Assume for now that we wish to solve the linear system

$$A\mathbf{x} = \mathbf{b},$$

where A is a symmetric positive definite matrix. In each iteration, the Conjugate Gradient multiplies a vector by A . The number of iterations taken by the algorithm may be bounded in terms of the eigenvalues of A . In particular, the Conjugate Gradient is guaranteed to produce an ϵ -approximate solution $\tilde{\mathbf{x}}$ in at most $O(\sqrt{\kappa_f(A)} \log(1/\epsilon))$ iterations, where we say that $\tilde{\mathbf{x}}$ is an ϵ -approximate solution if

$$\|\tilde{\mathbf{x}} - \mathbf{x}\|_A \leq \epsilon \|\mathbf{x}\|_A,$$

where \mathbf{x} is the actual solution and

$$\|\mathbf{x}\|_A = \sqrt{\mathbf{x}^T A \mathbf{x}}.$$

The Conjugate Gradient algorithm thereby reduces the problem of solving a linear system in A to the application of many multiplications by A . This can produce a significant speed improvement when A is sparse.

One can show that the Conjugate Gradient algorithm will never require more than n iterations to compute the exact solution (if one uses with exact arithmetic). Thus, if A has m non-zero entries, the Conjugate Gradient will produce solutions to systems in A in time at most $O(mn)$. In contrast, Lipton, Rose and Tarjan [LRT79] prove that if A is the Laplacian matrix of a good expander having $m = O(n)$ edges, then under every ordering the lower-triangular factor of the Laplacian has almost $n^2/2$ non-zero entries and requires almost $n^3/6$ operations to compute by the Cholesky factorization algorithm.

While Laplacian matrices are always singular, one can apply the Conjugate Gradient to solve linear equations in these matrices with only slight modification. In this case, we insist that \mathbf{b} be in the range of the matrix. This is easy to check as the null space of the Laplacian of a connected graph is spanned by the all-ones vector. The same bounds on the running time of the Conjugate Gradient then apply. Thus, when discussing Laplacian matrices we will use the *finite condition number* κ_f , which measures the largest eigenvalue divided by the smallest non-zero eigenvalue.

Recall that expander graphs have low condition number, and so linear equations in the Laplacians of expanders can be solved quickly by the Conjugate Gradient. On the other hand, if a graph and all of its subgraphs have cuts of small isoperimetric number, then one can apply Generalized Nested Dissection [LRT79] to solve linear equations in its Laplacian quickly. Intuitively, this tells us that it should be possible to solve every Laplacian system quickly as it seems that either the Conjugate Gradient or Cholesky factorization with the appropriate ordering should be fast. While this argument cannot be made rigorous, it does inform our design of a fast algorithm.

3.3. Preconditioned Iterative Methods. Iterative methods can be greatly accelerated through the use of preconditioning. A good preconditioner for a matrix A is another matrix B that approximates A and such that it is easy to solve systems of linear equations in B . A preconditioned iterative solver uses solutions to linear equations in B to obtain accurate solutions to linear equations in A .

For example, in each iteration the Preconditioned Conjugate Gradient (PCG) solves a system of linear equations in B and multiplies a vector by A . The number of iterations the algorithm needs to find an ϵ -accurate solution to a system in A may be bounded in terms of the *relative condition number* of A with respect to B , written $\kappa(A, B)$. For symmetric positive definite matrices A and B , this may be defined as the ratio of the largest to the smallest eigenvalue of AB^{-1} . One can show that PCG will find an ϵ -accurate solution in at most $O(\sqrt{\kappa(A, B)} \log \epsilon^{-1})$ iterations. Tighter bounds can sometimes be proved if one knows more about the eigenvalues of AB^{-1} . Preconditioners have proved incredibly useful in practice. For Laplacians, incomplete Cholesky factorization preconditioners [MV77] and Multigrid preconditioners [BHM01] have proved particularly useful.

The same analysis of the PCG applies when A and B are Laplacian matrices of connected graphs, but with $\kappa_f(A, B)$ measuring the ratio of the largest to smallest non-zero eigenvalue of AB^+ , where B^+ is the Moore-Penrose pseudoinverse of B . We recall that for a symmetric matrix B with spectral decomposition

$$B = \sum_i \lambda_i \mathbf{v}_i \mathbf{v}_i^T,$$

the pseudoinverse of B is given by

$$B^+ = \sum_{i:\lambda_i \neq 0} \frac{1}{\lambda_i} \mathbf{v}_i \mathbf{v}_i^T.$$

That is, B projects a vector onto the image of A and then acts as the inverse of A on its image. When A and B are the Laplacian matrices of graphs, we will view $\kappa_f(A, B)$ as a measure of how well those graphs approximate one another.

4. Approximation by Sparse Graphs

Sparsification is the process of approximating a given graph G by a sparse graph H . We will say that H is an α -approximation of G if

$$\kappa_f(L_G, L_H) \leq 1 + \alpha, \tag{1}$$

where L_G and L_H are the Laplacian matrices of G and H . This tells us that G and H are similar in many ways. In particular, they have similar eigenvalues and the effective resistances in G and H between every pair of nodes is approximately the same.

The most obvious way that sparsification can accelerate the solution of linear equations is by replacing the problem of solving systems in dense matrices by the problem of solving systems in sparse matrices. Recall that the Conjugate Gradient, used as a direct solver, can solve systems in n -dimensional matrices with m non-zero entries in time $O(mn)$. So, if we could find a graph H with $O(n)$ non-zero entries that was even a 1-approximation of G , then we could quickly solve systems in L_G by using the Preconditioned Conjugate Gradient with L_H as the preconditioner, and solving the systems in L_H by the Conjugate Gradient. Each solve in H would then take time $O(n^2)$, and the number of iterations of the PCG required to get an ϵ -accurate solution would be $O(\log \epsilon^{-1})$. So, the total complexity would be

$$O((m + n^2) \log \epsilon^{-1}).$$

Sparsifiers are also employed in the fastest algorithms for solving linear equations in Laplacians, as we will later see in Section 7.

But, why should we believe that such good sparsifiers should exist? We believed it because Benczur and Karger [BK96] developed something very similar in their design of fast algorithms for the minimum cut problem. Benczur and Karger proved that for every graph G there exists a graph H with $O(n \log n / \alpha^2)$ edges such that the weight of every cut in H is approximately the same as in G . This could either be expressed by writing

$$w(\delta_H(S)) \leq w(\delta_G(S)) \leq (1 + \alpha)w(\delta_H(S)), \quad \text{for every } S \subset V,$$

or by

$$\chi_S^T L_H \chi_S \leq \chi_S^T L_G \chi_S \leq (1 + \alpha) \chi_S^T L_H \chi_S, \quad \text{for every } \chi_S \in \{0, 1\}^V. \tag{2}$$

A sparsifier H satisfies (1) if it satisfies (2) for all vectors in \mathbb{R}^V , rather than just $\{0, 1\}^V$. To distinguish Benczur and Karger's type of sparsifiers from those we require, we call their sparsifiers *cut sparsifiers* and ours *spectral sparsifiers*.

Benczur and Karger proved their sparsification theorem by demonstrating that if one forms H at random by choosing each edge of G with an appropriate probability, and re-scales the weights of the chosen edges, then the resulting

graph probably satisfies their requirements. Spielman and Srivastava [SS10a] prove that a different choice probabilities results in spectral sparsifiers that also have $O(n \log n/\alpha^2)$ edges and are α -approximations of the original graph. The probability distribution turns out to be very natural: one chooses each edge with probability proportional to the product of its weight with the effective resistance between its endpoints. After some linear algebra, their theorem follows from the following result of Rudelson and Vershynin [RV07] that lies at the intersection of functional analysis with random matrix theory.

Lemma 4.1. *Let $\mathbf{y} \in \mathbb{R}^n$ be a random vector for which $\|\mathbf{y}\| \leq M$ and*

$$\mathbf{E} [\mathbf{y}\mathbf{y}^T] = I.$$

Let $\mathbf{y}_1, \dots, \mathbf{y}_k$ be independent copies of \mathbf{y} . Then,

$$\mathbf{E} \left[\left\| \frac{1}{k} \sum_{i=1}^k \mathbf{y}_i \mathbf{y}_i^T - I \right\| \right] \leq C \frac{\sqrt{\log k}}{\sqrt{k}} M,$$

for some absolute constant C , provided that the right hand side is at most 1.

For computational purposes, the drawback of the algorithm of Spielman and Srivastava is that it requires knowledge of the effective resistances of all the edges in the graph. While they show that it is possible to approximately compute all of these at once in time $m \log^{O(1)} n$, this computation requires solving many linear equations in the Laplacian of the matrix to be sparsified. So, it does not help us solve linear equations quickly. We now examine two directions in which sparsification has been improved: the discovery of sparsifiers with fewer edges and the direct construction of sparsifiers in nearly-linear time.

4.1. Sparsifiers with a linear number of edges. Batson, Spielman and Srivastava [BSS09] prove that for every weighted graph G and every $\beta > 0$ there is a weighted graph H with at most $\lceil n/\beta^2 \rceil$ edges for which

$$\kappa_f(L_G, L_H) \leq \left(\frac{1 + \beta}{1 - \beta} \right)^2.$$

For $\beta < 1/10$, this means that H is a $1 + 5\beta$ approximation of G . Thus, every Laplacian can be well-approximated by a Laplacian with a linear number of edges. Such approximations were previously known to exist for special families of graphs. For example, Ramanujan expanders [LPS88, Mar88] are optimal sparse approximations of complete graphs.

Batson, Spielman and Srivastava [BSS09] prove this result by reducing it to the following statement about vectors in isotropic position.

Theorem 4.2. *Let v_1, \dots, v_m be vectors in \mathbb{R}^n such that*

$$\sum_i v_i v_i^T = I.$$

For every $\beta > 0$ there exist scalars $s_i \geq 0$, at most n/β^2 of which are non-zero, such that

$$\kappa \left(\sum_i s_i \mathbf{v}_i \mathbf{v}_i^T \right) \leq \left(\frac{1 + \beta}{1 - \beta} \right)^2.$$

This theorem may be viewed as an extension of Rudelson’s lemma. It does not concern random sets of vectors, but rather produces one particular set. By avoiding the use of random vectors, it is possible to produce a set of $O(n)$ vectors instead of $O(n \log n)$. On the other hand, these vectors now appear with coefficients s_i . We believe that these coefficients are unnecessary if all the vectors \mathbf{v}_i have the same norm. However, this statement may be non-trivial to prove as it would imply Weaver’s conjecture KS_2 , and thereby the Kadison-Singer conjecture [Wea04].

The proof of Theorem 4.2 is elementary. It involves choosing the coefficient of one vector at a time. Potential functions are introduced to ensure that progress is being made. Success is guaranteed by proving that at every step there is a vector whose coefficient can be made non-zero without increasing the potential functions. The technique introduced in this argument has also been used [SS10b] to derive an elementary proof of Bourgain and Tzafiri’s restricted invertibility principle [BT87].

4.2. Nearly-linear time computation. Spielman and Teng [ST08b] present an algorithm that takes time $O(m \log^{13} n)$ and produces ϵ sparsifiers with $O(n \log^{29} n / \epsilon^2)$ edges. While this algorithm takes nearly-linear time and is asymptotically faster than any algorithm taking time $O(m^c)$ for any $c > 1$, it is too slow to be practical. Still, it is the asymptotically fastest algorithm for producing sparsifiers that we know so far. The algorithm relies upon other graph theoretic algorithms that are interesting in their own right.

The key insight in the construction of [ST08b] is that if G has high conductance, then one can find a good sparsifier of G through a very simple random sampling algorithm. On the other hand, if G does not have high conductance then one can partition the vertices of G into two parts without removing too many edges. By repeatedly partitioning in this way, one can divide any dense graph into parts of high conductance while removing only a small fraction of its edges (see also [Tre05] and [KVV04]). One can then produce a sparsifier by randomly sampling edges from the components of high conductance, and by recursively sparsifying the remaining edges.

However, in order to make such an algorithm fast, one requires a way of quickly partitioning a graph into subgraphs of high conductance without removing too many edges. Unfortunately, we do not yet know how to do this.

Problem 1. *Design a nearly-linear time algorithm that partitions the vertices of a graph G into sets V_1, \dots, V_k so that the conductance of the induced graph on each set V_i is high (say $\Omega(1/\log n)$) and at most half of the edges of G have endpoints in different components.*

Instead, Spielman and Teng [ST08b] show that the result of $O(\log n)$ iterations of repeated approximate partitioning suffice for the purposes of sparsification.

This leaves the question of how to approximately partition a graph in nearly-linear time. Spielman and Teng [ST08a] found such an algorithm by designing an algorithm for the local clustering problem, which we describe further in Section 8.

Problem 2. *Design an algorithm that on input a graph G and an $\alpha \leq 1$ produces an α -approximation of G with $O(n/\alpha^2)$ edges in time $O(m \log n)$.*

5. Subgraph Preconditioners and Support Theory

The breakthrough that led to the work described in the rest of this survey was Vaidya's idea of preconditioning Laplacian matrices of graphs by the Laplacians of subgraphs of those graphs [Vai90]. The family of preconditioners that followed have been referred to as subgraph or combinatorial preconditioners, and the tools used to analyze them are known as "support theory".

Support theory uses combinatorial techniques to prove inequalities on the Laplacian matrices of graphs. Given positive semi-definite matrices A and B , we write

$$A \succcurlyeq B$$

if $A - B$ is positive semi-definite. This is equivalent to saying that for all $\mathbf{x} \in \mathbb{R}^V$

$$\mathbf{x}^T A \mathbf{x} \succcurlyeq \mathbf{x}^T B \mathbf{x}.$$

Boman and Hendrickson [BH03] show that if $\sigma_{A,B}$ and $\sigma_{B,A}$ are the least constants such that

$$\sigma_{A,B} A \succcurlyeq B \quad \text{and} \quad \sigma_{B,A} B \succcurlyeq A,$$

then

$$\lambda_{\max}(AB^+) = \sigma_{B,A}, \quad \lambda_{\min}(AB^+) = \sigma_{A,B}, \quad \text{and} \quad \kappa(A, B) = \sigma_{A,B} \sigma_{B,A}.$$

Such inequalities are natural for the Laplacian matrices of graphs.

Let $G = (V, E, w)$ be a graph and $H = (V, F, w)$ be a subgraph, where we have written w in both to indicate that edges that appear in both G and H should have the same weights. Let L_G and L_H denote the Laplacian matrices

of these graphs. We then know that

$$\mathbf{x}^T L_G \mathbf{x} = \sum_{(u,v) \in E} w_{u,v} (\mathbf{x}(u) - \mathbf{x}(v))^2 \geq \sum_{(u,v) \in F} w_{u,v} (\mathbf{x}(u) - \mathbf{x}(v))^2 = \mathbf{x}^T L_H \mathbf{x}.$$

So, $L_G \succcurlyeq L_H$.

For example, Vaidya [Vai90] suggested preconditioning the Laplacian of graph by the Laplacian of a spanning tree. As we can use a direct method to solve linear equations in the Laplacians of trees in linear time, each iteration of the PCG with a spanning tree preconditioner would take time $O(m + n)$, where m is the number of edges in the original graph. In particular, Vaidya suggested preconditioning by the Laplacian of a maximum spanning tree. One can show that if T is a maximum spanning tree of G , then $(nm)L_T \succcurlyeq L_G$ (see [BGH⁺06] for details). While maximum spanning trees can be good preconditioners, this bound is not sufficient to prove it. From this bound, we obtain an upper bound of nm on the relative condition number, and thus a bound of $O(\sqrt{nm})$ on the number of iterations of PCG. However, we already know that PCG will not require more than n iterations. To obtain provably faster spanning tree preconditioners, we must measure their quality in a different way.

6. Low-stretch Spanning Trees

Boman and Hendrickson [BH01] recognized that for the purpose of preconditioning, one should measure the stretch of a spanning tree. The concept of the *stretch* of a spanning tree was first introduced by Alon, Karp, Peleg and West [AKPW95] in an analysis of algorithms for the k -server problem. However, it can be cleanly defined without reference that problem.

We begin by defining the stretch for graphs in which every edge has weight 1. If T is a spanning tree of $G = (V, E)$, then for every edge $(u, v) \in E$ there is a unique path in T connecting u to v . When all the weights in T and G are 1, the *stretch of (u, v) with respect to T* , written $st_T(u, v)$, is the number of edges in that path. The *stretch of G with respect to T* is then the sum of the stretches of all the edges in G :

$$st_T(G) = \sum_{(u,v) \in E} st_T(u, v).$$

For a weighted graph $G = (V, E, w)$ and spanning tree $T = (V, F, w)$, the stretch of an edge $e \in E$ with respect to T may be defined by assigning a length to every edge equal to the reciprocal of its weight. The stretch of an edge $e \in E$ is then just the length of the path in T between its endpoints divided by the length of e :

$$st_T(e) = w_e \left(\sum_{f \in P} \frac{1}{w_f} \right),$$

where P is the set of edges in the path in T from u to v . This may also be viewed as the effective resistance between u and v in T divided by the resistance of the edge e . To see this, recall that the resistances of edges are the reciprocals of their weights and that the effective resistance of a chain of resistors is the sum of their resistances.

Using results from [BH03], Boman and Hendrickson [BH01] proved that

$$\text{st}_T(G)L_T \succcurlyeq L_G.$$

Alon et al. [AKPW95] proved the surprising result that every weighted graph G has a spanning tree T for which

$$\text{st}_T(G) \leq m2^{O(\sqrt{\log n \log \log n})} \leq m^{1+o(1)},$$

where m is the number of edges in G . They also showed how to construct such a tree in time $O(m \log n)$. Using these low-stretch spanning trees as preconditioners, one can solve a linear system in a Laplacian matrix to accuracy ϵ in time

$$O(m^{3/2+o(1)} \log \epsilon^{-1}).$$

Presently the best construction of low-stretch spanning trees is that of Abraham, Bartal and Neiman [ABN08], who employ the star-decomposition of Elkin, Emek, Spielman and Teng [EEST08] to prove the following theorem.

Theorem 6.1. *Every weighted graph G has a spanning tree T such that*

$$\text{st}_T(G) \leq O(m \log n \log \log n (\log \log \log n)^3) \leq O(m \log n (\log \log n)^2)$$

where m is the number of edges G . Moreover, one can compute such a tree in time $O(m \log n + n \log^2 n)$.

This result is almost tight: one can show that there are graphs with $2n$ edges and no cycles of length less than $c \log n$ for some $c > 0$ (see [Mar82] or [Bol98, Section III.1]). For such a graph G and every spanning tree T ,

$$\text{st}_T(G) \geq \Omega(n \log n).$$

We ask if one can achieve this lower bound.

Problem 3. *Determine whether every weighted graph G has a spanning tree T for which*

$$\text{st}_T(G) \leq O(m \log n).$$

If so, find an algorithm that computes such a T in time $O(m \log n)$.

It would be particularly exciting to prove a result of this form with small constants.

Problem 4. *Is it true that every weighted graph G on n vertices has a spanning tree T such that*

$$\kappa_f(L_G, L_T) \leq O(n)?$$

It turns out that one can say much more about low-stretch spanning trees as preconditioners. Spielman and Woo [SW09] prove that $\text{st}_T(G)$ equals the trace of $L_G L_T^\dagger$. As the largest eigenvalue of $L_G L_T^\dagger$ is at most the trace, the bound on the condition number of the graph with respect to a spanning tree follows immediately. This bound proves useful in two other ways: it is the foundation of the best constructions of preconditioners, and it tells us that low-stretch spanning trees are even better preconditioners than we believed.

Once we know that $\text{st}_T(G)$ equals the trace of $L_G L_T^\dagger$, we know much more about the spectrum of $L_G L_T^\dagger$ than just lower and upper bounds on its smallest and largest eigenvalues. We know that $L_G L_T^\dagger$ cannot have too many large eigenvalues. In particular, we know that it has at most k eigenvalues larger than $\text{st}_T(G)/k$. Spielman and Woo [SW09] use this fact to prove that PCG actually only requires $O((\text{st}_T(G))^{1/3} \log 1/\epsilon)$ iterations. Kolla, Makarychev, Saberi and Teng [KMST09] observe that one could turn T into a much better preconditioner if one could just fix a small number of eigenvalues. We make their argument precise in the next section.

7. Ultra-sparsifiers

Perhaps because maximum spanning trees do not yield worst-case asymptotic improvements in the time required to solve systems of linear equations, Vaidya [Vai90] discovered ways of improving spanning tree preconditioners. He suggested augmenting a spanning tree preconditioner by adding $o(n)$ edges to it. In this way, one obtains a graph that looks mostly like a tree, but has a few more edges. We will see that it is possible to obtain much better preconditioners this way. It is intuitive that one could use this technique to find graphs with lower relative condition numbers. For example, if for every edge that one added to the tree one could “fix” one eigenvalue of $L_G L_T^\dagger$, then by adding $n^2/\text{st}_T(G)$ edges one could produce an augmented graph with relative condition number at most $(\text{st}_T(G)/n)^2$. We call a graph with $n + o(n)$ edges that provides a good approximation of G an *ultra-sparsifier* of G .

We must now address the question of how one would solve a system of linear equations in an ultra-sparsifier. As an ultra-sparsifier mostly looks like a tree, it must have many vertices of degree 1 and 2. Naturally, we use Cholesky factorization to eliminate all such nodes. In fact, we continue eliminating until no vertex of degree 1 or 2 remains. One can show that if the ultra-sparsifier has $n + t$ edges, then the resulting graph has at most $3t$ edges and vertices [ST09, Proposition 4.1]. If t is sufficiently small, we could solve this system directly either by Cholesky factorization or the Conjugate Gradient. As the matrix obtained after the elimination is still a Laplacian, we would do even better to

solve that system recursively. This approach was first taken by Joshi [Jos97] and Reif [Rei98]. Spielman and Teng [ST09, Theorem 5.5] prove that if one can find ultra-sparsifiers of every n -vertex graph with relative condition number $c\chi^2$ and at most $n + n/\chi$ edges, for some small constant c , then this recursive algorithm will solve Laplacian linear systems in time

$$O(m\chi \log 1/\epsilon).$$

Kolla et al. [KMST09] have recently shown that such ultrasparsifiers can be obtained from low-stretch spanning trees with

$$\chi = O(\text{st}_T(G)/n).$$

For graphs with $O(n)$ edges, this yields a Laplacian linear-equation solver with complexity

$$O(n \log n (\log \log n)^2 \log 1/\epsilon).$$

While the procedure of Kolla et al. for actually constructing the ultrasparsifiers is not nearly as fast, their result is the first to tell us that such good preconditioners exist. The next challenge is to construct them quickly.

The intuition behind the Kolla et al. construction of ultrasparsifiers is basically that explained in the first paragraph of this section. But, they cannot fix each eigenvalue of the low-stretch spanning tree by the addition of one edge. Rather, they must add a small constant number of edges to fix each eigenvalue. Their algorithm successively chooses edges to add to the low-stretch spanning tree. At each iteration, it makes sure that the edge it adds has the desired impact on the eigenvalues. Progress is measured by a refinement of the barrier function approach used by Batson, Spielman and Srivastava [BSS09] for constructing graph sparsifiers.

Spielman and Teng [ST09] obtained nearly-linear time constructions of ultra-sparsifiers by combining low-stretch spanning trees with nearly-linear time constructions of graph sparsifiers [ST08b]. They showed that in time $O(m \log^{c_1} n)$ one can produce graphs with $n + (m/k) \log^{c_2} n$ edges that k -approximate a given graph G having m edges, for some constants c_1 and c_2 . This construction of ultra-sparsifiers yielded the first nearly-linear time algorithm for solving systems of linear equations in Laplacian matrices. This has led to a search for even faster algorithms.

Two days before the day on which I submitted this paper, I was sent a paper by Koutis, Miller and Peng [KMP10] that makes tremendous progress on this problem. By exploiting low-stretch spanning trees and Spielman and Srivastava's construction of sparsifiers, they produce ultra-sparsifiers that lead to an algorithm for solving linear systems in Laplacians that takes time

$$O(m \log^2 n (\log \log n)^2 \log \epsilon^{-1}).$$

This is much faster than any algorithm known to date.

Problem 5. *Can one design an algorithm for solving linear equations in Laplacian matrices that runs in time $O(m \log n \log \epsilon^{-1})$ or even in time $O(m \log \epsilon^{-1})$?*

We remark that Koutis and Miller [KM07] have designed algorithms for solving linear equations in the Laplacians of planar graphs that run in time $O(m \log \epsilon^{-1})$.

8. Local Clustering

The problem of local graph clustering may be motivated by the following problem. Imagine that one has a massive graph, and is interested in finding a cluster of vertices near a particular vertex of interest. Here we will define a cluster to be a set of vertices of low conductance. We would like to do this without examining too many vertices of the graph. In particular, we would like to find such a small cluster while only examining a number of vertices proportional to the size of the cluster, if it exists.

Spielman and Teng [ST08a] introduced this problem for the purpose of designing fast graph partitioning algorithms. Their algorithm does not solve this problem for every choice of initial vertex. Rather, assuming that G has a set of vertices S of low conductance, they presented an algorithm that works when started from a random vertex v of S . It essentially does this by approximating the distribution of a random walk starting at v . Their analysis exploited an extension of the connection between the mixing rate of random walks and conductance established by Lovász and Simonovits [LS93]. Their algorithm and analysis was improved by Andersen, Chung and Lang [ACL06], who used approximations of the Personal PageRank vector instead of random walks and also analyzed these using the technique of Lovász and Simonovits [LS93].

So far, the best algorithm for this problem is that of Andersen and Peres [AP09]. It is based upon the volume-biased evolving set process [MP03]. Their algorithm satisfies the following guarantee. If it is started from a random vertex in a set of conductance ϕ , it will output a set of conductance at most $O(\phi^{1/2} \log^{1/2} n)$. Moreover, the running time of their algorithm is at most $O(\phi^{-1/2} \log^{O(1)} n)$ times the number of vertices in the set their algorithm outputs.

References

- [ABN08] I. Abraham, Y. Bartal, and O. Neiman. Nearly tight low stretch spanning trees. In *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*, pages 781–790, Oct. 2008.
- [ACF⁺04] Yossi Azar, Edith Cohen, Amos Fiat, Haim Kaplan, and Harald Rcke. Optimal oblivious routing in polynomial time. *Journal of Computer and System Sciences*, 69(3):383–394, 2004. Special Issue on STOC 2003.

- [ACL06] Reid Andersen, Fan Chung, and Kevin Lang. Local graph partitioning using pagerank vectors. In *FOCS '06: Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*, pages 475–486, Washington, DC, USA, 2006. IEEE Computer Society.
- [ADD96] Patrick R. Amestoy, Timothy A. Davis, and Iain S. Duff. An approximate minimum degree ordering algorithm. *SIAM Journal on Matrix Analysis and Applications*, 17(4):886–905, 1996.
- [AKPW95] Noga Alon, Richard M. Karp, David Peleg, and Douglas West. A graph-theoretic game and its application to the k -server problem. *SIAM Journal on Computing*, 24(1):78–100, February 1995.
- [Alo86] N. Alon. Eigenvalues and expanders. *Combinatorica*, 6(2):83–96, 1986.
- [AM85] Noga Alon and V. D. Milman. λ_1 , isoperimetric inequalities for graphs, and superconcentrators. *J. Comb. Theory, Ser. B*, 38(1):73–88, 1985.
- [AM07] Dimitris Achlioptas and Frank Mcsherry. Fast computation of low-rank matrix approximations. *J. ACM*, 54(2):9, 2007.
- [AP09] Reid Andersen and Yuval Peres. Finding sparse cuts locally using evolving sets. In *STOC '09: Proceedings of the 41st annual ACM symposium on Theory of computing*, pages 235–244, New York, NY, USA, 2009. ACM.
- [ARV09] Sanjeev Arora, Satish Rao, and Umesh Vazirani. Expander flows, geometric embeddings and graph partitioning. *J. ACM*, 56(2):1–37, 2009.
- [Bar96] Yair Bartal. Probabilistic approximation of metric spaces and its algorithmic applications. In *Proceedings of the 37th Annual Symposium on Foundations of Computer Science*, page 184. IEEE Computer Society, 1996.
- [Bar98] Yair Bartal. On approximating arbitrary metrics by tree metrics. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 161–168, 1998.
- [BGH⁺06] M. Bern, J. Gilbert, B. Hendrickson, N. Nguyen, and S. Toledo. Support-graph preconditioners. *SIAM Journal on Matrix Analysis and Applications*, 27(4):930–951, 2006.
- [BH74] James R. Bunch and John E. Hopcroft. Triangular factorization and inversion by fast matrix multiplication. *Mathematics of Computation*, 28(125):231–236, 1974.
- [BH01] Erik Boman and B. Hendrickson. On spanning tree preconditioners. Manuscript, Sandia National Lab., 2001.
- [BH03] Erik G. Boman and Bruce Hendrickson. Support theory for preconditioning. *SIAM Journal on Matrix Analysis and Applications*, 25(3):694–717, 2003.
- [BHM01] W. L. Briggs, V. E. Henson, and S. F. McCormick. *A Multigrid Tutorial, 2nd Edition*. SIAM, 2001.
- [BHV08] Erik G. Boman, Bruce Hendrickson, and Stephen Vavasis. Solving elliptic finite element systems in near-linear time with support preconditioners. *SIAM Journal on Numerical Analysis*, 46(6):3264–3284, 2008.

- [BK96] András A. Benczúr and David R. Karger. Approximating s-t minimum cuts in $O(n^2)$ time. In *Proceedings of The Twenty-Eighth Annual ACM Symposium On The Theory Of Computing (STOC '96)*, pages 47–55, May 1996.
- [Bol98] Béla Bollobás. *Modern graph theory*. Springer-Verlag, New York, 1998.
- [BSS09] Joshua D. Batson, Daniel A. Spielman, and Nikhil Srivastava. Twice-Ramanujan sparsifiers. In *Proceedings of the 41st Annual ACM Symposium on Theory of computing*, pages 255–262, 2009.
- [BT87] J. Bourgain and L. Tzafriri. Invertibility of “large” sumatriicies with applications to the geometry of banach spaces and harmonic analysis. *Israel Journal of Mathematics*, 57:137–224, 1987.
- [Che70] J. Cheeger. A lower bound for smallest eigenvalue of the Laplacian. In *Problems in Analysis*, pages 195–199, Princeton University Press, 1970.
- [Chu97] Fan R. K. Chung. *Spectral Graph Theory*. CBMS Regional Conference Series in Mathematics. American Mathematical Society, 1997.
- [CW82] D. Coppersmith and S. Winograd. On the asymptotic complexity of matrix multiplication. *SIAM Journal on Computing*, 11(3):472–492, August 1982.
- [Dod84] Jozef Dodziuk. Difference equations, isoperimetric inequality and transience of certain random walks. *Transactions of the American Mathematical Society*, 284(2):787–794, 1984.
- [DS08] Samuel I. Daitch and Daniel A. Spielman. Faster approximate lossy generalized flow via interior point algorithms. In *Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pages 451–460, 2008.
- [EEST08] Michael Elkin, Yuval Emek, Daniel A. Spielman, and Shang-Hua Teng. Lower-stretch spanning trees. *SIAM Journal on Computing*, 32(2):608–628, 2008.
- [FG07] A. Frangioni and C. Gentile. Prim-based support-graph preconditioners for min-cost flow problems. *Computational Optimization and Applications*, 36(2):271–287, 2007.
- [Fie73] M. Fiedler. Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal*, 23(98):298–305, 1973.
- [FK99] Alan Frieze and Ravi Kannan. Quick approximation to matrices and applications. *Combinatorica*, 19(2):175–220, 1999.
- [FRT04] Jittat Fakcharoenphol, Satish Rao, and Kunal Talwar. A tight bound on approximating arbitrary metrics by tree metrics. *Journal of Computer and System Sciences*, 69(3):485–497, 2004. Special Issue on STOC 2003.
- [Geo73] Alan George. Nested dissection of a regular finite element mesh. *SIAM Journal on Numerical Analysis*, 10(2):345–363, 1973.
- [GL81] J. A. George and J. W. H. Liu. *Computer Solution of Large Sparse Positive Definite Systems*. Prentice-Hall, Englewood Cliffs, NJ, 1981.
- [GW95] Michel X. Goemans and David P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145, 1995.

- [HLW06] Shlomo Hoory, Nathan Linial, and Avi Wigderson. Expander graphs and their applications. *Bulletin of the American Mathematical Society*, 43(4):439–561, 2006.
- [Jos97] Anil Joshi. *Topics in Optimization and Sparse Linear Systems*. PhD thesis, UIUC, 1997.
- [Kar00] David R. Karger. Minimum cuts in near-linear time. *J. ACM*, 47(1):46–76, 2000.
- [KM07] Ioannis Koutis and Gary L. Miller. A linear work, $o(n^{1/6})$ time, parallel algorithm for solving planar Laplacians. In *Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 1002–1011, 2007.
- [KMP10] Ioannis Koutis, Gary L. Miller, and Richard Peng. Approaching optimality for solving sdd systems. to appear on arXiv, March 2010.
- [KMST09] Alexandra Kolla, Yury Makarychev, Amin Saberi, and Shanghua Teng. Subgraph sparsification and nearly optimal ultrasparsifiers. *CoRR*, abs/0912.1623, 2009.
- [KVV04] Ravi Kannan, Santosh Vempala, and Adrian Vetta. On clusterings: Good, bad and spectral. *J. ACM*, 51(3):497–515, 2004.
- [LPS88] A. Lubotzky, R. Phillips, and P. Sarnak. Ramanujan graphs. *Combinatorica*, 8(3):261–277, 1988.
- [LR99] Tom Leighton and Satish Rao. Multicommodity max-flow min-cut theorems and their use in designing approximation algorithms. *Journal of the ACM*, 46(6):787–832, November 1999.
- [LRT79] Richard J. Lipton, Donald J. Rose, and Robert Endre Tarjan. Generalized nested dissection. *SIAM Journal on Numerical Analysis*, 16(2):346–358, April 1979.
- [LS88] Gregory F. Lawler and Alan D. Sokal. Bounds on the l^2 spectrum for Markov chains and Markov processes: A generalization of Cheeger’s inequality. *Transactions of the American Mathematical Society*, 309(2):557–580, 1988.
- [LS93] Lovasz and Simonovits. Random walks in a convex body and an improved volume algorithm. *RSA: Random Structures & Algorithms*, 4:359–412, 1993.
- [Mar82] G. A. Margulis. Graphs without short cycles. *Combinatorica*, 2:71–78, 1982.
- [Mar88] G. A. Margulis. Explicit group theoretical constructions of combinatorial schemes and their application to the design of expanders and concentrators. *Problems of Information Transmission*, 24(1):39–46, July 1988.
- [MP03] Ben Morris and Yuval Peres. Evolving sets and mixing. In *STOC ’03: Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, pages 279–286, New York, NY, USA, 2003. ACM.
- [MV77] J. A. Meijerink and H. A. van der Vorst. An iterative solution method for linear systems of which the coefficient matrix is a symmetric m -matrix. *Mathematics of Computation*, 31(137):148–162, 1977.

- [Rei98] John Reif. Efficient approximate solution of sparse linear systems. *Computers and Mathematics with Applications*, 36(9):37–58, 1998.
- [RV07] Mark Rudelson and Roman Vershynin. Sampling from large matrices: An approach through geometric functional analysis. *J. ACM*, 54(4):21, 2007.
- [SJ89] Alistair Sinclair and Mark Jerrum. Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation*, 82(1):93–133, July 1989.
- [SS10a] Daniel A. Spielman and Nikhil Srivastava. Graph sparsification by effective resistances. *SIAM Journal on Computing*, 2010. To appear.
- [SS10b] Daniel A. Spielman and Nikhil Srivastava. Title: An elementary proof of the restricted invertibility theorem. Available at <http://arxiv.org/abs/0911.1114>, 2010.
- [ST08a] Daniel A. Spielman and Shang-Hua Teng. A local clustering algorithm for massive graphs and its application to nearly-linear time graph partitioning. *CoRR*, abs/0809.3232, 2008. Available at <http://arxiv.org/abs/0809.3232>.
- [ST08b] Daniel A. Spielman and Shang-Hua Teng. Spectral sparsification of graphs. *CoRR*, abs/0808.4134, 2008. Available at <http://arxiv.org/abs/0808.4134>.
- [ST09] Daniel A. Spielman and Shang-Hua Teng. Nearly-linear time algorithms for preconditioning and solving symmetric, diagonally dominant linear systems. *CoRR*, abs/cs/0607105, 2009. Available at <http://www.arxiv.org/abs/cs.NA/0607105>.
- [Str86] Gilbert Strang. *Introduction to Applied Mathematics*. Wellesley-Cambridge Press, 1986.
- [SW09] Daniel A. Spielman and Jaehoo Woo. A note on preconditioning by low-stretch spanning trees. *CoRR*, abs/0903.2816, 2009. Available at <http://arxiv.org/abs/0903.2816>.
- [Tre05] Lucan Trevisan. Approximation algorithms for unique games. *Proceedings of the 46th Annual IEEE Symposium on Foundations of Computer Science*, pages 197–205, Oct. 2005.
- [TW67] W.F. Tinney and J.W. Walker. Direct solutions of sparse network equations by optimally ordered triangular factorization. *Proceedings of the IEEE*, 55(11):1801–1809, nov. 1967.
- [Vai90] Pravin M. Vaidya. Solving linear equations with symmetric diagonally dominant matrices by constructing good preconditioners. Unpublished manuscript UIUC 1990. A talk based on the manuscript was presented at the IMA Workshop on Graph Theory and Sparse Matrix Computation, October 1991, Minneapolis., 1990.
- [Var85] N. Th. Varopoulos. Isoperimetric inequalities and Markov chains. *Journal of Functional Analysis*, 63(2):215–239, 1985.

- [Wea04] Nik Weaver. The Kadison-Singer problem in discrepancy theory. *Discrete Mathematics*, 278(1–3):227–239, 2004.
- [ZGL03] Xiaojin Zhu, Zoubin Ghahramani, and John D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proc. 20th Int. Conf. on Mach. Learn.*, 2003.

The Unified Theory of Pseudorandomness[†]

Salil Vadhan*

Abstract

Pseudorandomness is the theory of efficiently generating objects that “look random” despite being constructed with little or no randomness. One of the achievements of this research area has been the realization that a number of fundamental and widely studied “pseudorandom” objects are all *almost equivalent* when viewed appropriately. These objects include pseudorandom generators, expander graphs, list-decodable error-correcting codes, averaging samplers, and hardness amplifiers. In this survey, we describe the connections between all of these objects, showing how they can all be cast within a single “list-decoding framework” that brings out both their similarities and differences.

Mathematics Subject Classification (2010). Primary 68Q01; Secondary 94B99, 68R01, 68Q87, 68-02.

Keywords. Pseudorandom generators, expander graphs, list decoding, error-correcting codes, samplers, randomness extractors, hardness amplification

1. Introduction

Pseudorandomness is the theory of efficiently generating objects that “look random” despite being constructed with little or no randomness. Over the past 25 years, it has developed into a substantial area of study, with significant implications for complexity theory, cryptography, algorithm design, combinatorics, and communications theory. One of the achievements of this line of work has been the realization that a number of fundamental and widely studied “pseudorandom” objects are all *almost equivalent* when viewed appropriately. These objects include:

[†]An earlier version of this article appeared in *SIGACT News* [Vad1].

*Supported by NSF grant CCF-0133096, ONR grant N00014-04-1-0478, and US-Israel BSF grant 2002246.

School of Engineering and Applied Sciences, Harvard University, 33 Oxford Street, Cambridge, MA 02138, USA. E-mail: salil@seas.harvard.edu.

Pseudorandom Generators These are procedures that stretch a short “seed” of truly random bits into a long string of “pseudorandom” bits that cannot be distinguished from truly random by any efficient algorithm. In this article, we focus on methods for constructing pseudorandom generators from boolean functions of high circuit complexity.

Expander Graphs Expanders are graphs that are sparse but nevertheless highly connected. There are many variants of expander graphs, but here we focus on the classical notion of *vertex expansion*, where every subset of not-too-many vertices has many neighbors in the graph.

Error-Correcting Codes These are methods for encoding messages so that even if many of the symbols are corrupted, the original message can still be recovered. Here we focus on *list decoding*, where there are so many corruptions that uniquely decoding the original message is impossible, but it is still may be possible to produce a short list of possible candidates.

Randomness Extractors These are procedures that extract almost uniformly distributed bits from sources of biased and correlated bits. Here we focus on extractors for general sources, where all we assume is a lower bound on the amount of “entropy” in the source and only get a single sample from the source. Extractors for such sources necessarily use a small number of additional truly random bits as a “seed” for extraction.

Samplers These are randomness-efficient methods for sampling elements of a large universe so that approximately the correct fraction of samples will land in any subset of the universe with high probability.

Hardness Amplifiers These are methods for converting worst-case hard boolean functions into ones that are average-case hard.

These objects are all “pseudorandom” in the sense that a randomly chosen object can be shown to have the desired properties with high probability, and the main goal is typically to find *explicit constructions* — ones that are deterministic and computationally efficient — achieving similar parameters. Each of these objects was introduced with a different motivation, and originally developed its own body of research. However, as mentioned above, research in the theory of pseudorandomness has uncovered intimate connections between all of them. In recent years, a great deal of progress has been made in understanding and constructing each of these objects by translating intuitions and techniques developed for one to the others.

The purpose of this survey is to present the connections between these objects in a single place, using a single language. Hopefully, this will make the connections more readily accessible and usable for non-experts and those familiar with some but not all of the objects at hand. In addition, it is also meant to clarify the *differences* between the objects, and explain why occasional claims of “optimal” constructions of one type of object do not always lead to improved constructions of the others.

Naturally, describing connections between six different notions in a short article makes it impossible to do justice to any of the objects in its own. Thus, for motivation, constructions, and applications, the reader is referred to existing surveys focused on the individual objects [CRT, Kab, HLW, Sud, Gur, NT, Sha, Gol1, Tre2] or the broader treatments of pseudorandomness in [Mil, Tre3, Gol3, AB, Vad2]. In particular, the monograph [Vad2] develops the subject in a way that emphasizes the connections described here.

The framework used in this survey extends to a number of other pseudorandom objects, such as “randomness condensers,” but we omit these extensions due to space constraints. (See [Vad2].)

Notation. For a natural number $N \in \mathbb{N}$, $[N]$ denotes the set $\{1, \dots, N\}$. For a discrete random variable X , $x \stackrel{R}{\leftarrow} X$ means that x is sampled according to X . For a set S , U_S is a random variable distributed uniformly over S . For convenience, we will sometimes write $x \stackrel{R}{\leftarrow} S$ as shorthand for $x \stackrel{R}{\leftarrow} U_S$. All logs are base 2.

We make extensive use of asymptotic notation. For a nonnegative function $f = f(x_1, \dots, x_k)$, we write $O(f)$ (resp., $\Omega(f)$) as shorthand for an unspecified nonnegative function $g = g(x_1, \dots, x_k)$ for which there is a constant $c > 0$ such that $g(x_1, \dots, x_k) \leq c \cdot f(x_1, \dots, x_k)$ (resp., $g(x_1, \dots, x_k) \geq c \cdot f(x_1, \dots, x_k)$) for all settings of x_1, \dots, x_k . We write $\text{poly}(f_1, \dots, f_t)$ for an unspecified function bounded by $(f_1 + \dots + f_t)^c + c$ for a positive constant c , and $\tilde{O}(f)$ for a function bounded by $f \cdot \text{poly}(\log f)$. For a nonnegative function $f(x)$ of one variable, we write $o(f)$ for an unspecified function $g(x)$ such that $\lim_{x \rightarrow \infty} g(x)/f(x) = 0$.

2. The Framework

As we will see, all of the objects we are discussing can be syntactically viewed as functions $\Gamma : [N] \times [D] \rightarrow [M]$. We will show how the defining properties of each of the objects can be cast in terms of the following notion.

Definition 1. For a function $\Gamma : [N] \times [D] \rightarrow [M]$, a set $T \subseteq [M]$, and an agreement parameter $\varepsilon \in [0, 1)$, we define

$$\text{LIST}_\Gamma(T, \varepsilon) = \{x \in [N] : \Pr[\Gamma(x, U_{[D]}) \in T] > \varepsilon\}$$

We also define $\text{LIST}_\Gamma(T, 1) = \{x \in [N] : \Pr[\Gamma(x, U_{[D]}) \in T] = 1\}$.

In general, it will be possible to characterize each of the pseudorandom objects by a condition of the following form:

“For every subset $T \in \mathcal{C}$, we have $|\text{LIST}_\Gamma(T, \varepsilon)| \leq K$.”

Here $\varepsilon \in [0, 1]$ and $K \in [0, N]$ will be parameters corresponding to the “quality” of the object, and we usually wish to minimize both. \mathcal{C} will be a class of subsets of $[M]$, sometimes governed by an additional “quality” parameter. Sometimes the requirement will be that the size of $\text{LIST}_\Gamma(T, \varepsilon)$ is *strictly* less than K , but this is just a matter of notation, amounting to replacing K in the above formulation by $\lceil K \rceil - 1$.

The notation “ $\text{LIST}_\Gamma(\cdot, \cdot)$ ” comes from the interpretation of list-decodable error-correcting codes in this framework (detailed in the next section), where T corresponds to a corrupted codeword and $\text{LIST}_\Gamma(T, \varepsilon)$ to the list of possible decodings. This list-decoding viewpoint turns out to be very useful for casting all of the objects in the same language. However, this is not the only way of looking at the objects, and indeed the power of the connections we describe in this survey comes from the variety of perspectives they provide. In particular, many of the connections were discovered through the study of randomness extractors, and extractors remain a powerful lens through which to view the area. The list-decoding view of extractors, and consequently of many of the other objects presented here, emerged through a sequence of works, and was crystallized in paper of Ta-Shma and Zuckerman [TZ].

Our notation (e.g. the parameters N, M, D, K, ε) follows the literature on extractors, and thus is nonstandard for some of the objects. We also follow the convention from the extractor literature that $n = \log N$, $d = \log D$, $m = \log M$, and $k = \log K$. While it is not necessary for the definitions to make sense, in some cases it is more natural to think of N , D , and/or M as a power of 2, and thus the sets $[N]$, $[D]$, and $[M]$ as corresponding to the set of bit-strings of length n , d , and m , respectively. In some cases (namely, list-decodable codes and hardness amplifiers), we will restrict to functions in which y is a prefix of $\Gamma(x, y)$, and then it will be convenient to denote the range by $[D] \times [q]$ rather than $[M]$. This syntactic constraint actually leads to natural variants (sometimes referred to as “strong” or “seed-extending” variants) of the other objects, too, but we do not impose it here for sake of generality and consistency with the most commonly used definitions.

Much of the work on the objects are discussing is concerned with giving *explicit* constructions, which correspond to the function $\Gamma : [N] \times [D] \rightarrow [M]$ being deterministically and efficiently computable, e.g. in time $\text{poly}(n, d)$. However, since our focus is on the connections between the objects rather than their constructions, we will generally not discuss explicitness except in passing.

3. List-decodable Codes

We begin by describing how the standard notion of list-decodable codes can be cast in the framework, because it motivates the notation $\text{LIST}_\Gamma(\cdot, \cdot)$ and provides a good basis for understanding the other objects.

A *code* is specified by an *encoding function* mapping n -bit messages to code-words consisting of D symbols over an alphabet of size q . More generally, it can

be a function $\text{Enc} : [N] \rightarrow [q]^D$. (In the coding literature, the message alphabet is usually taken to be the same as the codeword alphabet, which translates to a scaling of the message length by a factor of $\log q$. In addition, the message length is usually denoted by k rather than n and the codeword length is n rather than D .) The goal is to define the function Enc so that if a codeword $\text{Enc}(x)$ is corrupted in a significant number of symbols and one only receives the corrupted string $r \in [q]^D$, the message x can still be recovered. *List-decodable* codes are designed for a setting where the number of corruptions is too large to hope for uniquely decoding x , and thus we settle for getting a short list of possible candidates.

Definition 2. A code $\text{Enc} : [N] \rightarrow [q]^D$ is (ε, K) list-decodable if for every “received word” $r \in [q]^D$, there are at most K messages $x \in [N]$ such that $\text{Enc}(x)$ and r agree in greater than a $1/q + \varepsilon$ fraction of positions.

This definition says that if we receive a string $r \in [q]^D$ that we know has resulted from corrupting a codeword $\text{Enc}(x)$ in less than a $1 - (1/q + \varepsilon)$ fraction of positions, then we can pin down the message x to one of at most K possibilities. K is thus called the *list size*. Note that we expect a uniformly random string $r \stackrel{\mathbb{R}}{\leftarrow} [q]^D$ to agree with most codewords in roughly a $1/q$ fraction of positions so we cannot expect to do any meaningful decoding from agreement $1/q$; this is why we ask for agreement greater than $1/q + \varepsilon$.

Naturally, one wants the agreement parameter ε to be as small possible and the (relative) *rate* $\rho = \log N / (D \log q)$ of the code to be as large as possible.

In coding theory, one typically considers both ε and ρ to be fixed constants in $(0, 1)$, while the message length $n = \log N$ tends to infinity and the alphabet size remains small (ideally, $q = O(1)$). The main challenge is to achieve an optimal tradeoff between the rate and agreement, while maintaining a list size K polynomially bounded in the message length n . Indeed, we usually also want an efficient algorithm that enumerates all the possible decodings x in time polynomial in n , which implies a polynomial bound on the list size. There has been dramatic progress on this challenge in the recent years; see the surveys [Sud, Gur].

To cast list-decodable codes in our framework, note that given a code $\text{Enc} : [N] \rightarrow [q]^D$, we can define a function $\Gamma : [N] \times [D] \rightarrow [D] \times [q]$ by

$$\Gamma(x, y) = (y, \text{Enc}(x)_y). \tag{1}$$

Note that the range of Γ is $[D] \times [q]$ and it has the property that the first component of $\Gamma(x, y)$ is always y . Moreover, given any Γ with this property, we can obtain a corresponding code Enc .

Proposition 3. Let the code $\text{Enc} : [N] \rightarrow [q]^D$ correspond to the function $\Gamma : [N] \times [D] \rightarrow [D] \times [q]$ via Equation (1). Then Enc is (ε, K) list-decodable if and only if

$$\forall r \in [q]^D \quad |\text{LIST}_\Gamma(T_r, 1/q + \varepsilon)| \leq K,$$

where $T_r = \{(y, r_y) : y \in [D]\}$.

Proof. It suffices to show that for every $r \in [q]^D$ and $x \in [N]$, we have $x \in \text{LIST}_\Gamma(T_r, 1/q + \varepsilon)$ iff $\text{Enc}(x)$ agrees with r in greater than a $1/q + \varepsilon$ fraction of places. We show this as follows:

$$\begin{aligned} x \in \text{LIST}_\Gamma(T_r, 1/q + \varepsilon) &\Leftrightarrow \Pr_{y \stackrel{R}{\leftarrow} [D]} [\Gamma(x, y) \in T_r] > 1/q + \varepsilon \\ &\Leftrightarrow \Pr_{y \stackrel{R}{\leftarrow} [D]} [(y, \text{Enc}(x)_y) \in T_r] > 1/q + \varepsilon \\ &\Leftrightarrow \Pr_{y \stackrel{R}{\leftarrow} [D]} [\text{Enc}(x)_y = r_y] > 1/q + \varepsilon \end{aligned}$$

□

In addition to the particular range of parameters typically studied (e.g. the small alphabet size q), the other feature that distinguishes list-decodable codes from many of the other objects described below is that it only considers sets of the form $T_r \subseteq [D] \times [q]$ for received words $r \in [q]^D$. These sets contain only one element for each possible first component $y \in [D]$, and thus are of size exactly D . Note that as the alphabet size q grows, these sets contain a vanishingly small fraction of the range $[D] \times [q]$.

4. Samplers

Suppose we are interested in estimating the average value of a boolean function $T : [M] \rightarrow \{0, 1\}$ on a huge domain $[M]$, given an oracle for T . The Chernoff Bound tells us that if we take $D = O(\log(1/\delta)/\varepsilon^2)$ independent random samples from $[M]$, then with probability at least $1 - \delta$, the average of T on the sample will approximate T 's global average within an additive error of ε . However, it is well known that these samples need not be generated independently; for example, samples generated according to a k -wise independent distribution or by a random walk on an expander graph have similar properties [CG2, BR, SSS, Gil]. The advantage of using such correlated sample spaces is that the samples can be generated using many fewer random bits than independent samples; this can be useful for derandomization and or simply because it provides a compact representation of the sequence of samples.

The definition below abstracts this idea of a procedure that uses $n = \log N$ random bits to generate D samples from $[M]$ with the above average-approximation property.

Definition 4 ([BR]¹). A sampler Smp for domain size M is given “coin tosses” $x \stackrel{R}{\leftarrow} [N]$ and outputs samples $z_1, \dots, z_D \in [M]$. We say that Smp is a (δ, ε)

¹Bellare and Rogaway [BR] referred to these as *oblivious samplers*, but they were renamed *averaging samplers* by Goldreich [Gol1].

averaging sampler if for every function $T : [M] \rightarrow \{0, 1\}$, we have

$$\Pr_{(z_1, \dots, z_D) \leftarrow \text{Smp}(U_{[N]})} \left[\frac{1}{D} \sum_i T(z_i) \leq \mu(T) + \varepsilon \right] \geq 1 - \delta,$$

where $\mu(T) \stackrel{\text{def}}{=} \mathbb{E}[T(U_{[M]})]$.

Note that the definition only bounds the probability that the sample-average deviates from $\mu(T)$ from above. However, a bound in both directions can be obtained by applying the above definition also to the complement of T , at the price of a factor of 2 in the error probability δ . (Considering deviations in only one direction will allow us to cast samplers in our framework without any slackness in parameters.) We note that the above definition can be also generalized to functions T that are not necessarily boolean, and instead map to the real interval $[0, 1]$. Non-boolean samplers and boolean samplers turn out to be equivalent up to a small loss in the parameters [Zuc2].

We note that one can consider more general notions of samplers that make adaptive oracle queries to the function T and and/or produce their estimate of $\mu(T)$ by an arbitrary computation on the values returned (not necessarily taking the sample average). In fact, utilizing this additional flexibility, there are known explicit samplers that achieve better parameters than we know how to achieve with averaging samplers. (For these generalizations, constructions of such samplers, and discussion of other issues regarding samplers, see the survey [Gol1].) Nevertheless, some applications require averaging samplers, and averaging samplers are also more closely related to the other objects we are studying.

In terms of the parameters, one typically considers M , ε , and δ as given, and seeks to minimize both the number $n = \log N$ of random bits and the number D of samples. Usually, complexity is measured as a function of $m = \log M$, with ε ranging between constant and $1/\text{poly}(m)$, and δ ranging between $o(1)$ and $2^{-\text{poly}(m)}$.

Samplers can be cast rather directly into our framework as follows. Given a sampler Smp for domain size M that generates D samples using coin tosses from $[N]$, we can define $\Gamma : [N] \times [D] \rightarrow [M]$ by setting

$$\Gamma(x, y) = \text{the } y\text{'th sample of Smp on coin tosses } x. \tag{2}$$

Conversely, any function $\Gamma : [N] \times [D] \rightarrow [M]$ yields a sampler. The property of Smp being an averaging sampler can be translated to the “list-decodability” of Γ as follows.

Proposition 5. *Let Smp be a sampler for domain size M that generates D samples using coin tosses from $[N]$, and let $\Gamma : [N] \times [D] \rightarrow [M]$ be the function corresponding to Smp via Equation (2). Then Smp is a (δ, ε) averaging sampler*

if and only if

$$\forall T \subseteq [M] \quad |\text{LIST}_\Gamma(T, \mu(T) + \varepsilon)| \leq K,$$

where $K = \delta N$ and $\mu(T) \stackrel{\text{def}}{=} |T|/M$.

Proof. We can view a function $T : [M] \rightarrow \{0, 1\}$ as the characteristic function of a subset of $[M]$, which, by abuse of notation, we also denote by T . Note that $\text{LIST}_\Gamma(T, \mu(T) + \varepsilon)$ is precisely the set of coin tosses x for which $\text{Smp}(x)$ outputs a sample on which T 's average is greater than $\mu(T) + \varepsilon$. Thus, the probability of a bad sample is at most δ iff $|\text{LIST}_\Gamma(T, \mu(T) + \varepsilon)| \leq \delta N$. \square

Let's compare the characterization of samplers given by Proposition 5 to the characterization of list-decodable codes given by Proposition 3. One difference is that codes correspond to functions Γ where $\Gamma(x, y)$ always includes y as a prefix. This turns out to be a relatively minor difference, and most known samplers can be modified to have this property. A major difference, however, is that for list-decodable codes, we only consider decoding from sets of the form T_r for some received word $r \in [q]^D$. Otherwise, the two characterizations are identical. (Note that $\mu(T_r) = 1/q$, and bounding K and bounding δ are equivalent via the relation $K = \delta N$.) Still, the settings of parameters typically considered in the two cases are quite different. In codes, the main growing parameter is the message length $n = \log N$, and one typically wants the alphabet size q to be a constant (e.g. $q = 2$) and the codeword length D to be linear in n . Thus, the range of Γ is of size $M = D \cdot q = O(\log N)$. In samplers, the main growing parameter is $m = \log M$, which is the number of random bits needed to select a single element of the universe $[M]$ uniformly at random, and one typically seeks samplers using a number random bits $n = \log N$ that is linear (or possibly polynomial) in m . Thus, $M = N^{\Omega(1)}$, in sharp contrast to the typical setting for codes. Also in contrast to codes, samplers are interesting even when δ is a constant independent of N (or vanishes slowly as a function of N). In such a case, the number of samples can be independent of N (e.g. in an optimal sampler, $D = O(\log(1/\delta)/\varepsilon^2)$). But constant δ in codes means that the list size $K = \delta N$ is a constant fraction of the message space, which seems too large to be useful from a coding perspective. Instead, the list size for codes is typically required to be $K = \text{poly}(n) = \text{poly}(\log N)$, which forces the codeword length D to be at least as large as the message length $n = \log N$.

5. Expander Graphs

Expanders are graphs with two seemingly contradictory properties. On one hand, they have very low degree; on the other, they are extremely well-connected. Expanders have numerous applications in theoretical computer science, and their study has also turned out to be mathematically very rich; see the survey [HLW].

There are a variety of measures of expansion, with close relationships between them, but here we will focus on the most basic measure, known as *vertex expansion*. We restrict attention to bipartite graphs, where the requirement is that every set of left-vertices that is not too large must have “many” neighbors on the right. We allow multiple edges between vertices. We require the graph to be left-regular, but it need not be right-regular.

Definition 6. *Let G be a left-regular bipartite multigraph with left vertex set $[N]$, right vertex set $[M]$, and left degree D . G is an $(= K, A)$ expander if every left-set S of size at least K has at least $A \cdot K$ neighbors on the right. G is a (K, A) expander if it is a $(= K', A)$ expander for every $K' \leq K$.*

The classic setting of parameters for expanders is the balanced one, where $M = N$, and then the goal is to have the degree D and the expansion factor A to both be constants independent of the number of vertices, with $A > 1$ and expansion achieved for sets of size up to $K = \Omega(M)$. However, the imbalanced case $M < N$ is also interesting, and then even expansion factors A smaller than 1 are nontrivial (provided $A > M/N$).

We can cast expanders in our framework as follows. For a left-regular bipartite multigraph G with left vertex set $[N]$, right vertex set $[M]$, and left degree D , we define the *neighbor function* $\Gamma : [N] \times [D] \rightarrow [M]$ by

$$\Gamma(x, y) = \text{the } y\text{'th neighbor of } x \tag{3}$$

Proposition 7. *Let G be a left-regular bipartite multigraph with left vertex set $[N]$, right vertex set $[M]$, and left degree D , and let $\Gamma : [N] \times [D] \rightarrow [M]$ be the neighbor function corresponding to G via Equation (3). Then G is an $(= K, A)$ expander if and only if*

$$\forall T \subseteq [M] \text{ s.t. } |T| < AK \quad |\text{LIST}_\Gamma(T, 1)| < K. \tag{4}$$

Thus, G is a (K, A) expander iff for every $T \subseteq [M]$ of size less than AK , we have $|\text{LIST}_\Gamma(T, 1)| < |T|/A$.

Proof. We show that G fails to be an $(= K, A)$ expander iff Condition (4) is false.

If G is not an $(= K, A)$ expander, then there is a left-set $S \subseteq [N]$ of size at least K with fewer than AK neighbors on the right. Let T be the set of neighbors of S . Then $|T| < AK$ but $S \subseteq \text{LIST}_\Gamma(T, 1)$, so $|\text{LIST}_\Gamma(T, 1)| \geq K$, violating Condition (4).

Conversely, suppose that Condition (4) fails. Then there is a right-set $T \subseteq [M]$ of size less than AK for which $|\text{LIST}_\Gamma(T, 1)| \geq K$. But the neighbors of $\text{LIST}_\Gamma(T, 1)$ are all elements of T , violating expansion. \square

We now compare the characterization of expanders given in Proposition 7 to those for list-decodable codes and samplers. First, note that we quantify over all sets T of a bounded size (namely, smaller than AK). In codes, the sets T

were also of a small size but also restricted to be of the form T_r for a received word r . In samplers, there was no constraint on T . Second, we only need a bound on $|\text{LIST}_\Gamma(T, 1)|$, which is conceivably easier to obtain than a bound on $|\text{LIST}_\Gamma(T, \mu(T) + \varepsilon)|$ as in codes and samplers. Nevertheless, depending on the parameters, vertex expansion (as in Definition 6 and Proposition 7) often implies stronger measures of expansion (such as a spectral gap [Alo] and randomness condensing [TUZ]), which in turn imply bounds on $|\text{LIST}_\Gamma(T, \mu(T) + \varepsilon)|$.

The typical parameter ranges for expanders are more similar to those for samplers than for those of codes. Specifically, N and M tend to be of comparable size; indeed, the classic case is $N = M$, and even in the unbalanced case, they are typically polynomially related. However, for expanders, there is no parameter ε . On the other hand, there is something new to optimize, namely the expansion factor A , which is the ratio between the size of T and the list size K . In particular, to have expansion factor larger than 1 (the classic setting of parameters for expansion), we must have a list size that is *smaller* than $|T|$. In samplers, however, there is no coupling of the list size and $|T|$; the list size $K = \delta N$ depends on the error probability δ , and should be apply for every $T \subseteq [M]$. With list-decodable codes, the set $T = T_r$ is always small (of size D), but the difference between list size D and, say, $D/2$ is typically insignificant.

Despite the above differences between codes and expanders, recent constructions of list-decodable codes have proved useful in constructing expanders with near-optimal expansion factors (namely, $A = (1 - \varepsilon)D$) via Proposition 7 [GUV]. A formulation of expansion similar to Proposition 7 also appeared in [GT].

6. Randomness Extractors

A randomness extractor is a function that extracts almost-uniform bits from a source of biased and correlated bits. The original motivation for extractors was the simulation of randomized algorithms with physical sources of randomness, but they have turned out to have a wide variety of other applications in theoretical computer science. Moreover, they have played a unifying role in the theory of pseudorandomness, and have been the avenue through which many of the connections described in this survey were discovered. History, applications, and constructions of extractors are described in more detail in [NT, Sha].

To formalize the notion of an extractor, we need to model a “source of biased and correlated bits” and define what it means for the output of the extractor to be “almost uniform.” For the former, we adopt a very general notion, advocated in [CG1, Zuc1], where we only require that the source has enough randomness in it, as measured by the following variant of entropy.

Definition 8. *The min-entropy of a random variable X is*

$$H_\infty(X) = \min_{x \in \text{Supp}(X)} \log(1/\Pr[X = x]).$$

X is a k -source if $H_\infty(X) \geq k$. Equivalently, X is a k -source if $\Pr[X = x] \leq 2^{-k}$ for all x .

Intuitively, we think of a k -source as having “ k bits of randomness” in it. For example, a random variable that is uniformly distributed over any $K = 2^k$ strings is a k -source.

For the quality of the output of the extractor, we use a standard measure of distance between probability distributions.

Definition 9. *The statistical difference between random variables X and Y taking values in a universe $[M]$ is defined to be*

$$\Delta(X, Y) = \max_{T \subseteq [M]} |\Pr[X \in T] - \Pr[Y \in T]| = \max_{T \subseteq [M]} \Pr[X \in T] - \Pr[Y \in T].$$

X and Y are ε -close if $\Delta(X, Y) \leq \varepsilon$. Otherwise, we say they are ε -far.

The equivalence between the formulations of statistical difference with and without the absolute values can be seen by observing that $\Pr[X \in T] - \Pr[Y \in T] = -(\Pr[X \in \bar{T}] - \Pr[Y \in \bar{T}])$.

Ideally we’d like an extractor to be a function $\text{Ext} : [N] \rightarrow [M]$ such that for every k -source X taking values in $[N]$, the random variable $\text{Ext}(X)$ is ε -close to $U_{[M]}$. That is, given an n -bit string coming from an unknown random source with at least k bits of randomness, the extractor is guaranteed to produce m bits that are close to uniform. However, this is easily seen to be impossible even when $m = 1$: the uniform distribution on either $\text{Ext}^{-1}(0)$ or $\text{Ext}^{-1}(1)$ is an $(n - 1)$ -source on which the output of the extractor is constant.

Nisan and Zuckerman [NZ] proposed to get around this difficulty by allowing the extractor a small number of truly random bits as a *seed* for the extraction.² This leads to the following definition.

Definition 10 ([NZ]). *$\text{Ext} : [N] \times [D] \rightarrow [M]$ is a (k, ε) extractor if for every k -source X taking values in $[N]$, $\text{Ext}(X, U_{[D]})$ is ε -close to $U_{[M]}$.*

The reason extraction is still interesting is that the number $d = \log D$ of truly random bits can be much smaller than the number of almost-uniform bits extracted. Indeed, d can be even be logarithmic in $m = \log M$, and thus in many applications, the need for a seed can be eliminated by enumerating all 2^d possibilities.

The ranges of the min-entropy threshold k most commonly studied in the extractor literature are $k = \alpha n$ or $k = n^\alpha$ for constants $\alpha \in (0, 1)$, where $n = \log N$ is the length of the source. The error parameter ε is often taken to be a small constant, but vanishing ε is important for some applications (especially

²Another way around the difficulty is to consider more restricted classes of sources or to allow multiple *independent* sources. There is a large and beautiful literature on “deterministic” extractors for these cases, which we do not discuss here.

in cryptography). One usually aims to have a seed length $d = O(\log n)$ or $d = \text{polylog}(n)$, and have the output length $m = \log M$ be as close to k as possible, corresponding to extracting almost all of the randomness from the source. (Ideally, $m \approx k + d$, but $m = \Omega(k)$ or $m = k^{\Omega(1)}$ often suffices.)

Notice that the syntax of extractors already matches that of the functions $\Gamma : [N] \times [D] \rightarrow [M]$ studied in our framework. The extraction property can be captured, with a small slackness in parameters, as follows.

Proposition 11. *Let $\Gamma = \text{Ext} : [N] \times [D] \rightarrow [M]$ and let $K = 2^k$. Then:*

1. *If Ext is a (k, ε) extractor, then*

$$\forall T \subseteq [M] \quad |\text{LIST}_\Gamma(T, \mu(T) + \varepsilon)| < K, \tag{5}$$

where $\mu(T) = |T|/M$.

2. *Conversely, if Condition 5 holds, then Ext is a $(k + \log(1/\varepsilon), 2\varepsilon)$ extractor.*

Proof. 1. Suppose that Condition (5) fails. That is, there is a set $T \subseteq [M]$ such that $|\text{LIST}_\Gamma(T, \mu(T) + \varepsilon)| \geq K$. Let X be a random variable distributed uniformly over $\text{LIST}_\Gamma(T, \mu(T) + \varepsilon)$. Then X is a k -source, but

$$\begin{aligned} \Pr[\text{Ext}(X, U_{[D]}) \in T] &= \mathbb{E}_{x \stackrel{\text{R}}{\leftarrow} X} [\Pr[\text{Ext}(x, U_{[D]}) \in T]] \\ &> \mu(T) + \varepsilon \\ &= \Pr[U_{[M]} \in T] + \varepsilon, \end{aligned}$$

so $\text{Ext}(X, U_{[D]})$ is ε -far from $U_{[M]}$. Thus, Ext is not a (k, ε) extractor.

2. Suppose Condition (5) holds. To show that Ext is a $(k + \log(1/\varepsilon), 2\varepsilon)$ extractor, let X be any $(k + \log(1/\varepsilon))$ -source taking values in $[N]$. We need to show that $\text{Ext}(X, U_{[D]})$ is 2ε -close to $U_{[M]}$. That is, we need to show that for every $T \subseteq [M]$, $\Pr[\text{Ext}(X, U_{[D]}) \in T] \leq \mu(T) + 2\varepsilon$.

So let T be any subset of $[M]$. Then

$$\begin{aligned} \Pr[\text{Ext}(X, U_{[D]}) \in T] &\leq \Pr[X \in \text{LIST}(T, \mu(T) + \varepsilon)] \\ &\quad + \Pr[\text{Ext}(X, U_{[D]}) \in T | X \notin \text{LIST}(T, \mu(T) + \varepsilon)] \\ &\leq |\text{LIST}(T, \mu(T) + \varepsilon)| \cdot 2^{-(k + \log(1/\varepsilon))} + (\mu(T) + \varepsilon) \\ &\leq K \cdot 2^{-(k + \log(1/\varepsilon))} + \mu(T) + \varepsilon \\ &= \mu(T) + 2\varepsilon \quad \square \end{aligned}$$

The slackness in parameters in the above characterization is typically insignificant for extractors. Indeed, it is known that extractors must lose at least $\Theta(\log(1/\varepsilon))$ bits of the source entropy [RT], and the above slackness only affects the leading constant.

Notice that the condition characterizing extractors here is *identical* to the one characterizing averaging samplers in Proposition 5. Thus, the only real difference between extractors and averaging samplers is one of perspective, and both perspectives can be useful. For example, recall that in samplers, we measure the error probability $\delta = K/N = 2^k/2^n$, whereas in extractors we measure the min-entropy threshold k on its own. Thus, the sampler perspective can be more natural when δ is relatively large compared to $1/N$, and the extractor perspective when δ becomes quite close to $1/N$. Indeed, an extractor for min-entropy $k = o(n)$ corresponds to a sampler with error probability $\delta = 1/2^{(1-o(1))n}$, which means that each of the n bits of randomness used by the sampler reduces the error probability by almost a factor of 2!

This connection between extractors and samplers was proven and exploited by Zuckerman [Zuc2]. The characterization of extractors in Proposition 11 was implicit in [Zuc2, Tre1], and was explicitly formalized in coding-theoretic terms by Ta-Shma and Zuckerman [TZ].

7. Hardness Amplifiers

The connections described in following two sections, which emerged from the work of Trevisan [Tre1], are perhaps the most surprising of all, because they establish a link between complexity-theoretic objects (which refer to computational intractability) and the purely information-theoretic and combinatorial objects we have been discussing so far.

Complexity Measures. In this section, we will be referring to a couple of different measures of computational complexity, which we informally review here. A *boolean circuit* C computes a finite function $C : \{0, 1\}^\ell \rightarrow \{0, 1\}^m$ using bit operations (such as AND, OR, and NOT). The *size* of a circuit C is the number of bit operations it uses. When we say that a circuit C computes a function $C : [n] \rightarrow [q]$, we mean that it maps the $\lceil \log n \rceil$ -bit binary representation of any element $x \in [n]$ to the corresponding $\lceil \log q \rceil$ -bit binary representation of $C(x)$.

As a measure of computational complexity, boolean circuit size is known to be very closely related to the running time of algorithms. However, boolean circuits compute functions on finite domains, so one needs to design a circuit separately for each input length, whereas an algorithm is typically required to be a single “uniform” procedure that works for all input lengths. This gap can be overcome by considering algorithms that are augmented with a “nonuniform advice string” for each input length:

Fact 12 ([KL]). *Let $f : \{0, 1\}^* \rightarrow \{0, 1\}^*$ be a function defined on bit-strings of every length, and $s : \mathbb{N} \rightarrow \mathbb{N}$ (with $s(n) \geq n$). Then the following are equivalent:*

1. *There is a sequence of circuits C_1, C_2, \dots such that $C_n(x) = f(x)$ for every $x \in \{0, 1\}^n$, and the size of C_n is $\tilde{O}(s(n))$.*

2. There is an algorithm A and a sequence of advice strings $\alpha_1, \alpha_2, \dots \in \{0, 1\}^*$ such that $A(x, \alpha_n) = f(x)$ for every $x \in \{0, 1\}^n$, and both the running time of A on inputs of length n and $|\alpha_n|$ are $\tilde{O}(s(n))$.

Thus “circuit size” equals “running time of algorithms with advice,” up to polylogarithmic factors (hidden by the $\tilde{O}(\cdot)$ notation). Notice that, for the equivalence with circuit size, the running time of A and the length of its advice string are equated; below we will sometimes consider what happens when we decouple the two (e.g. having bounded-length advice but unbounded running time).

We will also sometimes refer to computations with “oracles”. Running an algorithm A with *oracle access* to a function f (denoted A^f) means that as many times as it wishes during its execution, A can make a query x to the function f and receive the answer $f(x)$ in one time step. That is, A can use f as a subroutine, but we do not charge A for the time to evaluate f . But note that if A runs in time t and f can be evaluated in time s , then A^f can be simulated by a non-oracle algorithm B that runs in time $t \cdot s$. The same is true if we use circuit size instead of running time as the complexity measure.

Hardness Amplification. Hardness amplification is the task of increasing the average-case hardness of a function. We measure the average-case hardness of a function by the fraction of inputs on which every efficient algorithm (or circuit) must err.

Definition 13. A function $f : [n] \rightarrow [q]$ is (s, δ) hard if for every boolean circuit C of size s , we have

$$\Pr[C(U_{[n]}) \neq f(U_{[n]})] > \delta.$$

Hardness amplification is concerned with transforming a function so as to increase δ , the fraction of inputs on which it is hard. Ideally, we would like to go from $\delta = 0$, corresponding to *worst-case* hardness, to $\delta = 1 - 1/q - \varepsilon$, which is the largest value we can hope for (since every function with a range of $[q]$ can be computed correctly on a $1/q$ fraction of inputs by a constant circuit). In addition to the basic motivation of relating worst-case and average-case hardness, such hardness amplifications also are useful in constructing pseudorandom generators (see Section 8), because it is easier to construct pseudorandom generators from average-case hard functions (specifically, when $q = 2$ and $\delta = 1/2 - \varepsilon$) [NW, BFNW].

To make the goal more precise, we are interested in transformations for converting a function $f : [n] \rightarrow \{0, 1\}$ that is $(s, 0)$ hard to a function $f' : [n'] \rightarrow [q]$ that is $(s', 1 - 1/q - \varepsilon)$ hard for a constant q (ideally $q = 2$) and small ε . (The restriction of f to have range $\{0, 1\}$ is without loss of generality when considering worst-case hardness; otherwise we can use the function that outputs the j 'th bit of $f(i)$ on input (i, j) .)

The price that we usually pay for such hardness amplifications is that the circuit size for which the function is hard decreases (i.e. $s' < s$) and the domain size increases (i.e. $n' > n$); we would like these to be losses to be moderate (e.g. polynomial). Also, the complexity of computing the function correctly often increases and we again would like this increase to be moderate (e.g. f' should be computable in exponential time if f is). However, this latter property turns out to correspond to the “explicitness” of the construction, and thus we will not discuss it further below.

Several transformations achieving the above goal of converting worst-case hardness into average-case hardness are known; see the surveys [Kab, Tre2]. Like most (but not all!) results in complexity theory, these transformations are typically “black box” in the following sense. First, a single “universal” transformation algorithm Amp is given that shows how to compute f' given oracle access to f , and this transformation is well-defined for every oracle f , regardless of its complexity (even though we are ultimately interested only in functions f' within some complexity class, such as exponential time). Second, the property that f' is average-case hard when f is worst-case hard is proven by giving an “reduction” algorithm Red that efficiently converts algorithms r computing f' well on average into algorithms computing f in the worst-case. (Thus if f is hard in the worst case, there can be no efficient r computing f' well on average.) Again, even though we are ultimately interested in applying the reduction to efficient algorithms r , this property of the reduction should hold given any oracle r , regardless of its efficiency. Since our notion of hardness refers to nonuniform circuits, we will allow the reduction Red to use some nonuniform advice, which may depend on both f and r .

Black-box worst-case-to-average-case hardness amplifiers as described here are captured by the following definition.

Definition 14. *Let $\text{Amp}^f : [D] \rightarrow [q]$ be an algorithm that is defined for every oracle $f : [n] \rightarrow \{0, 1\}$. We say that Amp is a (t, k, ε) black-box worst-case-to-average-case hardness amplifier if there is an oracle algorithm Red , called the reduction, running in time t such that for every function $r : [D] \rightarrow [q]$ such that*

$$\Pr[r(U_{[D]}) = \text{Amp}^f(U_{[D]})] > 1/q + \varepsilon,$$

there is an advice string $z \in [K]$, where $K = 2^k$, such that

$$\forall i \in [n] \quad \text{Red}^r(i, z) = f(i).$$

The amplified function is $f' = \text{Amp}^f$; we have denoted the domain size as D rather than n' for convenience below. Note that without loss of generality, $k \leq t$, because an algorithm running in time t cannot read more than t bits of its advice string.

The following proposition shows that transformations meeting Definition 14 suffice for amplifying hardness.

Proposition 15. *If Amp is a (t, t, ε) black-box hardness amplifier and f is $(s, 0)$ hard, then Amp^f is $(s/\tilde{O}(t), 1 - 1/q - \varepsilon)$ hard.*

Proof. Suppose for contradiction there is a circuit $r : [D] \rightarrow [q]$ of size s' computing Amp^f on greater than a $1 - 1/q + \varepsilon$ fraction of inputs. Then there is an advice string z such that $\text{Red}^r(\cdot, z)$ computes f correctly on all inputs. Hardwiring z and using the fact that algorithms running in time t can be simulated by circuits of size $\tilde{O}(t)$, we get a circuit of size $\tilde{O}(t) \cdot s'$ computing f correctly on all inputs. This is a contradiction for $s' = s/\tilde{O}(t)$. \square

Typical settings of parameters for hardness amplification are $q = 2$, ε ranging from $o(1)$ to $1/n^{\Omega(1)}$, and $t = \text{poly}(\log n, 1/\varepsilon)$. Note that we make no reference to the length k of the advice string, and it does not appear in the conclusion of Proposition 15. Indeed, for the purposes of hardness amplification against nonuniform circuits, k may as well be set equal to running time t of the reduction. However, below it will be clarifying to separate these two parameters.

Now we place black-box hardness amplifiers in our framework. Given $\text{Amp}^f : [D] \rightarrow [q]$ defined for every oracle $f : [n] \rightarrow \{0, 1\}$, we can define $\Gamma : [N] \times [D] \rightarrow [D] \times [q]$ by

$$\Gamma(f, y) = (y, \text{Amp}^f(y)), \quad (6)$$

where $N = 2^n$ and we view $[N]$ as consisting of all boolean functions on $[n]$. Just as with list-decodable codes, the second input y is a prefix of the output of Γ . Moreover, any function Γ with this property yields a corresponding amplification algorithm Amp in the natural way. This syntactic similarity between codes and hardness amplifiers is no coincidence. The next proposition shows that, if we allow reductions Red of *unbounded* running time t (but still bounded advice length k), then black-box hardness amplifiers are *equivalent* to list-decodable codes.

Proposition 16. *Let $\text{Amp}^f : [D] \rightarrow [q]$ be an algorithm that is defined for every oracle $f : [n] \rightarrow \{0, 1\}$. Let $\Gamma : [N] \times [D] \rightarrow [D] \times [q]$ be the function corresponding to Amp via (6), where $N = 2^n$. Then Amp is an (∞, k, ε) black-box hardness amplifier if and only if*

$$\forall r \in [q]^D \quad |\text{LIST}_\Gamma(T_r, 1/q + \varepsilon)| \leq K,$$

where $T_r = \{(y, r_y) : y \in [D]\}$.

Note that the characterization given here is indeed identical to that of list-decodable codes given in Proposition 3.

Proof. First note that, viewing strings $r \in [q]^D$ as functions $r : [D] \rightarrow [q]$, we have $f \in \text{LIST}_\Gamma(T_r, 1/q + \varepsilon)$ iff

$$\Pr[r(U_{[D]}) = \text{Amp}^f(U_{[D]})] > 1/q + \varepsilon. \quad (7)$$

So we need to show that Amp is an (∞, k, ε) black-box hardness amplifier if and only if, for every function $r : [D] \rightarrow [q]$, there are at most K functions f satisfying Inequality (7).

Suppose that Amp is an (∞, k, ε) black-box hardness amplifier, let Red be the associated reduction, and let $r : [D] \rightarrow [q]$ be any function. If a function f satisfies Inequality (7), then, by Definition 14, f is of the form $\text{Red}^r(\cdot, z)$ for some $z \in [K]$. Since there are at most K choices for z , there are at most K functions satisfying Inequality (7).

Conversely, suppose that for every function $r : [D] \rightarrow [q]$, there are at most K functions f satisfying Inequality (7). Let $f_{r,1}, \dots, f_{r,K}$ be these functions in lexicographic order (repeating the last one if necessary to have exactly K functions). Then we can define the reduction Red by $\text{Red}^r(i, z) = f_{r,z}(i)$. (Recall that Red has unbounded running time, so constructing the list $f_{r,1}, \dots, f_{r,K}$ can be done by brute force.) By construction, for every function f satisfying Inequality (7), there exists a $z \in [K]$ such that $\text{Red}^r(\cdot, z) = f(\cdot)$. Thus Amp is an (∞, k, ε) black-box amplifier. \square

What about black-box amplifiers with reductions of *bounded* running time, as are needed for complexity-theoretic applications? (Proposition 15 is vacuous for $t = \infty$.)

First, note that every (t, k, ε) amplifier is also an (∞, k, ε) amplifiers, so we conclude that black-box amplifiers with efficient reductions are stronger than list-decodable codes. However, the efficiency of the reduction does have a natural coding-theoretic interpretation. Combining Constructions (1) and (6), we can interpret the role of the reduction Red in the following manner.

Assume for starters that Red does not use any advice, i.e. $K = 1$. Then Red is given oracle access to a received word $r \in [q]^D$ (meaning that it can ask for the j 'th symbol of r in one time step) and is given a message coordinate $i \in [n]$, and should output the i 'th symbol of the message f in time t . This is precisely the notion of *local decoding* for an error-correcting code; see the survey [Tre2]. Normally, a decoding algorithm is given the received word r in its entirety and should output the corresponding message f (or the list of possible messages f) in its entirety, ideally in polynomial time (e.g. time $\text{poly}(D, \log q) \geq \text{poly}(n)$). Here, however, we are interested in much smaller running times, such as $t = \text{poly}(\log n, 1/\varepsilon)$, so the decoder does not even have time to read the entire received word or write the entire message. Instead we give it oracle access to the received word and only ask to decode a particular message symbol in which we are interested.

From the previous paragraph, we see that black-box worst-case-to-average-case hardness amplifiers with no advice and bounded running time are *equivalent* to locally decodable error-correcting codes. With advice, Definition 14 provides a natural formulation of locally *list*-decodable error-correcting codes, where the number k of advice bits corresponds to the list size $K = 2^k$. (It is sometimes useful to allow a more general formulation, where the correspon-

dence between the advice strings z and decodings f , can be determined by a randomized preprocessing phase, which is given oracle access to r ; see [STV].)

Despite their close relationship, there are some differences in the typical parameter ranges for list-decodable codes and hardness amplification. In list-decodable codes, one typically wants the agreement parameter ε to be a constant and the codeword length to be linear in the message length (i.e. $D \log q = O(n)$). In hardness amplification, ε is usually taken to be vanishingly small (even as small as $1/n^{\Omega(1)}$), and one can usually afford for the codeword length to be polynomial in the message length (i.e. $D \log q = \text{poly}(n)$), because this corresponds to a linear blow-up in the input length of the amplified function Amp^f as compared to f . Another difference is that in locally list-decodable codes, it is most natural to for the list size K to be comparable to the running time t of the decoder, so the decoder has time to enumerate the elements of the list. For hardness amplification against nonuniform circuits, we may as well allow for the number of advice bits k to be as large as the running time t , which means that the list size $K = 2^k$ can be exponential in t .

The fact that locally list-decodable codes imply worst-case-to-average-case hardness amplification was shown by Sudan et al. [STV]. The fact that black-box amplifications imply list-decodable codes was implicit in [Tre1], and was made explicit in [TV].

8. Pseudorandom Generators

A pseudorandom generator is a deterministic function that stretches a short seed of truly random bits into a long string of “pseudorandom” bits that “look random” to any efficient algorithm. The idea of bits “looking random” is formalized by the notion of computational indistinguishability, which is a computational analogue of statistical difference (cf., Definition 9).

Definition 17 ([GM]). *Random variables X and Y are (s, ε) indistinguishable if for every boolean circuit T of size s , we have*

$$\Pr[T(X) = 1] - \Pr[T(Y) = 1] \leq \varepsilon.$$

This is equivalent to the more standard definition in which we bound the absolute value of the left-hand side by replacing T with its complement (which does not affect standard measures of circuit size).

Now we can define a pseudorandom generator as a function stretching d truly random bits into $m > d$ bits that are computationally indistinguishable from m truly random bits.

Definition 18 ([BM, Yao]). *A function $G : [D] \rightarrow [M]$ is an (s, ε) pseudorandom generator if $G(U_{[D]})$ is (s, ε) indistinguishable from $U_{[M]}$.*

Pseudorandom generators are powerful tools for cryptography and for derandomization (converting randomized algorithms to deterministic algorithms).

See the surveys [CRT, Mil, Kab, Gol3]. As far as the parameters, we would like the seed length $d = \log D$ to be as small as possible relative to the output length $m = \log M$, and we typically want generators that fool circuits of size $s = \text{poly}(m)$. The error parameter ε is usually not too important for derandomization (e.g. constant ε) suffices, but vanishing ε (e.g. $\varepsilon = 1/\text{poly}(m)$) is typically achievable and is crucial for cryptographic applications.

Another important parameter is the complexity of computing the generator itself. Even though this will not be explicit below, our discussions are most relevant to pseudorandom generators whose running time may be larger than the distinguishers T they fool, e.g. polynomial in s or even exponential in the seed length $d = \log D$. The study of such generators was initiated by Nisan and Wigderson [NW]. They suffice for derandomization, where we allow a polynomial slowdown in the algorithm we derandomize and anyhow enumerate over all $D = 2^d$ seeds. They are not suitable, however, for most cryptographic applications, where the generator is run by the honest parties, and must fool adversaries that have much greater running time.

The advantage of “noncryptographic” pseudorandom generators, whose running time is greater than that of the distinguishers, is that they can be constructed under weaker assumptions. The existence of “cryptographic” generators is equivalent to the existence of one-way functions [HILL], whereas “non-cryptographic” generators can be constructed from any boolean function (computable in time $2^{O(n)}$) with high circuit complexity [NW, BFNW].

We formalize the notion of a black-box construction of pseudorandom generators from functions of high worst-case circuit complexity analogously to Definition 14.

Definition 19. Let $G^f : [D] \rightarrow [M]$ be an algorithm that is defined for every oracle $f : [n] \rightarrow \{0, 1\}$. We say that G is a (t, k, ε) black-box PRG construction if there is an oracle algorithm Red , running in time t , such that for every $T : [M] \rightarrow \{0, 1\}$ such that

$$\Pr[T(G^f(U_{[D]})) = 1] - \Pr[T(U_{[M]}) = 1] > \varepsilon,$$

there is an advice string $z \in [K]$ such that

$$\forall i \in [n] \quad \text{Red}^T(i, z) = f(i).$$

Analogously to Proposition 15, black-box constructions according to the above definition do suffice for constructing pseudorandom generators from functions of high circuit complexity.

Proposition 20. If Amp is a (t, k, ε) black-box hardness amplifier and f is $(s, 0)$ hard, then G^f is an $(s/\tilde{O}(t), \varepsilon)$ pseudorandom generator.

Again, for the purposes of this proposition, k may as well be taken to be equal to t , and the values of interest range from the “high end” $k = t = n^{\Omega(1)}$ (applicable for functions f whose circuit complexity is exponential in the input

length, $\log n$) to the “low end” $k = t = (\log n)^{\Omega(1)}$ (applicable for functions f of superpolynomial circuit complexity).

We place pseudorandom generators in our framework analogously to hardness amplifiers. Given $G^f : [D] \rightarrow [M]$ defined for every oracle $f : [n] \rightarrow \{0, 1\}$, we can define $\Gamma : [N] \times [D] \rightarrow [M]$ by

$$\Gamma(f, y) = G^f(y), \tag{8}$$

where $N = 2^n$. Again, if we allow the reduction unbounded running time, then black-box pseudorandom generator constructions can be characterized exactly in our framework.

Proposition 21. *Let $G^f : [D] \rightarrow [M]$ be an algorithm defined for every oracle $f : [n] \rightarrow \{0, 1\}$, and let $\Gamma : [N] \times [D] \rightarrow [M]$ be the function corresponding to G via Equation (8). Then G is an (∞, k, ε) black-box hardness amplifier if and only if*

$$\forall T \subseteq [M] \quad |\text{LIST}_\Gamma(T, \mu(T) + \varepsilon)| \leq K,$$

where $K = 2^k$ and $\mu(T) = |T|/M$.

The proof of Proposition 21 is similar to Proposition 16, noting that we can view a function $T : [M] \rightarrow \{0, 1\}$ as the characteristic function of a set $T \subseteq [M]$ and conversely.

Notice that the condition in Proposition 21 is identical to the ones in our characterizations of averaging samplers (Proposition 5) and randomness extractors (Proposition 11). Thus, black-box pseudorandom generator constructions with reductions of unbounded running time (but bounded advice length k) are equivalent to both averaging samplers and randomness extractors. Analogously to the discussion of hardness amplifiers, an *efficient* reduction corresponds to extractors and samplers with efficient “local decoding” procedures. Here the decoder is given oracle access to a statistical test T that is trying to distinguish the output of the extractor Ext from uniform. It should be able to efficiently compute any desired bit of any source string $x = f$ for which T succeeds in distinguishing the output $\text{Ext}(x, U_{[D]})$ from uniform given some $k = \log K$ bits of advice depending on x . Even though achieving this additional local decoding property seems to only make constructing extractors more difficult, the perspective it provides has proved useful in constructing extractors, because it suggests an algorithmic approach to establishing the extractor property (namely, designing an appropriate reduction/decoder).

In terms of parameters, black-box PRG constructions are closer to extractors than samplers. In particular, the “high end” of PRG constructions has $k = t = n^{\Omega(1)}$, corresponding to extracting randomness from sources whose min-entropy is polynomially smaller than the length. However, a difference with extractors is that in pseudorandom generator constructions, one typically only looks for an output length m that it is polynomially related to $t = k$. This corresponds to extractors that extract $m = k^{\Omega(1)}$ bits out of the k bits of min-entropy

in the source, but for extractors, achieving $m = \Omega(k)$ or even $m \approx k + d$ is of interest. The connection between pseudorandom generators and extractors described here was discovered and first exploited by Trevisan [Tre1], and has inspired many subsequent works.

References

- [Alo] N. Alon. Eigenvalues and expanders. *Combinatorica*, 6(2):83–96, 1986. Theory of computing (Singer Island, Fla., 1984).
- [AB] S. Arora and B. Barak. *Computational complexity*. Cambridge University Press, Cambridge, 2009. A modern approach.
- [BFNW] L. Babai, L. Fortnow, N. Nisan, and A. Wigderson. BPP Has Subexponential Time Simulations Unless EXPTIME has Publishable Proofs. *Computational Complexity*, 3(4):307–318, 1993.
- [BR] M. Bellare and J. Rompel. Randomness-Efficient Oblivious Sampling. In *35th Annual Symposium on Foundations of Computer Science*, pages 276–287, Santa Fe, New Mexico, 20–22 Nov. 1994. IEEE.
- [BM] M. Blum and S. Micali. How to Generate Cryptographically Strong Sequences of Pseudo-Random Bits. *SIAM Journal on Computing*, 13(4):850–864, Nov. 1984.
- [CG1] B. Chor and O. Goldreich. Unbiased Bits from Sources of Weak Randomness and Probabilistic Communication Complexity. *SIAM Journal on Computing*, 17(2):230–261, Apr. 1988.
- [CG2] B. Chor and O. Goldreich. On the Power of Two-Point Based Sampling. *Journal of Complexity*, 5(1):96–106, Mar. 1989.
- [CRT] A. E. F. Clementi, J. D. P. Rolim, and L. Trevisan. Recent advances towards proving $P=BPP$. *Bulletin of the European Association for Theoretical Computer Science. EATCS*, 64:96–103, 1998.
- [GT] D. Galvin and P. Tetali. Slow mixing of Glauber dynamics for the hardcore model on regular bipartite graphs. *Random Structures & Algorithms*, 28(4):427–443, 2006.
- [Gil] D. Gillman. A Chernoff bound for random walks on expander graphs. *SIAM J. Comput.*, 27(4):1203–1220 (electronic), 1998.
- [Gol1] O. Goldreich. A Sample of Samplers - A Computational Perspective on Sampling (survey). *Electronic Colloquium on Computational Complexity (ECCC)*, 4(20), 1997.
- [Gol2] O. Goldreich. *Computational complexity: a conceptual perspective*. Cambridge University Press, Cambridge, 2008.
- [Gol3] O. Goldreich. Pseudorandom Generators: A Primer. <http://www.wisdom.weizmann.ac.il/oded/prg-primer.html>, July 2008. Revised version of [Gol2, Ch. 8].
- [GM] S. Goldwasser and S. Micali. Probabilistic Encryption. *Journal of Computer and System Sciences*, 28(2):270–299, Apr. 1984.

- [Gur] V. Guruswami. *Algorithmic Results in List Decoding*, volume 2, number 2 of *Foundations and Trends in Theoretical Computer Science*. now publishers, 2006.
- [GUV] V. Guruswami, C. Umans, and S. Vadhan. Unbalanced Expanders and Randomness Extractors from Parvaresh–Vardy Codes. *Journal of the ACM*, 56(4):1–34, 2009.
- [HILL] J. Håstad, R. Impagliazzo, L. A. Levin, and M. Luby. A pseudorandom generator from any one-way function. *SIAM Journal on Computing*, 28(4):1364–1396 (electronic), 1999.
- [HLW] S. Hoory, N. Linial, and A. Wigderson. Expander graphs and their applications. *Bulletin of the AMS*, 43(4):439–561, 2006.
- [Kab] V. Kabanets. Derandomization: a brief overview. *Bulletin of the EATCS*, 76:88–103, 2002.
- [KL] R. M. Karp and R. J. Lipton. Turing machines that take advice. *L’Enseignement Mathématique. Revue Internationale. Ite Série*, 28(3–4):191–209, 1982.
- [Mil] P. Miltersen. *Handbook of Randomized Computing*, chapter Derandomizing Complexity Classes. Kluwer, 2001.
- [NT] N. Nisan and A. Ta-Shma. Extracting Randomness: A Survey and New Constructions. *Journal of Computer and System Sciences*, 58(1):148–173, February 1999.
- [NW] N. Nisan and A. Wigderson. Hardness vs Randomness. *Journal of Computer and System Sciences*, 49(2):149–167, Oct. 1994.
- [NZ] N. Nisan and D. Zuckerman. Randomness is Linear in Space. *Journal of Computer and System Sciences*, 52(1):43–52, Feb. 1996.
- [RT] J. Radhakrishnan and A. Ta-Shma. Bounds for dispersers, extractors, and depth-two superconcentrators. *SIAM Journal on Discrete Mathematics*, 13(1):2–24 (electronic), 2000.
- [SSS] J. P. Schmidt, A. Siegel, and A. Srinivasan. Chernoff-Hoeffding bounds for applications with limited independence. *SIAM Journal on Discrete Mathematics*, 8(2):223–250, 1995.
- [Sha] R. Shaltiel. Recent Developments in Extractors. In G. Paun, G. Rozenberg, and A. Salomaa, editors, *Current Trends in Theoretical Computer Science*, volume 1: Algorithms and Complexity. World Scientific, 2004.
- [Sud] M. Sudan. List decoding: Algorithms and applications. *SIGACT News*, 31(1):16–27, 2000.
- [STV] M. Sudan, L. Trevisan, and S. Vadhan. Pseudorandom Generators without the XOR Lemma. *Journal of Computer and System Sciences*, 62:236–266, 2001.
- [TUZ] A. Ta-Shma, C. Umans, and D. Zuckerman. Loss-less condensers, unbalanced expanders, and extractors. In *Proceedings of the Thirty-Third Annual ACM Symposium on Theory of Computing*, pages 143–152 (electronic), New York, 2001. ACM.

- [TZ] A. Ta-Shma and D. Zuckerman. Extractor codes. *IEEE Transactions on Information Theory*, 50(12):3015–3025, 2004.
- [Tre1] L. Trevisan. Extractors and pseudorandom generators. *Journal of the ACM*, 48(4):860–879 (electronic), 2001.
- [Tre2] L. Trevisan. Some Applications of Coding Theory in Computational Complexity. *Quaderni di Matematica*, 13:347–424, 2004.
- [Tre3] L. Trevisan. Pseudorandomness and combinatorial constructions. In *International Congress of Mathematicians. Vol. III*, pages 1111–1136. Eur. Math. Soc., Zürich, 2006.
- [TV] L. Trevisan and S. Vadhan. Pseudorandomness and Average-Case Complexity via Uniform Reductions. *Computational Complexity*, 16(4):331–364, December 2007.
- [Vad1] S. Vadhan. The Unified Theory of Pseudorandomness. *SIGACT News*, 38(3):39–54, September 2007.
- [Vad2] S. P. Vadhan. *Pseudorandomness*. Foundations and Trends in Theoretical Computer Science. now publishers, 2010. To appear.
See <http://seas.harvard.edu/~salil/pseudorandomness>.
- [Yao] A. C. Yao. Theory and Applications of Trapdoor Functions (Extended Abstract). In *23rd Annual Symposium on Foundations of Computer Science*, pages 80–91, Chicago, Illinois, 3–5 Nov. 1982. IEEE.
- [Zuc1] D. Zuckerman. Simulating BPP Using a General Weak Random Source. *Algorithmica*, 16(4/5):367–391, Oct./Nov. 1996.
- [Zuc2] D. Zuckerman. Randomness-Optimal Oblivious Sampling. *Random Structures & Algorithms*, 11(4):345–367, 1997.

This page is intentionally left blank

Section 16

Numerical Analysis and Scientific Computing

This page is intentionally left blank

The Hybridizable Discontinuous Galerkin Methods

Bernardo Cockburn*

Abstract

In this paper, we present and discuss the so-called hybridizable discontinuous Galerkin (HDG) methods. The discontinuous Galerkin (DG) methods were originally devised for numerically solving linear and then nonlinear hyperbolic problems. Their success prompted their extension to the compressible Navier-Stokes equations – and hence to second-order elliptic equations. The clash between the DG methods and decades-old, well-established finite element methods resulted in the introduction of the HDG methods. The HDG methods can be implemented more efficiently and are more accurate than all previously known DG methods; they represent a competitive alternative to the well established finite element methods. Here we show how to devise and implement the HDG methods, argue why they work so well and prove optimal convergence properties in the framework of diffusion and incompressible flow problems. We end by briefly describing extensions to other continuum mechanics and fluid dynamics problems.

Mathematics Subject Classification (2010). Primary 65N30; Secondary 65M60.

Keywords. Convection, diffusion, incompressible fluid flow, discontinuous Galerkin methods, mixed methods, finite element methods

1. Introduction

In this paper, we study a recently introduced family of methods for numerically solving partial differential equations called the hybridizable discontinuous Galerkin (HDG) method. To motivate the advent of these methods, let us place them into a historical context.

*School of Mathematics, University of Minnesota, Minneapolis, MN 55455, USA.
E-mail: cockburn@math.umn.edu.

DG methods for hyperbolic problems. The original discontinuous Galerkin (DG) method was introduced in [47] back in 1973 for numerically solving the neutron transport equation, a linear hyperbolic equation for a scalar variable. The importance of the method was soon recognized and its first theoretical analysis was carried out in 1974 in [35]. The method lay dormant until the 90's, where it was successfully extended to non-linear time-dependent hyperbolic systems of conservation laws in a series of papers [27, 26, 25, 23, 28].

DG methods for convection-dominated problems. In 1997, prompted by the success of the RKDG methods for purely convective problems, the method was successfully extended [2] to the compressible Navier-Stokes equations. Soon after, many DG methods appeared for discretizing second-order symmetric elliptic problems and in 2002 a unifying framework for all of them was proposed in [1].

The new DG methods for second-order symmetric elliptic problems were then compared with the well established finite element methods, namely, the mixed methods of Raviart-Thomas (RT) [46] and Brezzi-Douglas-Marini (BDM) [4], and the continuous Galerkin (CG) method. A definite advantage of the DG methods was their ability to handle *adaptive* algorithms, as they are able to easily work with meshes with hanging nodes and approximations of varying polynomial degrees. However, when compared with the CG method, the new DG methods were criticized for having too many degrees of freedom and for not being as easy to implement. And when compared with the mixed methods, for providing less accurate approximations and also for not being as efficiently implementable.

The HDG methods. As a response to these criticisms, the HDG methods were introduced in [16] in the framework of diffusion problems. Therein, it was shown that the RT and BDM mixed methods could be obtained as particular cases of these new HDG methods. This suggested that HDG methods *close* to the RT and BDM methods could be implemented as efficiently and could even share their superior convergence properties while retaining the advantages typical of the DG methods. It was soon proven that this was the case in [11, 22, 18].

This breakthrough opened the possibility of a *new* systematic approach to devising HDG methods geared towards *efficient* implementation and towards achieving optimal order of convergence for *all* the unknowns as well as super-convergence of some of them.

Organization of the paper. In this paper, we show how this approach is being developed. Thus, in Section 2, we begin by revisiting the original DG method for transport in order to display the features that will also be those of the HDG methods for diffusion and incompressible flow problems. In Section 3,

we consider in detail the HDG methods for diffusion problems. Then in Section 4, we consider the HDG methods for the Stokes equations of incompressible flow. We end in Section 5 by briefly describing on the ongoing work on HDG methods for problems arising in continuum mechanics and fluid flow.

A short bibliographical note. The reader interested in a detailed history of the development of the DG methods up to 1999 is referred to [24]. More information about the DG methods can be found in the 2001 review [29], the 2003 short essay [7] and the 2004 article [8]. The subsequent work on DG methods is impossible to cover in a few references. However, the reader might want to see the paper on stabilization mechanisms of the DG methods [3], as well as the special issues on DG methods in [30] and [31]. Finally, a short overview of the HDG methods can be found in [40].

2. The Original DG Method for Transport

In this section, we revisit the original DG method [47] which was devised to numerically solve the neutron transport equation,

$$\sigma u + \nabla \cdot (\mathbf{a} u) = f \quad \text{in } \Omega, \quad (2.1a)$$

$$u = u_D \quad \text{on } \partial\Omega_-, \quad (2.1b)$$

where σ is a positive number, \mathbf{a} a constant vector and $\partial\Omega_-$ the *inflow* boundary of $\Omega \subset \mathbb{R}^d$, that is, $\partial\Omega_- = \{\mathbf{x} \in \partial\Omega : \mathbf{a} \cdot \mathbf{n}(\mathbf{x}) < 0\}$. Here $\mathbf{n}(\mathbf{x})$ is the outward unit normal at \mathbf{x} . For simplicity, we assume that Ω is a bounded polyhedral domain.

Our intention is to present the features of the method which are also going to be present in the HDG methods we consider in the following sections.

2.1. The method.

Discretization of the domain. We begin by discretizing the domain Ω . We consider disjoint open sets K called *elements* such that $\overline{\Omega} = \cup_{K \in \Omega_h} \overline{K}$. Their outward unit normal will be denoted by \mathbf{n} . We denote by Ω_h the collection of all these elements and set $\partial\Omega_h := \{\partial K : K \in \Omega_h\}$. We say that F is an *interior face* of the triangulation Ω_h if there are two elements K^+ and K^- in Ω_h such that $F = \partial K^+ \cap \partial K^-$. In this case, we denote by \mathbf{n}^\pm the outward unit normal of K^\pm at F . The collection of all interior faces is denoted by \mathcal{E}_h° . We say that F is a *boundary face* of the triangulation Ω_h if there is an element K in Ω_h such that $F = \partial K \cap \partial\Omega$. The collection of all boundary faces is denoted by \mathcal{E}_h^∂ . The set $\mathcal{E}_h = \mathcal{E}_h^\circ \cup \mathcal{E}_h^\partial$ is called the set of the faces of the triangulation Ω_h .

Rewriting the equations. For each $K \in \Omega_h$, the DG method is devised to provide an approximation u_h to the *restriction* to K of the exact solution u as well as an approximation $\mathbf{a} \cdot \mathbf{n} \hat{u}_h$ to the *trace* on ∂K of the normal component of the flux $\mathbf{a} \cdot \mathbf{n} \hat{u}$. Thus, to define these approximations, we need to rewrite the original problem (2.1) in terms of those functions.

We do this as follows. On each element $K \in \Omega_h$, we obtain u in terms of \hat{u} on ∂K_- by solving

$$\sigma u + \nabla \cdot (\mathbf{a} u) = f \quad \text{in } K, \tag{2.2a}$$

$$u = \hat{u} \quad \text{on } \partial K_-. \tag{2.2b}$$

In turn, the function \hat{u} on ∂K is expressed in terms of u and u_D as follows:

$$\hat{u}(\mathbf{x}) := \begin{cases} u_D(\mathbf{x}) & \text{for } \mathbf{x} \in \partial K \cap \partial\Omega_-, \\ \lim_{\epsilon \downarrow 0} u(\mathbf{x} - \epsilon \mathbf{a}) & \text{otherwise.} \end{cases} \tag{2.2c}$$

Here $\partial K_- := \{\mathbf{x} \in \partial K : \mathbf{a} \cdot \mathbf{n}(\mathbf{x}) < 0\}$.

The Galerkin formulation and the numerical trace. Now we discretize the above equations by combining a Galerkin method with a suitable definition of the numerical trace of the flux. So, on the element K , we take the approximation u_h in the finite dimensional space $W(K)$ and determine it by requiring that it satisfy a formulation we describe next. If we multiply the equation (2.2a) by a smooth function w , integrate over K , integrate by parts and use equation (2.2b), we get that

$$\sigma(u, w)_K - (u, \mathbf{a} \cdot \nabla w)_K + \langle \mathbf{a} \cdot \mathbf{n} \hat{u}, w \rangle_{\partial K} = (f, w)_K.$$

Here $(u, w)_K$ is the integral of $u w$ over K and $\langle w, v \rangle_{\partial K}$ the integral of $w v$ over ∂K . Thus, we require that

$$\sigma(u_h, w)_K - (u_h, \mathbf{a} \cdot \nabla w)_K + \langle \mathbf{a} \cdot \mathbf{n} \hat{u}_h, w \rangle_{\partial K} = (f, w)_K, \tag{2.3a}$$

for all $w \in W(K)$. On ∂K , \hat{u}_h is expressed in terms of u_h and u_D by

$$\hat{u}_h(\mathbf{x}) := \begin{cases} u_D(\mathbf{x}) & \text{for } \mathbf{x} \in \partial K \cap \partial\Omega_-, \\ \lim_{\epsilon \downarrow 0} u_h(\mathbf{x} - \epsilon \mathbf{a}) & \text{otherwise.} \end{cases} \tag{2.3b}$$

This completes the definition of the original DG method [47].

The discrete energy identity and the existence and uniqueness of the approximation. Let us show that the method is well defined. To do that,

we use the following *energy* argument. We begin by noting that if we multiply the equation (2.1a) by u , integrate over Ω and carry out some simple algebraic manipulations, we obtain the following *energy* identity:

$$\sigma \|u - f/2\sigma\|_{L^2(\Omega)}^2 + \frac{1}{2} \langle |\mathbf{a} \cdot \mathbf{n}| u, u \rangle_{\partial\Omega \setminus \partial\Omega_-} = \Psi(f, u_d),$$

where $\Psi(f, u_d) := \frac{1}{4\sigma} \|f\|_{L^2(\Omega)}^2 + \frac{1}{2} \langle |\mathbf{a} \cdot \mathbf{n}| u_D, u_D \rangle_{\partial\Omega_-}$. A discrete version of this identity can be obtained by taking $w := u_h$ in the equation (2.3a), adding over all the elements $K \in \Omega_h$ and performing similar manipulations to get

$$\sigma \|u_h - f/2\sigma\|_{L^2(\Omega)}^2 + \frac{1}{2} \langle |\mathbf{a} \cdot \mathbf{n}| u_h, u_h \rangle_{\partial\Omega \setminus \partial\Omega_-} + \Theta_h(u_h - \hat{u}_h) = \Psi(f, u_D), \tag{2.4}$$

where $\Theta_h(u_h - \hat{u}_h) = \frac{1}{2} \sum_{K \in \Omega_h} \langle |\mathbf{a} \cdot \mathbf{n}| (u_h - \hat{u}_h), u_h - \hat{u}_h \rangle_{\partial K_-} \geq 0$.

Now, to prove that the DG method is well defined, we only have to show that if we set $f = 0$ and $u_D = 0$, we obtain the trivial solution. But, by the energy identity we immediately get that $u_h = 0$ since σ is a positive number. This proves that the approximate solution exists and is unique.

Implementation. Note that, by construction, we have that

$$(\sigma u_h, w)_K - (u_h, \mathbf{a} \cdot \nabla w)_K + \langle \mathbf{a} \cdot \mathbf{n} u_h, w \rangle_{\partial K \setminus \partial K_-} = (f, w)_K - \langle \mathbf{a} \cdot \mathbf{n} \hat{u}_h, w \rangle_{\partial K_-}.$$

Thus we see that if \hat{u}_h on ∂K_- is known, then u_h on K can be computed. This is a remarkable property as it allows for an efficient implementation of the method. One cannot praise enough the importance of a property like this in a numerical method.

2.2. The stabilization mechanism.

The relation between the residuals. Note that the DG method is actually defined by imposing a linear relation between the residual in the interior of the element K , $R_K := \sigma u_h + \nabla \cdot (\mathbf{a} u_h) - f$, and the residual on its boundary ∂K , $R_{\partial K} := \mathbf{a} \cdot \mathbf{n} (u_h - \hat{u}_h)$.

Indeed, in terms of these residuals, the first equation defining the DG method (2.3a) reads

$$(R_K, w)_K = \langle R_{\partial K}, w \rangle_{\partial K} \quad \forall w \in W(K).$$

Since this implies that $\|P_{W(K)} R_K\|_{L^2(K)} \leq C h_K^{-1/2} \|R_{\partial K}\|_{L^2(\partial K)}$, where $P_{W(K)}$ is the L^2 -projection into $W(K)$, and since

$$\|R_K\|_{L^2(K)} \leq \|P_{W(K)} R_K\|_{L^2(K)} + \|(\text{Id} - P_{W(K)}) R_K\|_{L^2(K)},$$

we see that the size of the residual in the interior R_K is controlled by the residual on the boundary $R_{\partial K}$ and by the approximation properties of the L^2 -projection

into $W(K)$. This means that the size of the residuals, and hence the quality of the approximation, depend only on the size of the jumps $|\mathbf{a} \cdot \mathbf{n}|^{1/2}(u_h - \hat{u}_h)$ and on the approximation properties of the space $W(K)$.

Stabilization by the jumps. The discrete energy identity (2.4) suggests that, since $\Theta(u_h - \hat{u}_h) \geq 0$, the size of the jumps $|\mathbf{a} \cdot \mathbf{n}|^{1/2}(u_h - \hat{u}_h)$ remains bounded.

In fact, the quantity $\Theta_h(u_h - \hat{u}_h)$ is a dissipative term that enhances the stability properties of the method. Indeed, note that there is more dissipation if the size of the jump $u_h - \hat{u}_h$ is big. This happens, for example, whenever the exact solution is discontinuous and consequently the interior residual is big. Thus, the DG method has a built-in mechanism that transforms its potential inability to obtain a good approximation into *numerical dissipation* and into improved stability properties.

The counterpart of this mechanism in finite difference and finite volume methods for scalar hyperbolic conservation laws is the one induced by the so-called *artificial viscosity* term. One of the main problems for those methods is to define it in such a way that high-order accuracy can be attained. The stabilization mechanism of the DG method has such highly valued property. Indeed, next we show that optimal accuracy can be reached whenever the exact solution is smooth enough.

2.3. Convergence properties. We end our review of the original DG method by showing that optimal convergence properties of u_h can be proven when Ω is a polyhedral domain and the triangulations Ω_h are made of shape-regular simplexes K satisfying two conditions: (i) Each simplex K has a unique *\mathbf{a} -outflow* face, F_K^+ , and (ii) each interior face is the *\mathbf{a} -inflow* face of another simplex. We say that the face F is an *\mathbf{a} -outflow* (*inflow*) face when $\mathbf{a} \cdot \mathbf{n} > (<) 0$, where \mathbf{n} is the outward unit normal at F .

The auxiliary projection. Indeed, in this case, and when $W(K)$ is the space of polynomials of degree k on K , $\mathcal{P}_k(K)$, for each $K \in \Omega_h$, we can find an auxiliary projection Π with which the error analysis is greatly simplified. It is defined as follows. On the element K , the projection of the function $u \in H^1(K)$, Πu , is defined as the element of $\mathcal{P}_k(K)$ satisfying

$$\begin{aligned} (\Pi u - u, w)_K &= 0 & \forall w \in \mathcal{P}_{k-1}(K), \\ \langle \Pi u - u, w \rangle_{F_K^+} &= 0 & \forall w \in \mathcal{P}_k(F_K^+). \end{aligned}$$

This projection is well defined and, for smooth functions u , it provides optimal approximation properties, that is,

$$\|\Pi u - u\|_{L^2(K)} \leq C |u|_{H^{k+1}(K)} h^{k+1}.$$

Estimate of the projection of the error. The main reason for considering the projection of the error, $\varepsilon_u := \Pi(u - u_h)$, is that the projection Π is tailored to the structure of the numerical trace of the method. Indeed, for $\mathbf{x} \in \mathcal{E}_h \setminus \partial\Omega_-$, we have, by construction, that

$$\begin{aligned} \mathbf{a} \cdot \mathbf{n} \mathbb{P}_{M_h} u(\mathbf{x}) &= \mathbf{a} \cdot \mathbf{n} \lim_{\epsilon \downarrow 0} \Pi u(\mathbf{x} - \epsilon \mathbf{a}) \\ \mathbf{a} \cdot \mathbf{n} \widehat{u}_h(\mathbf{x}) &= \mathbf{a} \cdot \mathbf{n} \lim_{\epsilon \downarrow 0} u_h(\mathbf{x} - \epsilon \mathbf{a}), \end{aligned}$$

and we see that $\mathbf{a} \cdot \mathbf{n} (\mathbb{P}_{M_h} u - \widehat{u}_h)(\mathbf{x}) = \mathbf{a} \cdot \mathbf{n} \widehat{\varepsilon}_u(\mathbf{x})$. As a consequence, ε_u is the solution of

$$\sigma(\varepsilon_u, w)_K - (\varepsilon_u, \mathbf{a} \cdot \nabla w)_K + \langle \mathbf{a} \cdot \mathbf{n} \widehat{\varepsilon}_u, w \rangle_{\partial K} = \sigma(\Pi u - u, w)_K \quad \forall w \in W(K),$$

where $\widehat{\varepsilon}_u = 0$ on $\partial\Omega_-$. We immediately see that the projection of the error must depend *only* on $\Pi u - u$. In particular, by following the process used to obtain the discrete energy identity, we get

$$\sigma \|\varepsilon_u - \frac{1}{2}(\Pi u - u)\|_{L^2(\Omega)}^2 + \Theta_h(\varepsilon_u - \widehat{\varepsilon}_u) = \frac{\sigma}{4} \|\Pi u - u\|_{L^2(\Omega)}^2,$$

and we can deduce that,

$$\|u - u_h\|_{L^2(\Omega)} + \sigma^{-1/2} \Theta_h^{1/2}(\varepsilon_u - \widehat{\varepsilon}_u) \leq C |u|_{H^{k+1}(\Omega_h)} h^{k+1}.$$

This result is optimal in both regularity and order of convergence; see more general results in [10, 9]. For arbitrary meshes, there is a loss in the order of convergence of 1/2; see [33]. Although in practice this loss is hard to observe, it has been proven to actually occur in [45] and [48].

Postprocessing. Finally, we show how to postprocess the approximate solution in order to get an optimally convergent approximation of $\partial_{\mathbf{a}} u := \mathbf{a} \cdot \nabla u$. We proceed as follows. First, for each simplex K , we define the approximation \mathbf{q}_h of the flux $\mathbf{a} u$ as the element of $\mathcal{P}_k(K) + \mathbf{x} \mathcal{P}_k(K)$ that is the solution of

$$\begin{aligned} (\mathbf{q}_h - \mathbf{a} u_h, \mathbf{v})_K &= 0, & \forall \mathbf{v} \in \mathcal{P}_{k-1}(K) \\ \langle (\mathbf{q}_h - \mathbf{a} \widehat{u}_h) \cdot \mathbf{n}, w \rangle_F &= 0, & \forall w \in \mathcal{P}_k(F), \text{ for all faces } F \text{ of } K. \end{aligned}$$

Here $\mathcal{P}_k(K) := [\mathcal{P}_k(K)]^d$. Then, if we set $\partial_{\mathbf{a}} u_h^* := \nabla \cdot \mathbf{q}_h$, it is easy to show [10, 9] that $\partial_{\mathbf{a}} u_h^* - \mathbb{P}_{W(K)}(\partial_{\mathbf{a}} u) = \sigma(\mathbb{P}_{W(K)} u - u_h)$ on each simplex $K \in \Omega_h$, and to conclude that

$$\|\partial_{\mathbf{a}} u_h^* - \partial_{\mathbf{a}} u\|_{L^2(\Omega_h)} \leq C(|u|_{H^{k+1}(\Omega_h)} + |\partial_{\mathbf{a}} u|_{H^{k+1}(\Omega_h)}) h^{k+1}.$$

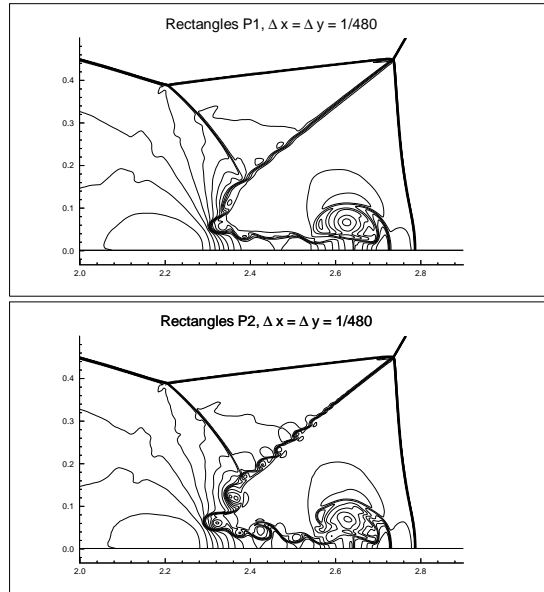


Figure 1. Euler equations of gas dynamics: Double Mach reflection problem. Isolines of the density around the double Mach stems. Linear polynomials on squares $\Delta x = \Delta y = \frac{1}{480}$ (top); and quadratic polynomials on squares $\Delta x = \Delta y = \frac{1}{480}$ (bottom).

2.4. The RKDG methods. Let us briefly point out that the extension of the original DG method to nonlinear hyperbolic conservation laws, called the Runge-Kutta discontinuous Galerkin (RKDG) methods, [27, 26, 25, 23, 28], shares with the original DG method many of the above-mentioned properties since it uses DG method to discretize the equations in space. To discretize the equations in time, a special type of *explicit* Runge-Kutta time marching methods is used. The distinctive feature of these Runge-Kutta methods is that their stability follows from the stability of a single Euler-forward step. A *crucial* component of the method is an operator (the so-called *slope limiter*) used to enforce the stability of the Euler-forward step and, as argued in [6], to ensure the convergence to the physically relevant solution. A rigorous convergence proof, however, remains a challenging open problem even for the scalar hyperbolic conservation law. See [20] for rigorous error estimates for an *implicit, shock-capturing* DG method for that equation.

In practice, however, the methods turned out to be optimally convergent as well as able to capture very well the discontinuities of the solution. For example, consider the classical double-Mach reflection problem for the Euler equations of gas dynamics. In Fig. 1, obtained in [28], details of the approximation of the density are shown. The strong shocks are very well resolved by the RKDG solution using piecewise linear and piecewise quadratic polynomials defined on

squares. Note that there is a remarkable improvement in the approximation of the density near the contacts when going from linear to quadratic polynomials.

3. HDG Methods for Diffusion

In this section, we consider HDG methods to numerically solve the diffusion model problem

$$c \mathbf{q} + \nabla u = 0 \quad \text{in } \Omega, \tag{3.1a}$$

$$\nabla \cdot \mathbf{q} = f \quad \text{in } \Omega, \tag{3.1b}$$

$$u = u_D \quad \text{on } \partial\Omega. \tag{3.1c}$$

Here c is a matrix-valued function which is symmetric and uniformly positive definite on Ω .

We are going to show that despite the fact that the nature of this problem is radically different from the one just considered, we can devise HDG methods by using a very similar approach. Most of the material for this section is contained in [16, 11, 22, 18].

3.1. The HDG methods.

Rewriting the equations. For each $K \in \Omega_h$, the methods provide an approximation to the restriction of (\mathbf{q}, u) to K as well as an approximation to the traces $(\hat{\mathbf{q}} \cdot \mathbf{n}, \hat{u})$ on ∂K . We are thus going to rewrite the original equations in terms of those functions in order to be able to define the HDG methods by discretizing them.

Thus, if for each $K \in \Omega_h$, we assume that we know the trace \hat{u} on ∂K , we can obtain (\mathbf{q}, u) inside K as the solution of

$$\begin{aligned} c \mathbf{q} + \nabla u &= 0 && \text{in } K, \\ \nabla \cdot \mathbf{q} &= f && \text{in } K, \\ u &= \hat{u} && \text{on } \partial K. \end{aligned}$$

The function \hat{u} can now be determined as the solution, on each $F \in \mathcal{E}_h$, of the equations

$$\begin{aligned} \llbracket \hat{\mathbf{q}} \cdot \mathbf{n} \rrbracket &= 0 && \text{if } F \in \mathcal{E}_h^o, \\ \hat{u} &= u_D && \text{if } F \in \mathcal{E}_h^\partial. \end{aligned}$$

Here we are using the notation $\llbracket \hat{\mathbf{q}} \cdot \mathbf{n} \rrbracket := \hat{\mathbf{q}}^+ \cdot \mathbf{n}^+ + \hat{\mathbf{q}}^- \cdot \mathbf{n}^-$.

The Galerkin formulation and the numerical traces. Now we discretize these equations by using a Galerkin method together with a suitable approximation of the traces. First, we take (\mathbf{q}_h, u_h) on the element $K \in \Omega_h$ in the finite dimensional space $\mathbf{V}(K) \times W(K)$ and \widehat{u}_h on the face $F \in \mathcal{E}_h^\circ$ in the finite dimensional space $M(F)$.

On each element $K \in \Omega_h$, the function (\mathbf{q}_h, u_h) is expressed in terms of (\widehat{u}_h, f) by using a Galerkin method. Since

$$\begin{aligned} (c \mathbf{q}, \mathbf{v})_K - (u, \nabla \cdot \mathbf{v})_K + \langle \widehat{u}, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K} &= 0, \\ -(\mathbf{q}, \nabla w)_K + \langle \widehat{\mathbf{q}} \cdot \mathbf{n}, w \rangle_{\partial K} &= (f, w)_K, \end{aligned}$$

for all sufficiently smooth functions (\mathbf{v}, w) , we determine (\mathbf{q}_h, u_h) in terms of (\widehat{u}_h, f) as the solution of

$$(c \mathbf{q}_h, \mathbf{v})_K - (u_h, \nabla \cdot \mathbf{v})_K + \langle \widehat{u}_h, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K} = 0, \tag{3.2a}$$

$$-(\mathbf{q}_h, \nabla w)_K + \langle \widehat{\mathbf{q}}_h \cdot \mathbf{n}, w \rangle_{\partial K} = (f, w)_K, \tag{3.2b}$$

for all $(\mathbf{v}, w) \in \mathbf{V}(K) \times W(K)$, where the numerical trace $\widehat{\mathbf{q}}_h \cdot \mathbf{n}$ is assumed to have the following simple form:

$$\widehat{\mathbf{q}}_h \cdot \mathbf{n} = \mathbf{q}_h \cdot \mathbf{n} + \tau(u_h - \widehat{u}_h) \quad \text{on } \partial K. \tag{3.2c}$$

Then the function \widehat{u}_h is determined by enforcing weakly the single-valuedness of the normal component of the numerical trace $\widehat{\mathbf{q}}_h$ and by capturing the Dirichlet boundary condition. Thus, for each face $F \in \mathcal{E}_h$, we require that

$$\langle \mu, \llbracket \widehat{\mathbf{q}}_h \cdot \mathbf{n} \rrbracket \rangle_F = 0 \quad \forall \mu \in M(F), \tag{3.2d}$$

$$\widehat{u}_h = u_D \quad \text{if } F \in \mathcal{E}_h^\partial. \tag{3.2e}$$

This completes the definition of the HDG methods. Note that equation (3.2d) is a condition on the single-valuedness of the normal component of $\widehat{\mathbf{q}}_h$ on \mathcal{E}_h° . Indeed, if the restriction of $\llbracket \widehat{\mathbf{q}}_h \cdot \mathbf{n} \rrbracket$ to F lies in $M(F)$ for all $F \in \mathcal{E}_h^\circ$, we have that $\llbracket \widehat{\mathbf{q}}_h \cdot \mathbf{n} \rrbracket = 0$ on \mathcal{E}_h° and the normal component of $\widehat{\mathbf{q}}_h$ is single valued. Studies of the importance of this property can be found in [22, 21].

The discrete energy identity and the existence and uniqueness of the approximation. The HDG methods are well defined under some very mild conditions, as we see in the next result.

Proposition 3.1. *The HDG method is well defined if (i) $\tau > 0$ on $\partial\Omega_h$, and if (ii) for any element $K \in \Omega_h$, $\nabla w \in \mathbf{V}(K)$ for all $w \in W(K)$.*

We can prove this result by using an energy argument. If we multiply the first two equations (3.1) defining the exact solution by \mathbf{q} and u , respectively,

integrate over Ω and add the equations, we obtain the following *energy identity*

$$(c \mathbf{q}, \mathbf{q})_\Omega = (f, u)_\Omega - \langle u_D, \mathbf{q} \cdot \widehat{\mathbf{n}} \rangle_{\partial\Omega}.$$

If we now apply a similar procedure to the HDG method, we get a *discrete energy identity*. So, taking $(\mathbf{v}, w) := (\mathbf{q}_h, u_h)$ in the first two equations defining the HDG method, adding over all the elements $K \in \Omega_h$ and then adding the resulting equations, we get

$$(c \mathbf{q}_h, \mathbf{q}_h)_\Omega + \Theta_\tau(u_h - \widehat{u}_h) = (f, u_h)_\Omega - \langle u_D, \widehat{\mathbf{q}}_h \cdot \widehat{\mathbf{n}} \rangle_{\partial\Omega}. \tag{3.3}$$

where $\Theta_\tau(u_h - \widehat{u}_h) := \sum_{K \in \Omega_h} \langle \tau(u_h - \widehat{u}_h), u_h - \widehat{u}_h \rangle_{\partial K}$.

To prove Proposition 3.1, we only have to show that if we set $f = 0$ and $u_D = 0$, the only solution is the trivial one. But, by the discrete energy identity and condition (i) of Proposition 3.1, we have that $\mathbf{q}_h = \mathbf{0}$ on Ω_h and that $u_h = \widehat{u}_h$ on \mathcal{E}_h . Now, by equation (3.2a), this implies that $(\nabla u_h, \mathbf{v})_K = 0 \quad \forall \mathbf{v} \in \mathbf{V}(K)$, and by condition (ii), we conclude that u_h is a constant on Ω . Since $u_h = \widehat{u}_h = 0$ on $\partial\Omega$, we see that $u_h = 0$ on Ω_h and hence that $\widehat{u}_h = 0$ on \mathcal{E}_h . The proof of Proposition 3.1 is complete.

Implementation. To describe the implementation of the HDG methods, we need to introduce some notation. We denote by $(\mathcal{Q}(\widehat{u}_h, f), \mathcal{U}(\widehat{u}_h, f))$ the linear mapping (see equations (3.2a), (3.2b) and (3.2c)) that associates (\widehat{u}_h, f) to (\mathbf{q}_h, u_h) and set

$$\begin{aligned} (\mathcal{Q}^{\widehat{u}_h}, \mathcal{U}^{\widehat{u}_h}) &:= (\mathcal{Q}(\widehat{u}_h, 0), \mathcal{U}(\widehat{u}_h, 0)), \\ (\mathcal{Q}^f, \mathcal{U}^f) &:= (\mathcal{Q}(0, f), \mathcal{U}(0, f)). \end{aligned}$$

We also introduce the space $M_h := \{\mu \in L^2(\mathcal{E}_h) : \mu|_F \in M(F) \quad F \in \mathcal{E}_h^o\}$ and set $M_h(\zeta) := \{\mu \in M_h : \mu|_{\partial\Omega} = \zeta\}$.

With this notation, we can characterize the approximate solution given by the HDG method as follows.

Proposition 3.2. *We have that $(\mathbf{q}_h, u_h) = (\mathcal{Q}^{\widehat{u}_h}, \mathcal{U}^{\widehat{u}_h}) + (\mathcal{Q}^f, \mathcal{U}^f)$, where $\widehat{u}_h \in M_h(u_D)$ is the solution of*

$$a_h(\widehat{u}_h, \mu) = b_h(\mu) \quad \forall \mu \in M_h(0).$$

Here, $a_h(\mu, \eta) := \sum_{K \in \Omega_h} ((c \mathcal{Q}^\mu, \mathcal{Q}^\eta)_K + \langle \tau(\mathcal{U}^\mu - \mu), \mathcal{U}^\eta - \eta \rangle_{\partial K})$, and $b_h(\mu) := (f, \mathcal{U}^\mu)_\Omega$, for any $\mu, \eta \in M_h$.

Note that the formulation characterizing \widehat{u}_h is a rewriting of equations (3.2d) and (3.2e).

We can thus see that the HDG method can be implemented as a typical finite element method. Once the function \widehat{u}_h is computed, we can readily compute (\mathbf{q}_h, u_h) . For details of the implementation, see [34], where the HDG methods are shown to be *more* efficient than the CG method for high-degree polynomial approximations.

3.2. The stabilization mechanism.

The relation between the residuals. Note that, also for diffusion problems, the HDG method is defined by imposing a linear relation between the residuals in the interior of the element K , $\mathbf{R}_K^u := \mathbf{c}\mathbf{q}_h + \nabla u_h$ and $R_K^q := \nabla \cdot \mathbf{q}_h - f$, and the residuals on its boundary ∂K , $R_{\partial K}^u := u_h - \widehat{u}_h$ and $R_{\partial K}^q := (\mathbf{q}_h - \widehat{\mathbf{q}}_h) \cdot \mathbf{n} = -\tau(u_h - \widehat{u}_h)$.

Indeed, in terms of these residuals, the first two equations defining the HDG method read

$$\begin{aligned} (\mathbf{R}_K^u, \mathbf{v})_K &= \langle R_{\partial K}^u, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K} \quad \forall \mathbf{v} \in \mathbf{V}(K), \\ (R_K^q, w)_K &= \langle R_{\partial K}^q, w \rangle_{\partial K} \quad \forall w \in W(K). \end{aligned}$$

Since this implies that

$$\begin{aligned} \|\mathbf{P}_{\mathbf{V}(K)} \mathbf{R}_K^u\|_{L^2(K)} &\leq C h_K^{-1/2} \|R_{\partial K}^u\|_{L^2(\partial K)}, \\ \|\mathbf{P}_{W(K)} R_K^q\|_{L^2(K)} &\leq C h_K^{-1/2} \|R_{\partial K}^q\|_{L^2(\partial K)}, \end{aligned}$$

where $\mathbf{P}_{\mathbf{V}(K)}$ and $\mathbf{P}_{W(K)}$ are the L^2 -projections into $\mathbf{V}(K)$ and $W(K)$, respectively, we see that the quality of the approximation depends only on $u_h - \widehat{u}_h$, τ and on the approximation properties of the spaces $\mathbf{V}(K)$ and $W(K)$.

Stabilization by the jumps. The discrete energy identity (3.3) indicates that we can control the jumps $u_h - \widehat{u}_h$ if we take τ to be strictly positive. In this case, $\Theta_\tau(u_h - \widehat{u}_h)$ becomes a dissipative term which enhances the stability of the numerical method, just as for the original DG method. That this stabilization does not affect in a negative manner the accuracy of the method is shown next.

3.3. Convergence properties. Here we discuss the convergence properties of the method when Ω is a polyhedral domain, the triangulations Ω_h are made of shape-regular simplexes K , and when we take $\mathbf{V}(K) \times W(K) := \mathbf{P}_k(K) \times P_k(K)$. For simplicity, we assume that the stabilization function τ is *constant* on each ∂K .

The auxiliary projection. To do this, we follow what was done for the original DG method and define an auxiliary projection. On any simplex K , the

projection of the function $(\mathbf{q}, u) \in \mathbf{H}^1(K) \times H^1(K)$, $\Pi(\mathbf{q}, u) := (\mathbf{\Pi q}, \Pi u)$ is the element of $\mathbf{P}_k(K) \times \mathcal{P}_k(K)$ which solves the equations

$$\begin{aligned} (\mathbf{\Pi q}, \mathbf{v})_K &= (\mathbf{q}, \mathbf{v})_K & \forall \mathbf{v} \in \mathbf{P}_{k-1}(K), \\ (\Pi u, w)_K &= (u, w)_K & \forall w \in \mathcal{P}_{k-1}(K), \\ \langle \mathbf{\Pi q} \cdot \mathbf{n} + \tau \Pi u, \mu \rangle_F &= \langle \mathbf{q} \cdot \mathbf{n} + \tau u, \mu \rangle_F & \forall \mu \in \mathcal{P}_k(F), \end{aligned}$$

for all faces F of the simplex K . This projection is well defined, as we see in the next result.

Theorem 3.3. *Suppose that $\tau_K := \tau|_K$ is positive. Then $(\mathbf{\Pi q}, \Pi u)$ is well defined for any $k \geq 0$. Furthermore, there is a constant C independent of K and τ_K such that*

$$\begin{aligned} \|\mathbf{\Pi q} - \mathbf{q}\|_K &\leq C h_K^{k+1} |\mathbf{q}|_{\mathbf{H}^{k+1}(K)} + C h_K^{k+1} \tau_K |u|_{H^{k+1}(K)}, \\ \|\Pi u - u\|_K &\leq C h_K^{k+1} |u|_{H^{k+1}(K)} + C \frac{h_K^{k+1}}{\tau_K} |\nabla \cdot \mathbf{q}|_{H^k(K)}. \end{aligned}$$

Estimate of the projection of the errors. This projection is fitted to the structure of the numerical trace $\widehat{\mathbf{q}}_h$ because, if we consider the projection of the errors $(\varepsilon_{\mathbf{q}}, \varepsilon_u) := (\mathbf{\Pi q} - \mathbf{q}_h, \Pi u - u_h)$, we see that we have, by the last property of the projection,

$$P_{M_h}(\mathbf{q} \cdot \mathbf{n}) = \mathbf{\Pi q} \cdot \mathbf{n} + \tau(\Pi u - P_{M_h} u) \quad \text{on } \partial K.$$

Comparing this expression with the definition of the numerical trace $\widehat{\mathbf{q}}_h$, (3.2c),

$$\widehat{\mathbf{q}}_h \cdot \mathbf{n} = \mathbf{q}_h \cdot \mathbf{n} + \tau(u_h - \widehat{u}_h) \quad \text{on } \partial K,$$

we obtain that $\varepsilon_{\widehat{\mathbf{q}}} \cdot \mathbf{n} := \varepsilon_{\mathbf{q}} \cdot \mathbf{n} + \tau(\varepsilon_u - \varepsilon_{\widehat{u}})$ on ∂K , provided we set $\varepsilon_{\widehat{u}} := P_{M_h} u - \widehat{u}_h$ and $\varepsilon_{\widehat{\mathbf{q}}} \cdot \mathbf{n} := P_{M_h}(\mathbf{q}_h \cdot \mathbf{n}) - \widehat{\mathbf{q}}_h \cdot \mathbf{n}$, where P_{M_h} is the L^2 -projection into the space M_h . This implies that the equations satisfied by the projection of the errors are the following. For each simplex $K \in \Omega_h$,

$$\begin{aligned} (c \varepsilon_{\mathbf{q}}, \mathbf{v})_K - (\varepsilon_u, \nabla \cdot \mathbf{v})_K + \langle \varepsilon_{\widehat{u}}, \mathbf{v} \cdot \mathbf{n} \rangle_{\partial K} &= (c(\mathbf{\Pi q} - \mathbf{q}), \mathbf{v})_K, \\ -(\varepsilon_{\mathbf{q}}, \nabla w)_K + \langle \varepsilon_{\widehat{\mathbf{q}}} \cdot \mathbf{n}, w \rangle_{\partial K} &= 0, \end{aligned}$$

for all $(\mathbf{v}, w) \in \mathbf{V}(K) \times W(K)$. Moreover, for each face $F \in \mathcal{E}_h$,

$$\begin{aligned} \langle \mu, \llbracket \varepsilon_{\widehat{\mathbf{q}}} \cdot \mathbf{n} \rrbracket \rangle_F &= 0 & \forall \mu \in M(F), \\ \varepsilon_{\widehat{u}} &= 0 & \text{if } F \in \mathcal{E}_h^\partial. \end{aligned}$$

We thus see that the projections of the error solely depend on $\mathbf{\Pi q} - \mathbf{q}$. In

particular, the discrete energy identity for these equations is

$$(c\varepsilon_{\mathbf{q}}, \varepsilon_{\mathbf{q}})_{\Omega} + \Theta_{\tau}(\varepsilon_u - \varepsilon_{\widehat{u}}) = (c(\mathbf{\Pi}\mathbf{q} - \mathbf{q}), \varepsilon_{\mathbf{q}})_{\Omega},$$

and we can estimate the projection of error in the flux and in the jumps. In fact, we have the following result.

Theorem 3.4. *For $k \geq 0$, we have that*

$$\|\varepsilon_{\mathbf{q}}\|_{L^2(\Omega)} + \Theta_{\tau}^{1/2}(\varepsilon_u - \varepsilon_{\widehat{u}}) \leq C \|\mathbf{\Pi}\mathbf{q} - \mathbf{q}\|_{L^2(\Omega)}.$$

Moreover, if the elliptic regularity estimate $\|u\|_{H^2(\Omega)} \leq C \|\nabla \cdot (c\nabla u)\|_{L^2(\Omega)}$ holds when $u = 0$ on $\partial\Omega$, we have that

$$\|\varepsilon_u\|_{L^2(\Omega)} \leq C C_{\tau} h^{\min\{k,1\}} \|\mathbf{\Pi}\mathbf{q} - \mathbf{q}\|_{L^2(\Omega)},$$

where $C_{\tau} = \max_{K \in \Omega_h} \{1, h_K \tau_K\}$.

We can now conclude that, whenever the exact solution (\mathbf{q}, u) is smooth enough, and the stabilization function τ is of order one, we have that

$$\|\varepsilon_{\mathbf{q}}\|_{L^2(\Omega)} \leq C h^{k+1} \quad \text{and} \quad \|\varepsilon_u\|_{L^2(\Omega)} \leq C h^{k+1+\min\{k,1\}},$$

and so,

$$\|\mathbf{q} - \mathbf{q}_h\|_{L^2(\Omega)} \leq C h^{k+1} \quad \text{and} \quad \|u - u_h\|_{L^2(\Omega)} \leq C h^{k+1}.$$

Postprocessing. We can take advantage of the superconvergence of the projection of the error ε_u for $k \geq 1$ to define a better approximation u_h^* to u . The approximation u_h^* is defined on the simplex $K \in \Omega_h$ as the unique function in $\mathcal{P}_{k+1}(K)$ satisfying

$$\begin{aligned} (\nabla u_h^*, \nabla w)_K &= - (c\mathbf{q}_h, \nabla w)_K && \text{for all } w \in \mathcal{W}_{k+1}(K), \\ (u_h^*, w)_K &= (u_h, w)_K && \text{for all } w \in \mathcal{P}_{k-1}(K). \end{aligned}$$

Here $\mathcal{W}_{k+1}(K)$ denotes the $L^2(K)$ -orthogonal complement of $\mathcal{P}_{k-1}(K)$ in $\mathcal{P}_{k+1}(K)$. This projection is a modification of the projection proposed in [51, 32, 52].

Theorem 3.5. *We have that*

$$\|u - u_h^*\|_{L^2(\Omega)} \leq C C_{\tau} h^{\min\{k,1\}} \|\mathbf{\Pi}\mathbf{q} - \mathbf{q}\|_{L^2(\Omega)} + C h^{k+2} |u|_{H^{\ell+2}(\Omega_h)},$$

for any $k \geq 0$ where $C_{\tau} = \max_{K \in \Omega_h} \{1, h_K \tau_K\}$.

We thus conclude that, when the exact solution (\mathbf{q}, u) is smooth enough, the stabilization function τ is of order one and $k \geq 1$, we have that

$$\|u - u_h^*\|_{L^2(\Omega)} \leq C h^{k+2}.$$

3.4. Comparison with other finite element methods.

Finite element methods fitting the HDG formulation. Many finite element methods fit the formulation (3.2); the main examples are displayed in Table 1. In fact, the RT and BDM mixed methods can be viewed as particular cases of HDG methods and the CG method can be considered as a limiting case. This suggests that we could consider that the HDG methods are *between* the RT and BDM mixed methods and the CG method.

Table 1. Methods fitting the formulation (3.2) for triangulations Ω_h of simplexes.

Method	$V(K)$	$W(K)$	$M(F)$	τ
RT	$\mathcal{P}_k(K) + \mathbf{x} \mathcal{P}_k(K)$	$\mathcal{P}_k(K)$	$\mathcal{P}_k(F)$	0
BDM	$\mathcal{P}_k(K)$	$\mathcal{P}_{k-1}(K)$	$\mathcal{P}_k(F)$	0
HDG	$\mathcal{P}_k(K)$	$\mathcal{P}_k(K)$	$\mathcal{P}_k(F)$	$(0, \infty)$
CG	$\mathcal{P}_{k-1}(K)$	$\mathcal{P}_k(K)$	$\mathcal{P}_k(F)$	∞

Boundary residuals and accuracy. To further elaborate this idea, we compare in Table 2 how these methods deal with the residuals at the boundary of the elements and how this is reflected in their convergence properties.

Table 2. Residuals, stabilization and order of accuracy (for $k \geq 1$).

Method	$R_{\partial K}^u$	$R_{\partial K}^q$	τ	\mathbf{q}_h	u_h	u_h^*
RT	–	0	0	$k + 1$	$k + 1$	$k + 2$
BDM	–	0	0	$k + 1$	k	$k + 2$
HDG	–	–	$\mathcal{O}(1)$	$k + 1$	$k + 1$	$k + 2$
HDG	–	–	$\mathcal{O}(1/h)$	k	$k + 1$	$k + 1$
CG	0	–	∞	k	$k + 1$	$k + 1$

We see, on the one hand, that the RT and BDM methods force the residual $R_{\partial K}^q := (\mathbf{q}_h - \widehat{\mathbf{q}}_h) \cdot \mathbf{n}$ to be equal to zero and obtain the orders of convergence of $k + 1$ and $k + 2$ for \mathbf{q}_h and u_h^* , respectively. On the other hand, the CG method forces the residual $R_{\partial K}^u := u_h - \widehat{u}_h$ to be equal to zero and obtain the orders of convergence of only k and $k + 1$ for \mathbf{q}_h and u_h^* , respectively. (For a comparison between the RT and CG methods, see [19].) However, unlike these methods, the HDG method does not force any of these two residuals to be zero. Instead, it plays with the stabilization function τ to *balance* their relative sizes so that the approximation error is optimal. As we see in Table 2, this happens when both residuals have a similar weight, that is, when the stabilization function τ is of order one. Note that we would expect, from Table 1, that taking τ *small* or *big* enough would guarantee that the convergence properties of the HDG

method are closer to those of the RT and BDM methods or to the CG method, respectively. The results displayed in Table 2 actually confirm this. A rigorous explanation of this fact can be found in [22, 18].

4. HDG Methods for Incompressible Fluid Flow

In this section, we extend the HDG methods to the more involved Stokes equations of incompressible fluid flow,

$$-\nu \Delta \mathbf{u} + \nabla p = \mathbf{f} \quad \text{on } \Omega, \tag{4.1a}$$

$$\nabla \cdot \mathbf{u} = 0 \quad \text{on } \Omega, \tag{4.1b}$$

$$\mathbf{u} = \mathbf{u}_D \quad \text{on } \partial\Omega, \tag{4.1c}$$

$$(p, 1)_\Omega = 0, \tag{4.1d}$$

where $\langle \mathbf{u}_D \cdot \mathbf{n}, 1 \rangle_{\partial\Omega} = 0$. Most of the material for this section is contained in [43, 17].

4.1. The HDG methods.

Rewriting the equations. For each $K \in \Omega_h$, the method provides an approximation to the restriction of $(\mathbf{L}, \mathbf{u}, p)$ to K , where \mathbf{L} is the gradient of the velocity \mathbf{u} , as well as to the traces $(-\nu \widehat{\mathbf{L}} \mathbf{n} + \widehat{p} \mathbf{n}, \widehat{\mathbf{u}})$ on ∂K . So, we first rewrite the Stokes equations in a manner that will be suitable to defining the HDG methods.

If we assume that we know the trace of the velocity on ∂K , $\widehat{\mathbf{u}}$, as well as the average of the pressure on K , \bar{p} , we can obtain $(\mathbf{L}, \mathbf{u}, p)$ inside K as the solution of

$$\begin{aligned} \mathbf{L} - \nabla \mathbf{u} &= 0 && \text{in } K, \\ -\nu \nabla \cdot \mathbf{L} + \nabla p &= \mathbf{f} && \text{in } K, \\ \nabla \cdot \mathbf{u} &= \frac{1}{|K|} \langle \widehat{\mathbf{u}} \cdot \mathbf{n}, 1 \rangle_{\partial K} && \text{in } K, \\ \mathbf{u} &= \widehat{\mathbf{u}} && \text{on } \partial K, \\ \frac{1}{|K|} (p, 1)_K &= \bar{p}. \end{aligned}$$

The functions $\widehat{\mathbf{u}}$ and \bar{p} can now be obtained as the solution of

$$\begin{aligned} \llbracket -\nu \widehat{\mathbf{L}} \mathbf{n} + \widehat{p} \mathbf{n} \rrbracket &= 0 && \text{for all } F \in \mathcal{E}_h^o, \\ \langle \widehat{\mathbf{u}} \cdot \mathbf{n}, 1 \rangle_{\partial K} &= 0 && \text{for all } K \in \Omega_h, \\ \widehat{\mathbf{u}} &= \mathbf{u}_D && \text{on } \partial\Omega, \\ (\bar{p}, 1)_\Omega &= 0. \end{aligned}$$

The Galerkin method and the numerical traces. We now discretize the equations by using a Galerkin method together with a suitable approximation of the traces. On the element $K \in \Omega_h$, we take $(\mathbf{L}_h, \mathbf{u}_h, p_h)$ in the space $\mathbf{G}(K) \times \mathbf{V}(K) \times Q(K)$, and on the face $F \in \mathcal{E}_h^o$, we take $\widehat{\mathbf{u}}_h$ in the space $\mathbf{M}(F)$.

On each element $K \in \Omega_h$, the function $(\mathbf{L}_h, \mathbf{u}_h, p_h)$ is expressed in terms of $(\widehat{\mathbf{u}}_h, \bar{p}_h, f)$ as follows. Since

$$\begin{aligned} (\mathbf{L}, \mathbf{G})_K + (\mathbf{u}, \nabla \cdot \mathbf{G})_K - \langle \widehat{\mathbf{u}}, \mathbf{G}\mathbf{n} \rangle_{\partial K} &= 0, \\ (\nu \mathbf{L}, \nabla \mathbf{v})_K - (p_h, \nabla \cdot \mathbf{v})_K - \langle \nu \widehat{\mathbf{L}}\mathbf{n} - \widehat{p}_h\mathbf{n}, \mathbf{v} \rangle_{\partial K} &= (\mathbf{f}, \mathbf{v})_K, \\ -(\mathbf{u}, \nabla q)_K + \langle \widehat{\mathbf{u}} \cdot \mathbf{n}, q - \bar{q} \rangle_{\partial K} &= 0, \end{aligned}$$

for smooth enough $(\mathbf{G}, \mathbf{v}, q)$, we determine $(\mathbf{L}_h, \mathbf{u}_h, p_h)$ in terms of $(\widehat{\mathbf{u}}_h, \bar{p}_h, f)$ as the solution of

$$(\mathbf{L}_h, \mathbf{G})_K + (\mathbf{u}_h, \nabla \cdot \mathbf{G})_K - \langle \widehat{\mathbf{u}}_h, \mathbf{G}\mathbf{n} \rangle_{\partial K} = 0, \tag{4.2a}$$

$$(\nu \mathbf{L}_h, \nabla \mathbf{v})_K - (p_h, \nabla \cdot \mathbf{v})_K - \langle \nu \widehat{\mathbf{L}}_h\mathbf{n} - \widehat{p}_h\mathbf{n}, \mathbf{v} \rangle_{\partial K} = (\mathbf{f}, \mathbf{v})_K, \tag{4.2b}$$

$$-(\mathbf{u}_h, \nabla q)_{\Omega_h} + \langle \widehat{\mathbf{u}}_h \cdot \mathbf{n}, q - \bar{q} \rangle_{\partial K} = 0, \tag{4.2c}$$

$$\frac{1}{|K|} (p_h, 1)_K = \bar{p}_h, \tag{4.2d}$$

for all $(\mathbf{G}, \mathbf{v}, q) \in \mathbf{G}(K) \times \mathbf{V}(K) \times P_h(K)$, where the numerical trace $-\nu \widehat{\mathbf{L}}_h\mathbf{n} + \widehat{p}_h\mathbf{n}$ is assumed to be given by

$$-\nu \widehat{\mathbf{L}}_h\mathbf{n} + \widehat{p}_h\mathbf{n} = -\nu \mathbf{L}_h\mathbf{n} + p_h\mathbf{n} + \nu \tau (\mathbf{u}_h - \widehat{\mathbf{u}}_h) \quad \text{on } \partial K. \tag{4.2e}$$

Then we determine $(\widehat{\mathbf{u}}_h, \bar{p}_h)$ by enforcing the remaining equations, that is, by requiring that

$$\langle [-\nu \widehat{\mathbf{L}}_h\mathbf{n} + \widehat{p}_h\mathbf{n}], \boldsymbol{\mu} \rangle_F = 0 \quad \forall \boldsymbol{\mu} \in \mathbf{M}(F) \quad \forall F \in \mathcal{E}_h^o, \tag{4.2f}$$

$$\langle \widehat{\mathbf{u}}_h \cdot \mathbf{n}, 1 \rangle_{\partial K} = 0 \quad \forall K \in \Omega_h, \tag{4.2g}$$

$$\widehat{\mathbf{u}}_h = \mathbf{u}_D \quad \text{on } \partial\Omega, \tag{4.2h}$$

$$(\bar{p}_h, 1)_{\Omega} = 0. \tag{4.2i}$$

This completes the definition of the HDG methods.

The discrete energy identity and the existence and uniqueness of the approximation. These methods are well defined under very mild conditions, as we see in the next result.

Proposition 4.1. *The HDG method is well defined if (i) the stabilization function τ is strictly positive on $\partial\Omega_h$, (ii) $\nabla \mathbf{v} \in \mathbf{G}(K)$ for any $\mathbf{v} \in \mathbf{V}(K)$, and if (iii) $\nabla q \in \mathbf{V}(K)$ for all $q \in Q(K)$.*

To prove this result, we begin by establishing an energy identity. Note that for the exact solution we have

$$(L, L)_\Omega = (\mathbf{f}, \mathbf{u})_\Omega + \langle -\nu L\mathbf{n} + p \mathbf{n}, \mathbf{u}_D \rangle_{\partial\Omega},$$

and so we should have a similar energy identity for the solution of the HDG method. Indeed, it is not difficult to obtain

$$(L_h, L_h)_\Omega + \Theta_\tau(\mathbf{u}_h - \widehat{\mathbf{u}}_h) = (\mathbf{f}, \mathbf{u}_h)_\Omega + \langle (-\nu \widehat{L}_h + \widehat{p}_h \mathbf{I})\mathbf{n}, \mathbf{u}_D \rangle_{\partial\Omega}, \tag{4.3}$$

where $\Theta_\tau(\mathbf{u}_h - \widehat{\mathbf{u}}_h) := \sum_{K \in \Omega_h} \langle \tau(\mathbf{u}_h - \widehat{\mathbf{u}}_h), \mathbf{u}_h - \widehat{\mathbf{u}}_h \rangle_{\partial K}$.

Once again, to prove Proposition 4.1, we only have to show that when if we set $\mathbf{f} = \mathbf{0}$ and $\mathbf{u}_D = \mathbf{0}$, the only solution is the trivial one. By the discrete energy identity, we see that in this case we have that $L_h = 0$ on Ω_h and that $\mathbf{u}_h = \widehat{\mathbf{u}}_h$ on \mathcal{E}_h . By equation (4.2a), this implies that

$$(\nabla \mathbf{u}, \mathbf{G})_K = 0 \quad \forall \mathbf{G} \in \mathbf{G}(K),$$

and by condition (ii), we conclude that \mathbf{u}_h is constant on Ω . Since $\mathbf{u}_h = \widehat{\mathbf{u}}_h = \mathbf{0}$ on $\partial\Omega$, we see that $\mathbf{u}_h = \mathbf{0}$ on Ω and, as a consequence, that $\widehat{\mathbf{u}}_h = \mathbf{0}$ on \mathcal{E}_h . Finally, by equation (4.2b),

$$(\nabla p_h, \mathbf{v})_K = 0 \quad \forall \mathbf{v} \in \mathbf{V}(K),$$

and by condition (iii), we have that p_h is a constant on Ω . By equations (4.2d) and (4.2i), we conclude that $p_h = \bar{p}_h = 0$ on Ω_h . This completes the proof of Proposition 4.1.

Implementation. To describe the implementation of the HDG methods, we need to introduce some notation. We denote by $(\mathcal{L}, \mathcal{U}, \mathcal{P})$ the linear mapping (given by equations (4.2a) to (4.2e)) that associates $(\widehat{\mathbf{u}}_h, \bar{p}_h, f)$ to (L_h, \mathbf{u}_h, p_h) , and set

$$\begin{aligned} (\mathcal{L}^{\widehat{\mathbf{u}}_h}, \mathcal{U}^{\widehat{\mathbf{u}}_h}, \mathcal{P}^{\widehat{\mathbf{u}}_h}) &:= (\mathcal{L}, \mathcal{U}, \mathcal{P})(\widehat{\mathbf{u}}_h, 0, 0), \\ (\mathcal{L}^{\bar{p}_h}, \mathcal{U}^{\bar{p}_h}, \mathcal{P}^{\bar{p}_h}) &:= (\mathcal{L}, \mathcal{U}, \mathcal{P})(0, \bar{p}_h, 0), \\ (\mathcal{L}^f, \mathcal{U}^f, \mathcal{P}^f) &:= (\mathcal{L}, \mathcal{U}, \mathcal{P})(0, 0, f). \end{aligned}$$

We also introduce the spaces

$$\begin{aligned} \mathbf{M}_h &:= \{ \boldsymbol{\mu} \in \mathbf{L}^2(\mathcal{E}_h) : \boldsymbol{\mu}|_F \in \mathbf{M}(F) \quad \forall F \in \mathcal{E}_h^o \}, \\ \bar{\mathcal{P}}_h &:= \{ \bar{q}_h \in L^2(\Omega) : \bar{q}_h \in \mathcal{P}_0(K) \quad \forall K \in \Omega_h \}, \end{aligned}$$

and set $\mathbf{M}_h(\boldsymbol{\zeta}) := \{ \boldsymbol{\mu} \in \mathbf{M}_h : \boldsymbol{\mu}|_{\partial\Omega} = \boldsymbol{\zeta} \}$.

With this notation, we can characterize the approximate solution given by the HDG method as follows.

Theorem 4.2. *We have that*

$$(\mathbf{L}_h, \mathbf{u}_h, p_h) = (\mathcal{L}^{\hat{\mathbf{u}}_h}, \mathbf{U}^{\hat{\mathbf{u}}_h}, \mathcal{P}^{\hat{\mathbf{u}}_h}) + (\mathcal{L}^{\bar{p}_h}, \mathbf{U}^{\bar{p}_h}, \mathcal{P}^{\bar{p}_h}) + (\mathcal{L}^f, \mathbf{U}^f, \mathcal{P}^f),$$

where $(\hat{\mathbf{u}}_h, \bar{p}_h)$ is the only element in $\mathbf{M}_h(\mathbf{u}_D) \times \bar{P}_h$ such that

$$\begin{aligned} a_h(\hat{\mathbf{u}}_h, \boldsymbol{\mu}) + b_h(\bar{p}_h, \boldsymbol{\mu}) &= \ell_h(\boldsymbol{\mu}), \quad \forall \boldsymbol{\mu} \in \mathbf{M}_h(\mathbf{0}), \\ b_h(\bar{q}, \hat{\mathbf{u}}_h) &= 0, \quad \forall \bar{q} \in \bar{P}_h, \\ (\bar{p}_h, 1)_\Omega &= 0. \end{aligned}$$

Here the forms are given by

$$\begin{aligned} a_h(\boldsymbol{\eta}, \boldsymbol{\mu}) &= \sum_{K \in \Omega_h} ((\nu \mathcal{L}^\boldsymbol{\eta}, \mathcal{L}^\boldsymbol{\mu})_K + \langle \nu \boldsymbol{\tau}(\mathbf{U}^\boldsymbol{\eta} - \boldsymbol{\eta}), (\mathbf{U}^\boldsymbol{\mu} - \boldsymbol{\mu}) \rangle_{\partial K}), \\ b_h(\bar{q}, \boldsymbol{\mu}) &= - \sum_{K \in \Omega_h} \langle \bar{q}, \boldsymbol{\mu} \cdot \mathbf{n} \rangle_{\partial K}, \\ \ell_h(\boldsymbol{\mu}) &= (\mathbf{f}, \mathbf{U}^\boldsymbol{\mu})_\Omega, \end{aligned}$$

for all $\boldsymbol{\eta} \in \mathbf{M}_h, \boldsymbol{\mu} \in \mathbf{M}_h$, and $\bar{q} \in \bar{P}_h$.

Note that the first equation of the formulation characterizing $(\hat{\mathbf{u}}_h, \bar{p}_h)$ is a rewriting of equation (4.2f) whereas the second is a rewriting of equation (4.2g).

We see that the HDG method can be implemented as a typical mixed method. In fact, an augmented lagrangian algorithm can be used to further improve its implementation; see [43].

4.2. The stabilization mechanism.

The relation between the residuals. Note that also for The Stokes problem, the HDG method is defined by imposing a linear relation between the residuals in the interior of the element K , $\mathbf{R}_K^{\mathbf{u}} := \mathbf{L}_h - \nabla \mathbf{u}_h$, $\mathbf{R}_K^{\mathbf{L},p} := \nabla \cdot (-\nu \mathbf{L}_h + p_h \mathbf{I}) - \mathbf{f}$, and $\mathbf{R}_K^{\nabla \cdot \mathbf{u}} := \nabla \cdot \mathbf{u}_h$ and the residuals on its boundary ∂K , $\mathbf{R}_{\partial K}^{\mathbf{u}} := (\hat{\mathbf{u}}_h - \mathbf{u}_h) \otimes \mathbf{n}$ and $\mathbf{R}_{\partial K}^{\mathbf{L},p} := (-\nu \mathbf{L}_h \mathbf{n} + p_h \mathbf{n}) - (-\nu \hat{\mathbf{L}}_h \mathbf{n} + \hat{p}_h \mathbf{n}) = -\nu \boldsymbol{\tau}(\mathbf{u}_h - \hat{\mathbf{u}}_h)$.

Indeed, the equations of the Galerkin method defining the HDG method can be rewritten as follows:

$$\begin{aligned} (\mathbf{R}_K^{\mathbf{u}}, \mathbf{G})_K &= \langle \mathbf{R}_{\partial K}^{\mathbf{u}}, \mathbf{G} \rangle_{\partial K} \\ (\mathbf{R}_K^{\mathbf{L},p}, \mathbf{v})_K &= \langle \mathbf{R}_{\partial K}^{\mathbf{L},p}, \mathbf{v} \rangle_{\partial K}, \\ (\mathbf{R}_K^{\nabla \cdot \mathbf{u}}, q)_K &= \langle \text{tr} \mathbf{R}_{\partial K}^{\mathbf{u}}, q \rangle_{\partial K}. \end{aligned}$$

Since this implies that

$$\begin{aligned} \|\mathbf{P}_{\mathbf{G}(K)} \mathbf{R}_K^{\mathbf{u}}\|_{L^2(K)} &\leq C h_K^{-1/2} \|\mathbf{R}_{\partial K}^{\mathbf{u}}\|_{L^2(\partial K)}, \\ \|\mathbf{P}_{\mathbf{V}(K)} \mathbf{R}_K^{\mathbf{L},p}\|_{L^2(K)} &\leq C h_K^{-1/2} \|\mathbf{R}_{\partial K}^{\mathbf{L},p}\|_{L^2(\partial K)}, \\ \|\mathbf{P}_{\mathbf{Q}(K)} \mathbf{R}_K^{\nabla \cdot \mathbf{u}}\|_{L^2(K)} &\leq C h_K^{-1/2} \|\text{tr} \mathbf{R}_{\partial K}^{\mathbf{u}}\|_{L^2(\partial K)}, \end{aligned}$$

we see that the quality of the approximation depends only on $\mathbf{u} - \widehat{\mathbf{u}}_h$, τ and on the approximation properties of the spaces $\mathbf{G}(K)$, $\mathbf{V}(K)$ and $Q(K)$.

Stabilization by the jumps. By the energy identity (4.3), we see that we can control the jumps $\mathbf{u}_h - \widehat{\mathbf{u}}_h$ if we require the stabilization function τ to be positive on $\partial\Omega_h$. Next, we show that this stabilization mechanism does not spoil the accuracy of the method.

4.3. Convergence properties. Here we discuss the convergence properties of the method when Ω is a polyhedral domain, the triangulations Ω_h are made of shape-regular simplexes K , and when we take

$$\mathbf{G}(K) \times \mathbf{V}(K) \times W(K) := \mathbf{P}_K(K) \times \mathbf{P}_k(K) \times P_k(K).$$

Here $\mathbf{P}_k(K)$ is the space of matrix-valued functions whose components belong to $P_k(K)$. Once again, for simplicity, we assume that the stabilization function τ is constant on each ∂K .

The auxiliary projection. Given a simplex $K \in \Omega_h$ and a function $(\mathbf{L}, \mathbf{u}, p)$ in $H^1(K) \times \mathbf{H}^1(K) \times H^1(K)$, we define its projection $\Pi(\mathbf{L}, \mathbf{u}, p) := (\Pi\mathbf{L}, \Pi\mathbf{u}, \Pi p)$ as the element of $\mathbf{G}_h \times \mathbf{V}_h \times P_h$ that solves the equations

$$\begin{aligned} (\Pi\mathbf{L}, \mathbf{G})_K &= (\mathbf{L}, \mathbf{G})_K & \forall \mathbf{G} \in \mathbf{P}_{k-1}(K), \\ (\Pi\mathbf{u}, \mathbf{v})_K &= (\mathbf{u}, \mathbf{v})_K & \forall \mathbf{v} \in \mathcal{P}_{k-1}(K), \\ (\Pi p, q)_K &= (p, q)_K & \forall q \in \mathcal{P}_{k-1}(K), \\ (\text{tr } \Pi\mathbf{L}, q)_K &= (\text{tr } \mathbf{L}, q)_K & \forall q \in \mathcal{P}_k(K), \\ \langle \nu \Pi\mathbf{L}\mathbf{n} - \Pi p \mathbf{n} - \nu \tau \Pi\mathbf{u}, \boldsymbol{\mu} \rangle_F &= \langle \nu \mathbf{L}\mathbf{n} - p \mathbf{n} - \nu \tau \mathbf{u}, \boldsymbol{\mu} \rangle_F & \forall \boldsymbol{\mu} \in \mathcal{P}_k(F), \end{aligned}$$

for all faces F of the simplex K . This projection is actually well defined.

Theorem 4.3. *Suppose that $\tau_K := \tau|_K$ is a positive constant on ∂K . Then the projection Π is well defined. Moreover, on each element $K \in \Omega_h$, we have that*

$$\begin{aligned} \|\Pi\mathbf{u} - \mathbf{u}\|_K &\leq C h^{k+1} (|\mathbf{u}|_{\mathbf{H}^{k+1}(K)} + \tau_K^{-1} |\mathbf{u}|_{\mathbf{H}^{k+2}(K)}), \\ \|\nu \Pi\mathbf{L} - \nu \mathbf{L}\|_K &\leq C h^{k+1} \nu (|\mathbf{u}|_{\mathbf{H}^{k+2}(K)} + \tau_K |\mathbf{u}|_{\mathbf{H}^{k+1}(K)}), \\ \|\Pi p - p\|_K &\leq C h^{k+1} |p|_{H^{k+1}(K)} + C \|\nu \Pi\mathbf{L} - \nu \mathbf{L}\|_K. \end{aligned}$$

We have assumed that $\text{tr } \mathbf{L} = 0$ for the last two inequalities and that $\nabla \cdot \mathbf{u} = 0$ in the last one.

Estimates of the projection of the errors. This projection is fitted to the structure of the numerical trace $-\nu \widehat{\mathbf{L}}\mathbf{n} + \widehat{p} \mathbf{n}$ in the following sense. Consider

the projection of the errors $(E_L, \boldsymbol{\varepsilon}_u, \varepsilon_p) := (\Pi L - L_h, \mathbf{I}\mathbf{u} - \mathbf{u}_h, \Pi p - p_h)$. Then, by the last equation defining this projection, we have that

$$\mathbf{P}_M(-\nu L\mathbf{n} + p\mathbf{n}) = -\nu \Pi L\mathbf{n} + \Pi p\mathbf{n} + \nu \tau(\mathbf{I}\mathbf{u} - \mathbf{P}_M\mathbf{u}),$$

where \mathbf{P}_M is the L^2 -projection into \mathbf{M}_h . Comparing this with the definition of the numerical trace

$$-\nu \widehat{L}_h\mathbf{n} + \widehat{p}_h\mathbf{n} = -\nu L_h\mathbf{n} + p_h\mathbf{n} + \nu \tau(\mathbf{u}_h - \widehat{\mathbf{u}}_h),$$

we get that $-\nu \varepsilon_{\widehat{L}}\mathbf{n} + \varepsilon_{\widehat{p}}\mathbf{n} = -\nu E_L\mathbf{n} + \varepsilon_p\mathbf{n} + \tau(\boldsymbol{\varepsilon}_u - \boldsymbol{\varepsilon}_{\widehat{u}})$ provided $-\nu \varepsilon_{\widehat{L}}\mathbf{n} + \varepsilon_{\widehat{p}}\mathbf{n} := \mathbf{P}_M(-\nu L\mathbf{n} + p\mathbf{n}) - (-\nu \widehat{L}_h\mathbf{n} + \widehat{p}_h\mathbf{n})$ and $\boldsymbol{\varepsilon}_{\widehat{u}} := \mathbf{P}_M(\mathbf{u} - \widehat{\mathbf{u}})$.

The equations satisfied by the projection of the errors are then the following. For each simplex $K \in \Omega_h$,

$$\begin{aligned} (E_L, \mathbf{G})_K + (\boldsymbol{\varepsilon}_u, \nabla \cdot \mathbf{G})_K - \langle \boldsymbol{\varepsilon}_{\widehat{u}}, \mathbf{G}\mathbf{n} \rangle_{\partial K} &= (\Pi L - L, \mathbf{G})_K, \\ -(\nabla \cdot (\nu E_L), \mathbf{v})_K + (\nabla \varepsilon_p, \mathbf{v})_K + \langle \nu \tau(\boldsymbol{\varepsilon}_u - \boldsymbol{\varepsilon}_{\widehat{u}}), \mathbf{v} \rangle_{\partial K} &= 0, \\ -(\boldsymbol{\varepsilon}_u, \nabla q)_K + \langle \boldsymbol{\varepsilon}_{\widehat{u}}, q\mathbf{n} \rangle_{\partial K} &= 0, \end{aligned}$$

for all $(\mathbf{G}, \mathbf{v}, q)$ in $\mathbf{G}(K) \times \mathbf{V}(K) \times Q(K)$. Moreover,

$$\begin{aligned} \langle -\nu E_L\mathbf{n} + \varepsilon_p\mathbf{n} + \nu \tau(\boldsymbol{\varepsilon}_u - \boldsymbol{\varepsilon}_{\widehat{u}}), \boldsymbol{\mu} \rangle_F &= 0 \quad \forall \boldsymbol{\mu} \in \mathbf{M}(F) \quad \forall F \in \mathcal{E}_h^o, \\ \boldsymbol{\varepsilon}_{\widehat{u}} &= 0 \quad \text{on } \partial\Omega, \\ (\varepsilon_p, 1)_\Omega &= (\Pi p - p, 1)_\Omega. \end{aligned}$$

We thus see that the projection of the errors only depend on $\Pi L - L$ and on $(\Pi p - p, 1)_\Omega$, the latter quantity being equal to zero for $k \geq 1$.

In particular, the discrete energy identity for the equations is

$$(E_L, E_L)_\Omega + \Theta_\tau(\boldsymbol{\varepsilon}_u - \boldsymbol{\varepsilon}_{\widehat{u}}) = (\Pi L - L, E_L)_\Omega,$$

and we immediately obtain an estimate of the projection of the error in the gradient and in the jumps of the velocity. In fact, we can prove the following result.

Theorem 4.4. *We have*

$$\begin{aligned} \|E_L\|_{L^2(\Omega)} + \Theta_\tau^{1/2}(\mathbf{u} - \widehat{\mathbf{u}}_h) &\leq C \|\Pi L - L\|_{L^2(\Omega)}, \\ \|\varepsilon_p\|_{L^2(\Omega)} &\leq |(\Pi p - p, 1)_\Omega| |\Omega|^{-1/2} + C \sqrt{C_\tau} \nu \|\Pi L - L\|_{L^2(\Omega)}, \end{aligned}$$

where $C_\tau := \max_{K \in \Omega_h} \{1, \tau_K h_K\}$. Moreover, if the elliptic regularity estimate $\nu \|\mathbf{u}\|_{\mathbf{H}^2(\Omega)} \leq C \|-\nu \Delta \mathbf{u} + \nabla p\|_{L^2(\Omega)}$ holds whenever $\mathbf{u} = \mathbf{0}$ on $\partial\Omega$, we have that

$$\|\boldsymbol{\varepsilon}_u\|_{L^2(\Omega)} \leq C C_\tau h^{\min\{k, 1\}} \|\Pi L - L\|_{L^2(\Omega)}.$$

We can now conclude that, whenever the exact solution (\mathbf{q}, u) is smooth enough, and the stabilization function τ is of order one, we have that

$$\nu \|E_L\|_{L^2(\Omega)} + \|\varepsilon_p\|_{L^2(\Omega)} \leq C h^{k+1} \quad \text{and} \quad \|\varepsilon_u\|_{L^2(\Omega)} \leq C h^{k+1+\min\{k,1\}},$$

and so,

$$\nu \|L - L_h\|_{L^2(\Omega)} + \|p - p_h\|_{L^2(\Omega)} \leq C h^{k+1} \quad \text{and} \quad \|u - u_h\|_{L^2(\Omega)} \leq C h^{k+1}.$$

Postprocessing. Here we show how to obtain a new approximate velocity which is exactly divergence-free, $\mathbf{H}(\text{div})$ -conforming, and converges with an additional order for $k \geq 1$. We only describe the three dimensional case, as the two dimensional case is much simpler. We quote [17] almost *verbatim*.

In the *three dimensional case* we define the postprocessed approximate velocity \mathbf{u}_h^* on the tetrahedron $K \in \Omega_h$ as the element of $\mathcal{P}_{k+1}(K)$ such that

$$\begin{aligned} \langle (\mathbf{u}_h^* - \widehat{\mathbf{u}}_h) \cdot \mathbf{n}, \mu \rangle_F &= 0 \quad \forall \mu \in \mathcal{P}_k(F), \\ \langle (\mathbf{n} \times \nabla)(\mathbf{u}_h^* \cdot \mathbf{n}) - \mathbf{n} \times (\{\{L_h^t\}\}\mathbf{n}), (\mathbf{n} \times \nabla)\mu \rangle_F &= 0 \quad \forall \mu \in \mathcal{P}_{k+1}(F)^\perp, \end{aligned}$$

for all faces F of K , and such that

$$\begin{aligned} (\mathbf{u}_h^* - \mathbf{u}_h, \nabla w)_K &= 0 \quad \forall w \in \mathcal{P}_k(K), \\ (\nabla \times \mathbf{u}_h^* - \mathbf{w}_h, (\nabla \times \mathbf{v}) B_K)_K &= 0 \quad \forall \mathbf{v} \in \mathcal{S}_k(K). \end{aligned}$$

Here $\mathcal{P}_{k+1}(F)^\perp := \{\mu \in \mathcal{P}_{k+1}(F) : \langle \mu, \widetilde{\mu} \rangle_F = 0, \quad \forall \widetilde{\mu} \in \mathcal{P}_k(F)\}$, $\mathbf{n} \times \nabla$ is the tangential gradient rotated $\pi/2$ in the positive sense (from the point of view of the normal vector) and the function $\{\{L_h^t\}\}$ is the single-valued function on \mathcal{E}_h equal to $((L_h^t)^+ + (L_h^t)^-)/2$ on the set $\mathcal{E}_h \setminus \partial\Omega$ and equal to L_h^t on $\partial\Omega$. In the last equation, we have that $\mathbf{w}_h := (L_{32}^h - L_{23}^h, L_{13}^h - L_{31}^h, L_{21}^h - L_{12}^h)$ is the approximation to the vorticity and $B_K := \sum_{\ell=0}^3 \lambda_{\ell-3} \lambda_{\ell-2} \lambda_{\ell-1} \nabla \lambda_\ell \otimes \nabla \lambda_\ell$ is the so-called *symmetric bubble matrix* introduced in [15]. Here the λ_i 's are the barycentric coordinates associated with the tetrahedron K , the subindices being counted modulo 4. Finally, to define $\mathcal{S}_k(K)$, recall the Nédélec space of the first kind [36], defined by $\mathbf{N}_k = \mathcal{P}_{k-1}(K) \oplus \mathcal{S}_k$, where \mathcal{S}_ℓ is the space of vector-valued homogeneous polynomials \mathbf{v} of degree ℓ such that $\mathbf{v} \cdot \mathbf{x} = 0$. Then, define $\mathcal{S}_k(K) := \{\mathbf{p} \in \mathbf{N}_k : (\mathbf{p}, \nabla \phi)_K = 0 \text{ for all } \phi \in \mathcal{P}_{k+1}(K)\}$.

Theorem 4.5. *We have that $\mathbf{u}_h^* \in \mathbf{H}(\text{div}, \Omega)$ and that $\nabla \cdot \mathbf{u}_h^* = 0$ on Ω . Moreover,*

$$\|\mathbf{u}_h^* - \mathbf{u}\|_{L^2(\Omega)} \leq C h^{k+2} \|\mathbf{u}\|_{\mathbf{H}^{\ell_{\mathbf{u}+2}(\Omega)}} + C C_\tau h^{\min\{k,1\}} \|\Pi L - L\|_{L^2(\Omega)}.$$

We thus conclude that, when the exact solution (L, \mathbf{u}, p) is smooth enough, the stabilization function τ is of order one and $k \geq 1$, we have that

$$\|\mathbf{u} - \mathbf{u}_h^*\|_{L^2(\Omega)} \leq C h^{k+2}.$$

5. Conclusion and Ongoing Work

The described approach to devise HDG methods has proven to be very powerful for the model problems considered in the previous sections. We believe that it can be used in a systematic manner to obtain efficiently implementable and accurate HDG methods for a wide variety of problems of practical interest. In fact, many HDG methods have already been defined and numerically tested on a variety of problems; their analyses constitute the subject of ongoing research. To end this paper, we describe them and briefly discuss their main convergence properties.

HDG methods have been devised for linear, steady-state convection-diffusion problems in [13], and for time-dependent linear and nonlinear convection-diffusion problems in [41] and [42], respectively. The convergence properties for HDG methods for the purely diffusive case seem to carry over to all these problems in the diffusion-dominated regime.

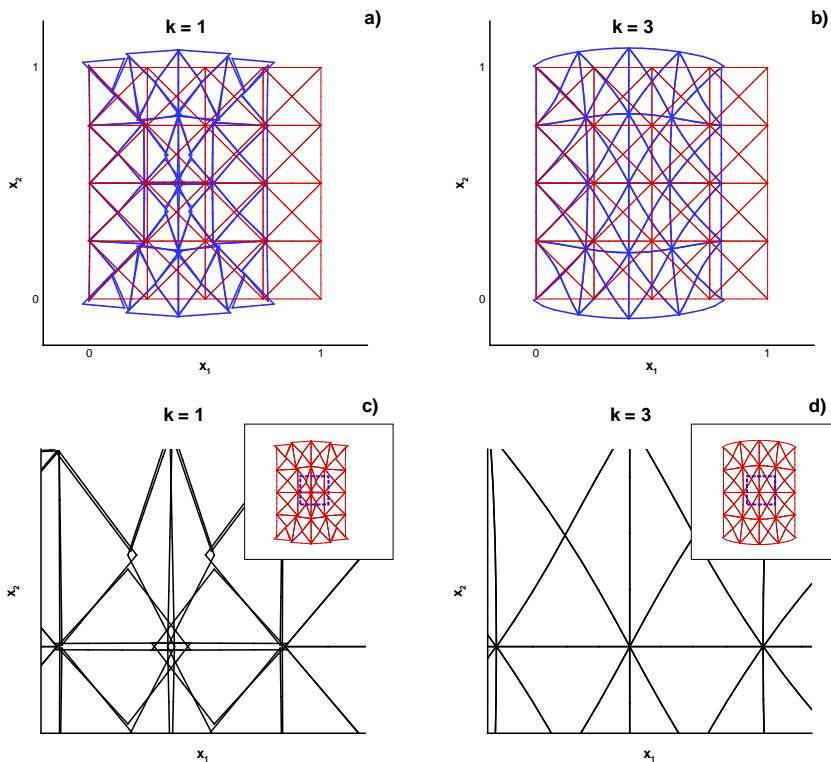


Figure 2. a) deformed shape using \mathcal{P}^1 , b) deformed shape using \mathcal{P}^3 , c) closeup view of Figure a), d) closeup view of Figure b).

HDG methods for linear and nonlinear elasticity have been devised in [50] and [49] with very good results. Indeed, by using polynomial approximations of degree k for all the components of the stress and displacement, the order of convergence of $k + 1$ for $k \geq 0$ in all variables is obtained. Moreover, by means of a local postprocessing, a new approximation of the displacement can be computed which converges with order $k + 2$ for $k \geq 2$. In Fig. 2, we show the approximate displacement of the borders of the elements for a nonlinear elasticity problem; see [49] for a detailed description. The approximation with polynomials of degree one is not as good as the approximation using polynomials of degree three. In full agreement with the properties of the HDG methods, this is reflected in the fact that the jumps for the former are highly visible whereas those of the latter are not.

HDG methods for Timoshenko beams have been developed in [5] with optimal convergence results. Indeed, if polynomials of degree k are used to approximate the displacement, rotation angle, bending moment and shear stress, numerical experiments suggest that all of these variables converge with order $k + 1$ for $k \geq 0$. For biharmonic problems [12], the HDG methods provide the optimal order of convergence of $k + 1$ for the scalar variable and its gradient. However, the approximation of the laplacian is only of order $k + 1/2$ and that of its gradient of only order $k - 1/2$. On the other hand, on strict subdomains, the order is the optimal $k + 1$ for all these variables. Further analysis of this phenomenon is required to obtain an optimally convergent HDG method on the the whole domain.

HDG methods for vorticity-velocity-pressure formulations of the Stokes equations have been proposed in [14] and later numerically compared with other HDG methods in [38]. The results indicate that the HDG method considered in the previous section performs better. Extensions of this HDG method for the incompressible Navier-Stokes has been recently proposed in [39, 37]. Once again, all the convergence properties of the HDG methods seem to carry over to these equations.

Finally, we would like to report that the HDG methods for both the Euler equations of gas dynamics and the compressible Navier-Stokes equations been devised in [44] seem to provide approximations with optimally convergent properties.

References

- [1] D. N. Arnold, F. Brezzi, B. Cockburn, and L. D. Marini, *Unified analysis of discontinuous Galerkin methods for elliptic problems*, SIAM J. Numer. Anal. **39** (2002), 1749–1779.
- [2] F. Bassi and S. Rebay, *A high-order accurate discontinuous finite element method for the numerical solution of the compressible Navier-Stokes equations*, J. Comput. Phys. **131** (1997), 267–279.

- [3] F. Brezzi, B. Cockburn, L. D. Marini, and E. Süli, *Stabilization mechanisms in discontinuous Galerkin finite element methods*, Comput. Methods Appl. Mech. Engrg. **195** (2006), 3293–3310, C. Dawson, Ed.
- [4] F. Brezzi, J. Douglas, Jr., and L. D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math. **47** (1985), 217–235.
- [5] F. Celiker, B. Cockburn, and K. Shi, *Hybridizable discontinuous Galerkin methods for Timoshenko beams*, J. Sci. Comput., to appear.
- [6] B. Cockburn, *Devising discontinuous Galerkin methods for non-linear hyperbolic conservation laws*, Journal of Computational and Applied Mathematics **128** (2001), 187–204.
- [7] ———, *Discontinuous Galerkin methods*, ZAMM Z. Angew. Math. Mech. **83** (2003), 731–754.
- [8] ———, *Discontinuous Galerkin Methods for Computational Fluid Dynamics*, Encyclopedia of Computational Mechanics (R. de Borst E. Stein and T.J.R. Hughes, eds.), vol. 3, John Wiley & Sons, Ltd., England, 2004, pp. 91–123.
- [9] B. Cockburn, B. Dong, and J. Guzmán, *Optimal convergence of the original DG method on special meshes for variable convective velocity*, SIAM J. Numer. Anal., to appear.
- [10] ———, *Optimal convergence of the original DG method for the transport-reaction equation on special meshes*, SIAM J. Numer. Anal. **46** (2008), 1250–1265.
- [11] ———, *A superconvergent LDG-hybridizable Galerkin method for second-order elliptic problems*, Math. Comp. **77** (2008), 1887–1916.
- [12] ———, *A hybridizable and superconvergent discontinuous Galerkin method for biharmonic problems*, J. Sci. Comput. **40** (2009), 141–187.
- [13] B. Cockburn, B. Dong, J. Guzmán, M. Restelli, and R. Sacco, *Superconvergent and optimally convergent LDG-hybridizable discontinuous Galerkin methods for convection-diffusion-reaction problems*, SIAM J. Sci. Comput., to appear.
- [14] B. Cockburn and J. Gopalakrishnan, *The derivation of hybridizable discontinuous Galerkin methods for Stokes flow*, SIAM J. Numer. Anal. **47** (2009), 1092–1125.
- [15] B. Cockburn, J. Gopalakrishnan, and J. Guzmán, *A new elasticity element made for enforcing weak stress symmetry*, Math. Comp., to appear.
- [16] B. Cockburn, J. Gopalakrishnan, and R. Lazarov, *Unified hybridization of discontinuous Galerkin, mixed and continuous Galerkin methods for second order elliptic problems*, SIAM J. Numer. Anal. **47** (2009), 1319–1365.
- [17] B. Cockburn, J. Gopalakrishnan, N.C. Nguyen, J. Peraire, and F.J. Sayas, *Analysis of an HDG method for Stokes flow*, Math. Comp., to appear.
- [18] B. Cockburn, J. Gopalakrishnan, and F.-J. Sayas, *A projection-based error analysis of HDG methods*, Math. Comp., to appear.
- [19] B. Cockburn, J. Gopalakrishnan, and H. Wang, *Locally conservative fluxes for the continuous Galerkin method*, SIAM J. Numer. Anal. **45** (2007), 1742–1776.
- [20] B. Cockburn and P.-A. Gremaud, *Error estimates for finite element methods for nonlinear conservation laws*, SIAM J. Numer. Anal. **33** (1996), 522–554.

- [21] B. Cockburn, J. Guzmán, S.-C. Soon, and H.K. Stolarski, *An analysis of the embedded discontinuous Galerkin method for second-order elliptic problems*, SIAM J. Numer. Anal. **47** (2009), 2686–2707.
- [22] B. Cockburn, J. Guzmán, and H. Wang, *Superconvergent discontinuous Galerkin methods for second-order elliptic problems*, Math. Comp. **78** (2009), 1–24.
- [23] B. Cockburn, S. Hou, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws IV: The multidimensional case*, Math. Comp. **54** (1990), 545–581.
- [24] B. Cockburn, G.E. Karniadakis, and C.-W. Shu, *The development of discontinuous Galerkin methods*, Discontinuous Galerkin Methods. Theory, Computation and Applications (Berlin) (B. Cockburn, G.E. Karniadakis, and C.-W. Shu, eds.), Lect. Notes Comput. Sci. Engrg., vol. 11, Springer Verlag, February 2000, pp. 3–50.
- [25] B. Cockburn, S.Y. Lin, and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws III: One dimensional systems*, J. Comput. Phys. **84** (1989), 90–113.
- [26] B. Cockburn and C.-W. Shu, *TVB Runge-Kutta local projection discontinuous Galerkin finite element method for scalar conservation laws II: General framework*, Math. Comp. **52** (1989), 411–435.
- [27] ———, *The Runge-Kutta local projection P^1 -discontinuous Galerkin method for scalar conservation laws*, RAIRO Modél. Math. Anal. Numér. **25** (1991), 337–361.
- [28] ———, *The Runge-Kutta discontinuous Galerkin finite element method for conservation laws V: Multidimensional systems*, J. Comput. Phys. **141** (1998), 199–224.
- [29] ———, *Runge-Kutta discontinuous Galerkin methods for convection-dominated problems*, J. Sci. Comput. **16** (2001), 173–261.
- [30] B. Cockburn and C.-W. Shu (eds.), *Special issue on discontinuous Galerkin methods*, J. Sci. Comput., vol. 22 and 23, Springer, 2005.
- [31] C. Dawson (ed.), *Special issue on discontinuous Galerkin methods*, Comput. Methods Appl. Mech. Engrg., vol. 195, Elsevier, 2006.
- [32] L. Gastaldi and R.H. Nochetto, *Sharp maximum norm error estimates for general mixed finite element approximations to second order elliptic equations*, RAIRO Modél. Math. Anal. Numér. **23** (1989), 103–128.
- [33] C. Johnson and J. Pitkäranta, *An analysis of the discontinuous Galerkin method for a scalar hyperbolic equation*, Math. Comp. **46** (1986), 1–26.
- [34] R.M. Kirby, S.J. Sherwin, and B. Cockburn, *To HDG or to CG: A comparative study*, submitted.
- [35] P. Lesaint and P. A. Raviart, *On a finite element method for solving the neutron transport equation*, Mathematical aspects of finite elements in partial differential equations (C. de Boor, ed.), Academic Press, 1974, pp. 89–145.
- [36] J.-C. Nédélec, *Mixed finite elements in \mathbf{R}^3* , Numer. Math. **35** (1980), 315–341.

- [37] N. C. Nguyen, J. Peraire, and B. Cockburn, *A hybridizable discontinuous Galerkin method for the incompressible Navier-Stokes equations (AIAA Paper 2010-362)*, Proceedings of the 48th AIAA Aerospace Sciences Meeting and Exhibit (Orlando, Florida), January 2010.
- [38] N.C. Nguyen, J. Peraire, and B. Cockburn, *A comparison of HDG methods for Stokes flow*, J. Sci. Comput., to appear.
- [39] ———, *An implicit high-order hybridizable discontinuous Galerkin method for the incompressible Navier-Stokes equations*, submitted.
- [40] ———, *Hybridizable discontinuous Galerkin methods*, Proceedings of the International Conference on Spectral and High Order Methods (Trondheim, Norway), Lect. Notes Comput. Sci. Engrg., Springer Verlag, June 2009.
- [41] ———, *An implicit high-order hybridizable discontinuous Galerkin method for linear convection-diffusion equations*, J. Comput. Phys. **228** (2009), 3232–3254.
- [42] ———, *An implicit high-order hybridizable discontinuous Galerkin method for nonlinear convection-diffusion equations*, J. Comput. Phys. **228** (2009), 8841–8855.
- [43] ———, *A hybridizable discontinuous Galerkin method for Stokes flow*, Comput. Methods Appl. Mech. Engrg. **199** (2010), 582–597.
- [44] J. Peraire, N. C. Nguyen, and B. Cockburn, *A hybridizable discontinuous Galerkin method for the compressible Euler and Navier-Stokes equations (AIAA Paper 2010-363)*, Proceedings of the 48th AIAA Aerospace Sciences Meeting and Exhibit (Orlando, Florida), January 2010.
- [45] T. Peterson, *A note on the convergence of the discontinuous Galerkin method for a scalar hyperbolic equation*, SIAM J. Numer. Anal. **28** (1991), 133–140.
- [46] P. A. Raviart and J. M. Thomas, *A mixed finite element method for second order elliptic problems*, Mathematical Aspects of Finite Element Method, Lecture Notes in Math. 606 (I. Galligani and E. Magenes, eds.), Springer-Verlag, New York, 1977, pp. 292–315.
- [47] W.H. Reed and T.R. Hill, *Triangular mesh methods for the neutron transport equation*, Tech. Report LA-UR-73-479, Los Alamos Scientific Laboratory, 1973.
- [48] G.R. Richter, *On the order of convergence of the discontinuous Galerkin method for hyperbolic equations*, Math. Comp. **77** (2008), 1871–1885.
- [49] S.-C. Soon, *Hybridizable discontinuous Galerkin methods for solid mechanics*, Ph.D. thesis, University of Minnesota, 2008.
- [50] S.-C. Soon, B. Cockburn, and H.K. Stolarski, *A hybridizable discontinuous Galerkin method for linear elasticity*, Internat. J. Numer. Methods Engrg., to appear.
- [51] R. Stenberg, *A family of mixed finite elements for the elasticity problem*, Numer. Math. **53** (1988), 513–538.
- [52] R. Stenberg, *Postprocessing schemes for some mixed finite elements*, RAIRO Modél. Math. Anal. Numér. **25** (1991), 151–167.

Numerical Analysis of Schrödinger Equations in the Highly Oscillatory Regime

Peter A. Markowich*

Abstract

Linear (and nonlinear) Schrödinger equations in the semiclassical (small dispersion) regime pose a significant challenge to numerical analysis and scientific computing, mainly due to the fact that they propagate high frequency spatial and temporal oscillations. At first we prove using Wigner measure techniques that finite difference discretisations in general require a disproportionate amount of computational resources, since underlying numerical meshes need to be fine enough to resolve all oscillations of the solution accurately, even if only accurate observables are required. This can be mitigated by using a spectral (in space) discretisation, combined with appropriate time splitting. Such discretisations are time-transverse invariant and allow for much coarser meshes than finite difference discretisations.

In many physical applications highly oscillatory periodic potentials occur in Schrödinger equations, still aggravating the oscillatory solution structure. For such problems we present a numerical method based on the Bloch decomposition of the wave function.

Mathematics Subject Classification (2010). 65M06, 65M12, 65M70, 35Q41, 35Q83

Keywords. Schrödinger equation, Wigner measure, semiclassical asymptotics, discretisation schemes, spectral methods, Bloch decomposition

Dedicated to the Memory of Frederic (Fredo) Poupaud

*Peter A. Markowich, Professor of Applied Mathematics, Department of Applied Mathematics and Theoretical Physics (DAMTP), Centre for Mathematical Sciences, University of Cambridge, Wilberforce Road, Cambridge CB3 0WA, United Kingdom and Professor of Applied Analysis, Faculty of Mathematics, University of Vienna, Nordbergstrasse 15, Room C706, A1090 Vienna, Austria. E-mail: P.A.Markowich@damtp.cam.ac.uk.

1. Introduction

We consider the (numerical) solution of the Schrödinger equation in the case of a small (scaled) Planck constant:

$$\varepsilon u_t^\varepsilon - i \frac{\varepsilon^2}{2} \Delta u^\varepsilon + iV(x)u^\varepsilon = 0, \quad x \in \mathbb{R}^m, t \in \mathbb{R} \tag{1.1a}$$

$$u^\varepsilon(x, t = 0) = u_I^\varepsilon(x), \quad x \in \mathbb{R}^m. \tag{1.1b}$$

Here V is a given electrostatic potential, $0 < \varepsilon \ll 1$ is the scaled Planck constant and $u^\varepsilon = u^\varepsilon(x, t)$ is the (generally complex-valued) wave function. By classical quantum physics [51] the wave function is an auxiliary quantity, used to compute the primary physical quantities, which are quadratic function(al)s of u^ε , e.g. the position density

$$n^\varepsilon(x, t) = |u^\varepsilon(x, t)|^2, \tag{1.2a}$$

the current density (where “ $-$ ” denotes complex conjugation)

$$J^\varepsilon(x, t) = \varepsilon \text{Im}(\bar{u}^\varepsilon(x, t) \partial_x u^\varepsilon(x, t)), \tag{1.2b}$$

and the energy density

$$e^\varepsilon(x, t) = \varepsilon^2 |\partial_x u^\varepsilon(x, t)|^2 + V(x) |u^\varepsilon(x, t)|^2. \tag{1.2c}$$

The equation (1.1a) propagates oscillations of wave length ε , which inhibit u^ε from converging strongly in, say, $L_t^\infty(L_x^2)$. Clearly, weak convergence of u^ε is not sufficient for passing to the limit in the macroscopic densities (1.2), which makes the analysis of the so-called semi-classical limit a mathematically complex issue.

Recently, much progress has been made in this area, particularly by the introduction of tools from microlocal analysis (defect measures [29], H-measures [72] and Wigner measures [30], [53], [56], [31], [54]). These techniques, which go far beyond classical WKB-methods, have shown the right way to exploit properties of the Schrödinger equation which allow the passage to the limit $\varepsilon \rightarrow 0$ in the macroscopic densities, revealing a hidden kinetic equation, whose solution is the Wigner measure associated to the sequence u^ε . Clearly, the oscillations in the wave-function have to be dealt with, too, when the Schrödinger equation with small ε is solved numerically.

For the linear Schrödinger equation classical numerical analysis methods (like the stability-consistency concept) are sufficient to derive meshing strategies for discretizations which guarantee (locally) strong convergence of the discrete wave functions to u^ε when $\varepsilon > 0$ is fixed (c.f. [77], [17], [18], [22]). Extensions to nonlinear Schrödinger equations can be found in [23], [73], [62], [65].

However, the classical strategies cannot be employed to analyse uniform properties of discretization schemes for $\varepsilon \rightarrow 0$.

At first we use microlocal techniques to analyse finite difference discretizations of linear Schrödinger equations. We choose the Crank-Nicolson time discretization scheme which is one of the most often used schemes for numerical

simulations. Spatial discretizations are general arbitrary-order symmetric finite difference schemes. We identify the semiclassical Wigner measure (on the scale ε) for all combinations of ε and of the time and space mesh sizes. We have uniform convergence for the average values of all (regular) observables in exactly those cases, for which the Wigner measure of the numerical scheme is identical to the Wigner measure of the Schrödinger equation itself. Thus, from this theory we obtain **sharp** (i.e. necessary and sufficient) conditions on the mesh sizes which guarantee good approximation quality of all observables uniformly as $\varepsilon \rightarrow 0^+$. For the Crank-Nicolson scheme we prove that spatial and temporal oscillations have to be asymptotically resolved in order to obtain accurate numerically computed observables. From this analysis (which can be generalized to other time-discretizations) it follows that finite difference methods have a very poor convergence behaviour for small values of ε .

This clearly shows the big *risk* in using FD-methods for Schrödinger calculations in the semiclassical regime. Even stable schemes may produce completely wrong observables under seemingly reasonable meshing strategies (i.e. asymptotic resolution of the oscillation is not always enough). Worse enough, in these cases there is no warning from the scheme (like blow-up) that something went wrong (since local error control cannot be used anyway). The only safety anchor here lies in analysis and in physical insight.

In the second part of the paper we consider time splitting-trigonometric spectral schemes which have much better asymptotic properties as $\varepsilon \rightarrow 0$. For analytical results on time-splitting spectral methods for linear and nonlinear Schrödinger equation (not in the semiclassical regime, though) we refer to [16], [28], [24]. The third part of the paper is concerned with an extension of the spectral-time splitting scheme to Schrödinger equations with periodic highly oscillatory potentials, typically occurring in solid state physics.

We emphasize that the first part of this paper is based on work reported in [58], the second on [5] and the third on [40].

2. Schrödinger-type Equations, Observables and Wigner Transforms

We consider the following scalar IVP (generalized linear Schrödinger equation)

$$\varepsilon u_t^\varepsilon + iQ(x, \varepsilon D)^W u^\varepsilon = 0, \quad x \in \mathbb{R}^m, t \in \mathbb{R} \quad (2.1a)$$

$$u^\varepsilon(x, t = 0) = u_I^\varepsilon(x), \quad x \in \mathbb{R}^m. \quad (2.1b)$$

Here $\varepsilon \in (0, \varepsilon_0]$, $\varepsilon_0 > 0$, is a small parameter (e.g. the scaled Planck-constant), and $Q(\cdot, \varepsilon D)^W$ is the Weyl-operator associated to the symbol $Q(x, \varepsilon \xi)$:

$$Q(x, \varepsilon D)^W \varphi(x) := \frac{1}{(2\pi)^m} \int_{\mathbb{R}_y^m} \int_{\mathbb{R}_\xi^m} Q\left(\frac{x+y}{2}, \varepsilon \xi\right) \varphi(y) e^{i(x-y)\cdot\xi} d\xi dy. \quad (2.2)$$

For the following we assume that the symbol $Q = Q(x, \xi)$ is polynomial in ξ with C^∞ -coefficients:

$$Q(x, \xi) = \sum_{|k| \leq K} Q_k(x) \xi^k \quad , \quad (2.3)$$

where $k = (k_1, \dots, k_m) \in \mathbb{N}_0^m$ denotes a multi-index, K is the order of the differential operator (2.2) and $|k| := k_1 + \dots + k_m$ the order of the multi-index k . The DO (2.2) can now be written as

$$Q(x, \varepsilon D)^W \varphi(x) = \sum_{|k| \leq K} \varepsilon^{|k|} D_y^k \left(Q_k \left(\frac{x+y}{2} \right) \varphi(y) \right) \Big|_{y=x} . \quad (2.4)$$

We denoted $D_y = -i\partial_y$.

The convenience in the Weyl-calculus lies in the fact that an essentially selfadjoint Weyl-operator has a realvalued symbol (cf. [38]).

Being interested in generalizations of the Schrödinger-equation we assume for the following

(A1) (i) $Q(x, \varepsilon D)^W$ is essentially selfadjoint on $L^2(\mathbb{R}^m)$

and, in order not to complicate the analysis unnecessarily

(A1) (ii) $\forall k, \alpha \in \mathbb{N}_0^m$ with $|k| \leq K \exists C_{k,\alpha} > 0 : |\partial_x^\alpha Q_k(x)| \leq C_{k,\alpha} \forall x \in \mathbb{R}^m$.

This implies in particular

$$Q_k \text{ is real valued for } 0 \leq |k| \leq K. \quad (2.5)$$

By Stone's Theorem $\exp(-i \frac{t}{\varepsilon} Q(\cdot, \varepsilon D)^W)$ is a strongly continuous group of unitary operators on $L^2(\mathbb{R}^m)$. Thus we conclude the $L^1(\mathbb{R}^m)$ -conservation in time of the position-density (1.2a):

$$n^\varepsilon(x, t) := |u^\varepsilon(x, t)|^2, \quad (2.6)$$

i.e. we have

$$\int_{\mathbb{R}^m} n^\varepsilon(x, t) dx = \int_{\mathbb{R}^m} n_I^\varepsilon(x) dx \quad \forall t \in \mathbb{R}, \quad (2.7)$$

where we set $n_I^\varepsilon := |u_I^\varepsilon|^2$.

In quantum mechanics the wave function $u^\varepsilon = u^\varepsilon(x, t)$ (i.e. the solution of the Schrödinger-equation) is usually considered an auxiliary quantity. It facilitates the calculation of physical observables of the system under consideration [51] corresponding to actual measurements. An observable A^ε , which depends on the position variable x and on the momentum operator εD , is given by the Weyl-operator

$$A^\varepsilon = a(\cdot, \varepsilon D)^W \quad (2.8)$$

with the realvalued symbol $a(x, \varepsilon\xi)$. Of particular physical interest is the average value of the observable A^ε in the state $u^\varepsilon(t)$ (i.e. the mean value of the performed measurement) given by:

$$E_a^\varepsilon(t) := \left(a(\cdot, \varepsilon D)^W u^\varepsilon(t), u^\varepsilon(t) \right). \tag{2.9}$$

Here (\cdot, \cdot) stands for the $L^2(\mathbb{R}^m)$ -scalar product and, of course, it is assumed that $u^\varepsilon(t)$ lies in the domain of $a(\cdot, \varepsilon D)^W$.

A good framework for manipulating quantities which are quadratic in the wave function (e.g. (2.9)), is given by the Wigner-transform [31, 75]. For given $f, g \in \mathcal{S}'(\mathbb{R}^m)$ and a given scale $\varepsilon \in (0, \varepsilon_0]$ we define the Wigner-transform (on the scale ε) by

$$w^\varepsilon(f, g)(x, \xi) = \frac{1}{(2\pi)^m} \int_{\mathbb{R}^m} f\left(x - \varepsilon \frac{v}{2}\right) \overline{g\left(x + \varepsilon \frac{v}{2}\right)} e^{iv \cdot \xi} dv. \tag{2.10}$$

For fixed ε this defines a bilinear continuous mapping from $\mathcal{S}'(\mathbb{R}^m) \times \mathcal{S}'(\mathbb{R}^m)$ into $\mathcal{S}'(\mathbb{R}_x^m \times \mathbb{R}_\xi^m)$. Also we have

$$w^\varepsilon(f, g) = \overline{w^\varepsilon(g, f)} \tag{2.11a}$$

and (by a simple calculation)

$$\langle w^\varepsilon(f, g), a \rangle = \langle \bar{g}, a(\cdot, \varepsilon D)^W f \rangle. \tag{2.11b}$$

Here we assume $a \in \mathcal{S}(\mathbb{R}_x^m \times \mathbb{R}_\xi^m)$ and denote by $\langle \cdot, \cdot \rangle$ the duality bracket between \mathcal{S}' and \mathcal{S} (linear in both arguments).

Obviously, (2.11b) implies

$$E_a^\varepsilon(t) = \int_{\mathbb{R}_x^m \times \mathbb{R}_\xi^m} w^\varepsilon[u^\varepsilon(t)](x, \xi) a(x, \xi) dx d\xi \tag{2.12}$$

where we denoted $w^\varepsilon[f] = w^\varepsilon(f, f)$.

For the following we denote the Fourier-transform by

$$\hat{f}(\xi) := (F_{x \rightarrow \xi} f)(\xi) = \int_{\mathbb{R}^m} f(x) e^{-ix \cdot \xi} dx. \tag{2.13}$$

The following proposition holds [31]:

Proposition 2.1. *Let f, g lie in a bounded subset of $L^2(\mathbb{R}_x^m)$. Then $w^\varepsilon(f, g)$ is bounded uniformly in $\mathcal{S}'(\mathbb{R}_x^m \times \mathbb{R}_\xi^m)$ as $\varepsilon \rightarrow 0$.*

Now let $f^\varepsilon \in L^2(\mathbb{R}^m)$ uniformly as $\varepsilon \rightarrow 0$. Then, by compactness, there exists a subsequence ε_k and $w^0 \in \mathcal{S}'$ such that

$$w^{\varepsilon_k}[f^{\varepsilon_k}] \xrightarrow{k \rightarrow \infty} w^0 \quad \text{in } \mathcal{S}'. \tag{2.14}$$

It is well known that w^0 is non-negative, i.e. it is a positive Borel-measure [31]. In the sequel we write $w^0 = w^0[f^{\varepsilon_k}]$ and call it the Wigner-measure of f^{ε_k} (for the scale ε_k).

Also we shall use (see [31] for the proof)

Proposition 2.2. *Let $q \in C^\infty(\mathbb{R}^m_x \times \mathbb{R}^m_\xi)$ satisfy for some $M \geq 0, C_\alpha \geq 0$:*

$$|\partial_{x,\xi}^\alpha q(x, \xi)| \leq C_\alpha(1 + |\xi|)^M \quad \forall \alpha \in \mathbb{N}_0^m \times \mathbb{N}_0^m. \tag{2.15}$$

Then, if f, g lie in a bounded subset of $L^2(\mathbb{R}^m)$, the expansion

$$w^\varepsilon \left(q(\cdot, \varepsilon D)^W f, g \right) = q w^\varepsilon(f, g) + \frac{\varepsilon}{2i} \{q, w^\varepsilon(f, g)\} + O(\varepsilon^2) \tag{2.16}$$

holds in $\mathcal{S}'(\mathbb{R}^m_x \times \mathbb{R}^m_\xi)$ uniformly for all symbols $q = q(x, \xi)$ which satisfy (2.15).

Here $\{\cdot, \cdot\}$ denotes the Poisson bracket:

$$\{f, g\} = \partial_\xi f \cdot \partial_x g - \partial_x f \cdot \partial_\xi g.$$

Now let the initial datum u_I^ε of (2.1) satisfy

$$(A2) \quad u_I^\varepsilon \in L^2(\mathbb{R}^m) \text{ uniformly as } \varepsilon \rightarrow 0.$$

Then (2.7) and Proposition 2.1 imply the uniform boundedness of $w^\varepsilon := w^\varepsilon[u^\varepsilon(t)]$ in $L^\infty(\mathbb{R}_t; \mathcal{S}'(\mathbb{R}^m_x \times \mathbb{R}^m_\xi))$ and the existence of $w^0 \in L^\infty(\mathbb{R}_t; \mathcal{M}^+(\mathbb{R}^m_x \times \mathbb{R}^m_\xi))$ (where \mathcal{M}^+ stands for the cone of positive Borel-measures) such that after selection of a subsequence

$$w^\varepsilon \xrightarrow{\varepsilon \rightarrow 0} w^0 \quad \text{in } L^\infty(\mathbb{R}_t; \mathcal{M}^+(\mathbb{R}^m_x \times \mathbb{R}^m_\xi)) \text{ weak-}^*. \tag{2.17}$$

To derive an equation for w^0 we differentiate

$$\frac{\partial}{\partial t} w^\varepsilon = w^\varepsilon(u_t^\varepsilon, u^\varepsilon) + w^\varepsilon(u^\varepsilon, u_t^\varepsilon) = 2 \operatorname{Re} w^\varepsilon(u_t^\varepsilon, u^\varepsilon),$$

use (2.1a) and Proposition 2.2

$$\frac{\partial}{\partial t} w^\varepsilon = \frac{2}{\varepsilon} \operatorname{Im} w^\varepsilon \left(Q(\cdot, \varepsilon D)^W u^\varepsilon, u^\varepsilon \right) = -\{Q, w^\varepsilon\} + O(\varepsilon)$$

(w^ε and Q are realvalued!). Passing to the limit $\varepsilon \rightarrow 0$ gives the transport equation

$$\frac{\partial}{\partial t} w^0 + \{Q, w^0\} = 0, \quad (x, \xi) \in \mathbb{R}^m_x \times \mathbb{R}^m_\xi, t \in \mathbb{R} \tag{2.18a}$$

subject to the initial condition

$$w^0(t = 0) = w_I^0 := w^0[u_I^\varepsilon]. \tag{2.18b}$$

Results along these lines for even more general IVP's can be found in [31].

The unique solution of (2.18) allows the calculation of the limit $\varepsilon \rightarrow 0$ of the average value of an observable A^ε determined by a symbol $a = a(x, \xi) \in \mathcal{S}$ in the state u^ε (cf. (2.9) and (2.12)). We obtain

$$E_a^\varepsilon \rightarrow E_a^0 := \int a(x, \xi) w^0(dx, d\xi, t) \tag{2.19}$$

after selection of a subsequence. Since the limit process (2.17) is actually locally uniform in t [31], the convergence (2.19) takes place in $C_{loc}(\mathbb{R}_t)$.

In order to avoid having to take subsequences we shall assume for the following

$$(A3) \quad w_I^0 \text{ is the unique Wigner-measure of } u_I^\varepsilon.$$

3. Finite Difference Schemes

Let

$$\Gamma = \left\{ \mu = l_1 a_1 + \dots + l_m a_m \mid l_j \in \mathbb{Z} \text{ for } 1 \leq j \leq m \right\} \subseteq \mathbb{R}^m \tag{3.1}$$

be the lattice generated by the linearly independent vectors $a_1, \dots, a_m \in \mathbb{R}^m$. For a multi-index $k \in \mathbb{N}_0^m$ we construct a discretization of the order N of the operator ∂_x^k as follows:

$$\partial_x^k \varphi(x) \sim \frac{1}{h^{|k|}} \sum_{\mu \in \Gamma_k} a_{\mu,k} \varphi(x + h\mu). \tag{3.2}$$

Here $h \in (0, h_0]$ is the mesh-size, $\Gamma_k \subseteq \Gamma$ is the finite set of discretization points and $a_{\mu,k} \in \mathbb{R}$ are coefficients satisfying

$$(D1) \quad \sum_{\mu \in \Gamma_k} a_{\mu,k} \mu^l = \delta_{l,k} k^l, \quad 0 \leq |l| \leq N + |k| - 1$$

where $\delta_{l,k} = 1$ if $l = k$ and 0 otherwise. It is an easy exercise to show that the local discretization error of (3.2) is $O(h^N)$ for all smooth functions if (D1) holds. For a detailed discussion of the linear problem (D1) (i.e. possible choices of the coefficients $a_{\mu,k}$) we refer to [55].

We now define the corresponding finite difference discretization of $Q(\cdot, \varepsilon D)^W$ by applying (3.2) (with $\partial = iD$) directly to (2.4). Denoting

$$Q_{h,\varepsilon}(x, \xi) = \sum_{|k| \leq K} \varrho^{|k|} (-i)^{|k|} \sum_{\mu \in \Gamma_k} a_{\mu,k} e^{i \frac{\mu}{\varepsilon} \cdot \xi} Q_k(x) \tag{3.3}$$

with

$$\varrho = \frac{\varepsilon}{h} \tag{3.4}$$

we obtain the finite difference discretization of (2.4) in the form

$$Q_{h,\varepsilon}(x, \varepsilon D)^W \varphi(x) = \sum_{|k| \leq K} \varrho^{|k|} (-i)^{|k|} \sum_{\mu \in \Gamma_k} a_{\mu,k} Q_k \left(x + \frac{h\mu}{2} \right) \varphi(x + h\mu). \tag{3.5}$$

Since $Q_{h,\varepsilon}(x, \varepsilon D)^W$ is a bounded operator on $L^2(\mathbb{R}^m)$, it is selfadjoint if

$$(D2) \quad i^{|k|} \sum_{\mu \in \Gamma_k} a_{\mu,k} e^{i\mu \cdot \xi} \text{ is realvalued for } 0 \leq |k| \leq K.$$

As temporal discretizations we consider the Crank-Nicolson scheme with time step $\Delta t > 0$:

$$\varepsilon \frac{u_{n+1}^\sigma - u_n^\sigma}{\Delta t} + iQ_{h,\varepsilon}(x, \varepsilon D)^W \left(\frac{1}{2}u_{n+1}^\sigma + \frac{1}{2}u_n^\sigma \right) = 0, \quad n = 0, 1, 2, \dots \quad (3.6a)$$

$$u_0^\sigma = u_I^\varepsilon. \quad (3.6b)$$

Here (and in the sequel) we denote the vector of small parameters by $\sigma = (\varepsilon, h, \Delta t)$.

Note that the selfadjointness of $Q_{h,\varepsilon}(x, \varepsilon D)^W$ implies that the operator $Id + i\omega Q_{h,\varepsilon}(x, \varepsilon D)^W$ is boundedly invertible on $L^2(\mathbb{R}_x^m)$ for all $\omega \in \mathbb{R}$. Therefore the scheme (3.6) gives well-defined approximations u_n^σ for $n = 1, 2, \dots$ if $u_I^\varepsilon \in L^2(\mathbb{R}_x^m)$. Moreover we remark that it is sufficient to evaluate (3.6) at $x \in h\Gamma$ in order to obtain discrete equations for $\{u_n^\sigma(h\mu) | \mu \in \Gamma\}$. Clearly, artificial ‘far out’ boundary conditions have to be imposed for practical computations. Their impact will not be taken into account in the subsequent analysis.

We now collect properties of the finite difference schemes. We start with the spatial discretizations:

Lemma 3.1. *Let (A1), (D1), (D2) hold.*

(i) *Assume that there is $C > 0$ such that $\varrho = \frac{\varepsilon}{h} \geq C$. Then for every $\delta \in \mathbb{N}_0^m \times \mathbb{N}_0^m$ there is $C_\delta > 0$ independent of ε, h such that*

$$|\partial_{x,\xi}^\delta Q_{h,\varepsilon}(x, \xi)| \leq C_\delta (1 + |\xi|^{N+K}). \quad (3.7)$$

(ii) *Assume that $\varphi \in \mathcal{S}(\mathbb{R}_x^m \times \mathbb{R}_\xi^m)$. For $\varrho = \frac{\varepsilon}{h} \xrightarrow{\varepsilon, h \rightarrow 0} \infty$ we have*

$$Q_{h,\varepsilon} \varphi \xrightarrow{\varepsilon, h \rightarrow 0} Q \varphi \quad \text{in } \mathcal{S}(\mathbb{R}_x^m \times \mathbb{R}_\xi^m). \quad (3.8)$$

For the proof we refer to [58].

Choosing h such that $\varrho = \frac{\varepsilon}{h} \rightarrow \infty$ corresponds to asymptotically resolving the oscillations of wave-length $O(\varepsilon)$ of the solution $u^\varepsilon(t)$ of (2.1). In the case $\varrho = \text{const.}$ (i.e. ‘placing a fixed number of gridpoints per oscillation’) the symbol $Q_{h,\varepsilon}(x, \xi)$ is independent of h and ε :

$$Q_{h,\varepsilon}(x, \xi) \equiv Q_\varrho(x, \xi) := \sum_{|k| \leq K} \varrho^{|k|} \sum_{\mu \in \Gamma_k} a_{\mu,k} (-i)^{|k|} e^{i\frac{\mu}{\varrho} \cdot \xi} Q_k(x). \quad (3.9)$$

In the case $\varrho \xrightarrow{\varepsilon, h \rightarrow 0} 0$ (which corresponds to ‘ignoring’ the oscillations) we have

$$Q_{h,\varepsilon} \underset{h, \varepsilon \rightarrow 0}{\sim} \sum_{\mu \in \Gamma_0} a_{\mu,0} \cos\left(\frac{\mu \cdot \xi}{\varrho}\right) Q_0(x)$$

and, thus, $Q_{h,\varepsilon}(x, \varepsilon D)^W$ does not approximate $Q(x, \varepsilon D)^W$. Therefore, we cannot expect reasonable numerical results in this case (which will not be investigated further).

The next Lemma concerns the temporal stability of the discretization. Here and in the sequel we use the notation $\|\cdot\|$ for the norm in $L^2(\mathbb{R}_x^m)$:

Lemma 3.2. *We have for the solution of (3.6):*

$$\|u_n^\sigma\| = \|u_l^\varepsilon\|; \quad n = 1, 2, 3, \dots \tag{3.10}$$

The proof follows from taking the L^2 -scalar product of (3.6a) with

$$\frac{1}{2}u_{n+1}^\sigma + \frac{1}{2}u_n^\sigma.$$

A comment on the time-transverse non-invariance of the discretization schemes is in order. It is clear that the average values of the observables defined in (2.12) are invariant under the substitution $v^\varepsilon(t) = u^\varepsilon(t)e^{i\frac{\omega}{\varepsilon}t}$ for $\omega \in \mathbb{R}$, i.e. the average value of the observable in the state $u^\varepsilon(t)$ is equal to its average value in the state $v^\varepsilon(t)$. Also, the Wigner-function is invariant under this substitution:

$$w^\varepsilon[u^\varepsilon(t)] = w^\varepsilon[v^\varepsilon(t)e^{i\frac{\omega}{\varepsilon}t}] \quad \forall \omega \in \mathbb{R}.$$

The PDO (2.1a) transforms to

$$\varepsilon v_t^\varepsilon + i\left(Q(x, \varepsilon D)^W + \omega Id\right)v^\varepsilon = 0,$$

which implies that the zeroth order term $Q_0(x)$ (physically a potential) is replaced by $Q_0(x) + \omega$ while the other coefficients $Q_k(x), k \neq 0$, remain unchanged. The situation is completely different for the difference schemes.

A simple calculation shows that the discrete gauge transformation $v_n^\sigma = u_n^\sigma e^{i\frac{\omega}{\varepsilon}t_n}$ does not ‘commute (modulo adding a real constant to the potential) with the discretizations’ (3.6). Thus, the discrete approximations of average values of observables depend on the gauging of the potential.

The consistency-stability concept of classical numerical analysis provides a framework for the convergence analysis of finite difference discretizations of linear partial differential equations. Thus, for $\varepsilon > 0$ fixed it is easy to prove that the scheme (3.6) is convergent of order N in space and order 2 in time if the solution u^ε is sufficiently smooth. Therefore, again for fixed $\varepsilon > 0$ we conclude convergence of the same order for averages of the observables defined in (2.9) assuming that a is smooth. Due to the oscillatory nature of the solutions of (2.1) the local discretization error of the finite difference schemes and, consequently, also the global discretization error, generally tend to infinity as ε tends to 0. Thus, the classical theory does not provide uniform results as $\varepsilon \rightarrow 0$. Indeed, under the reasonable assumption

$$\frac{\partial^{|j_1|+|j_2|}}{\partial x^{j_1} \partial t^{j_2}} u^\varepsilon \sim \varepsilon^{-|j_1|-|j_2|} \text{ in } L^2(\mathbb{R}^m)$$

locally uniformly in t for all multi-indices j_1 and $j_2 \in \mathbb{N} \cup \{0\}$, (which is satisfied for ε - oscillatory initial data) the classical stability-consistency analysis gives for the global L^2 -discretization error

$$O\left(\frac{(\Delta t)^2}{\varepsilon^3}\right) + O\left(\frac{h^N}{\varepsilon^{N+1}}\right). \tag{3.11}$$

The situation is further complicated by the fact that the solution u^ε of (2.1) and their discrete approximations u_n^σ , which solve (3.6), generally only converge weakly in $L^2(\mathbb{R}^m)$ as $\varepsilon \rightarrow 0$ and, resp., $\sigma \rightarrow 0$. The limit processes $\varepsilon \rightarrow 0, \sigma \rightarrow 0$ do not commute with the quadratically nonlinear operation which has to be carried out to compute the average values of observables.

In practice one is interested in finding conditions on the mesh sizes h and Δt , in dependence of ε and the used discretization such that the average values of the observables in the discrete state converge uniformly as $\varepsilon \rightarrow 0$ to E_a^ε given by (2.9).

Let us set for $n \in \mathbb{N}, t_n = nk$:

$$E_a^\sigma(t_n) := \left(a(\cdot, \varepsilon D)^W u_n^\sigma, u_n^\sigma\right). \tag{3.12}$$

The function $E_a^\sigma(t), t \in \mathbb{R}^+$, then is defined by piecewise linear interpolation of the values $E_a^\sigma(t_n)$. As discussed above, we want to find conditions on h, k such that for all $a \in \mathcal{S}(\mathbb{R}_x^m \times \mathbb{R}_\xi^m)$:

$$E_a^\sigma \xrightarrow{k, h \rightarrow 0} E_a^\varepsilon \quad \text{uniformly in } \varepsilon \in (0, \varepsilon_0] \tag{3.13}$$

and locally uniformly in t . Denoting

$$w^\sigma(t_n) := w^\varepsilon[u_n^\sigma] \tag{3.14}$$

and again defining $w^\sigma(t), t \in \mathbb{R}$, by piecewise linear interpolation of the values (3.14), we conclude from (2.12) that (3.13) is equivalent to proving

$$w^\sigma \xrightarrow{h, k \rightarrow 0} w^\varepsilon \quad \text{in } \mathcal{S}'(\mathbb{R}_x^m \times \mathbb{R}_\xi^m) \text{ uniformly in } \varepsilon \in (0, \varepsilon_0], \tag{3.15}$$

locally uniformly in t , where $w^\varepsilon = w^\varepsilon[u^\varepsilon]$ is the Wigner-transform of the solution u^ε of (2.1). Note that $w^\sigma(t_n)$ denotes the Wigner-transform of the finite difference solution u_n^σ **on the scale** ε .

In this Section we shall compute the accumulation points of w^σ as $\sigma \rightarrow 0$. We shall see that the set of Wigner-measures of the difference schemes

$$A := \left\{ W^0 \mid \exists \text{ subsequence } (\sigma_l) \text{ of } (\sigma) : W^0 = \lim_{l \rightarrow \infty} w^{\sigma_l} \right\} \tag{3.16}$$

depends decisively on the relative sizes of ε, h and Δt . In those cases, in which $W^0 = w^0 (= \lim_{\varepsilon \rightarrow 0} w^\varepsilon)$ holds, (3.15) follows, while (3.15) does not hold if the measures W^0 and w^0 are different.

Theorem 3.1. *Assume that (A1),(A2),(A3), (D1),(D2) hold. Then the following cases occur for the unique Wigner-measure $W^0 \in A$ of (3.6):*

Case (I) $\frac{h}{\varepsilon} \rightarrow 0$ ($\varrho \rightarrow \infty$).

(i) $\frac{\Delta t}{\varepsilon} \rightarrow 0$. Then W^0 satisfies:

$$\begin{aligned} \frac{\partial}{\partial t} W^0 + \{Q, W^0\} &= 0 \\ W^0(t=0) &= w_I^0 \end{aligned}$$

(ii) $\frac{\Delta t}{\varepsilon} \rightarrow \omega \in \mathbb{R}^+$. W^0 solves the IVP:

$$\begin{aligned} \frac{\partial}{\partial t} W^0 + \left\{ \frac{2}{\omega} \arctan \left(\frac{\omega}{2} Q \right), W^0 \right\} &= 0 \\ W^0(t=0) &= w_I^0 \end{aligned}$$

(iii) $\frac{\Delta t}{\varepsilon} \rightarrow \infty$. If there exists $D > 0$ such that:

$$|Q(x, \xi)| \geq D \quad \forall x, \xi \in \mathbb{R}^m$$

then W^0 is constant in time

$$W^0(x, \xi, t) \equiv W_I^0(x, \xi).$$

Case (II) $\frac{h}{\varepsilon} \rightarrow \frac{1}{\varrho} \in \mathbb{R}^+$. Then the assertions (i), (ii), (iii) hold true, if Q is replaced by Q_ϱ .

We recall that Q_ϱ is defined in (3.9).

The proof of the theorem proceeds similarly to the derivation of (2.18) from the IVP for the Schrödinger equation (see [57] for details). Note that Theorem 3.1 implies that asymptotically correct observables (as $\varepsilon \rightarrow 0$) can only be computed from the Crank-Nicolson scheme if both spacial and temporal oscillations of wave-length ε are accurately resolved by the grid. Time-irreversible finite difference schemes behave much worse (e.g. the explicit or implicit Euler scheme), they require $\Delta t = o(\varepsilon^2)$ in order to guarantee asymptotically correct numerically computed observables.

4. Time-splitting Spectral Approximations

In this section we present time-splitting trigonometric spectral approximations of the problem (1.1) with periodic boundary conditions. For the simplicity of

notation we shall introduce the method for the case of one space dimension ($m = 1$). Generalizations to $m > 1$ are straightforward for tensor product grids and the results remain valid without modifications. For $m = 1$, the problem becomes

$$\varepsilon \partial_t u^\varepsilon - i \frac{\varepsilon^2}{2} \partial_{xx} u^\varepsilon + iV(x)u^\varepsilon = 0, \quad a < x < b, \quad t > 0, \quad (4.1)$$

$$\begin{aligned} u^\varepsilon(x, t = 0) &= u_I^\varepsilon(x), \quad a \leq x \leq b, \quad u^\varepsilon(a, t) \\ &= u^\varepsilon(b, t), \partial_x u^\varepsilon(a, t) = \partial_x u^\varepsilon(b, t), \quad t > 0. \end{aligned} \quad (4.2)$$

We choose the spatial mesh size $h = \Delta x > 0$ with $h = (b - a)/M$ for M an even positive integer, the time step $k = \Delta t > 0$ and let the grid points and the time step be

$$x_j := a + j h, \quad t_n := n k, \quad j = 0, 1, \dots, M, \quad n = 0, 1, 2, \dots$$

Let $U_j^{\varepsilon, n}$ be the numerical approximation of $u^\varepsilon(x_j, t_n)$ and $\mathbf{u}^{\varepsilon, n}$ be the vector with components $U_j^{\varepsilon, n}$.

The first-order time-splitting spectral method (SP1). From time $t = t_n$ to time $t = t_{n+1}$, the Schrödinger equation (4.1) is solved in two steps. One solves

$$\varepsilon \partial_t u^\varepsilon - i \frac{\varepsilon^2}{2} \partial_{xx} u^\varepsilon = 0, \quad (4.3)$$

for one time step, followed by solving

$$\varepsilon \partial_t u^\varepsilon + iV(x)u^\varepsilon = 0, \quad (4.4)$$

again for the same time step. The solution of (4.3) of $t = t_{n+1}$ is taken as initial value for (4.4) at $t = t_n$. Equation (4.3) will be discretized in space by the spectral method and integrated in time *exactly*. The ODE (4.4) will then be solved exactly. The detailed method is given by:

$$\begin{aligned} U_j^{\varepsilon, *}&= \frac{1}{M} \sum_{l=-M/2}^{M/2-1} e^{-i\varepsilon k \mu_l^2/2} \widehat{U}_l^{\varepsilon, n} e^{i\mu_l(x_j-a)}, \quad j = 0, 1, 2, \dots, M-1, \\ U_j^{\varepsilon, n+1}&= e^{-iV(x_j)k/\varepsilon} U_j^{\varepsilon, *}; \end{aligned} \quad (4.5)$$

where $\widehat{U}_l^{\varepsilon, n}$, the Fourier coefficients of $U^{\varepsilon, n}$, are defined as

$$\mu_l = \frac{2\pi l}{b-a}, \quad \widehat{U}_l^{\varepsilon, n} = \sum_{j=0}^{M-1} U_j^{\varepsilon, n} e^{-i\mu_l(x_j-a)}, \quad l = -\frac{M}{2}, \dots, \frac{M}{2} - 1, \quad (4.6)$$

with

$$U_j^{\varepsilon, 0} = u^\varepsilon(x_j, 0) = u_I^\varepsilon(x_j), \quad j = 0, 1, 2, \dots, M. \quad (4.7)$$

Note that the only time discretization error of this method is the splitting error, which is first order in k for any fixed $\varepsilon > 0$.

The Strang splitting spectral method (SP2). From time $t = t_n$ to time $t = t_{n+1}$, we split the Schrödinger equation (4.1) via the well-known Strang splitting:

$$\begin{aligned}
 U_j^{\varepsilon,*} &= e^{-iV(x_j)k/2\varepsilon} U_j^{\varepsilon,n}, & j = 0, 1, 2, \dots, M-1, \\
 U_j^{\varepsilon,**} &= \frac{1}{M} \sum_{l=-M/2}^{M/2-1} e^{-i\varepsilon k \mu_l^2/2} \widehat{U}_l^{\varepsilon,*} e^{i\mu_l(x_j-a)}, & j = 0, 1, 2, \dots, M-1, \\
 U_j^{\varepsilon,n+1} &= e^{-iV(x_j)k/2\varepsilon} U_j^{\varepsilon,**}, & j = 0, 1, 2, \dots, M-1,
 \end{aligned}
 \tag{4.8}$$

where $\widehat{U}_l^{\varepsilon,*}$, the Fourier coefficients of $U^{\varepsilon,*}$, are defined as

$$\widehat{U}_l^{\varepsilon,*} = \sum_{j=0}^{M-1} U_j^{\varepsilon,*} e^{-i\mu_l(x_j-a)}, \quad l = -\frac{M}{2}, \dots, \frac{M}{2} - 1.
 \tag{4.9}$$

Again, the overall time discretization error comes solely from the splitting, which is now second order in k for fixed $\varepsilon > 0$. Note that the main advantage of (SP1) and (SP2) over FD-methods is their gauge invariance with respect to adding constants to the potential V .

Let $\mathbf{u} = (U_0, \dots, U_{M-1})^T$ and $\|\cdot\|_{l^2}$ the usual discrete l^2 -norm on the interval (a, b) , i.e.

$$\|\mathbf{u}\|_{l^2} = \sqrt{\frac{b-a}{M} \sum_{j=0}^{M-1} |U_j|^2}.
 \tag{4.10}$$

For the *stability* of the time-splitting spectral approximations (SP1) and (SP2), with variable potential $V = V(x)$, we prove the following lemma, which shows that the total charge is conserved.

Lemma 4.1. *The time-splitting spectral schemes (SP1) (4.5) and (SP2) (4.8) are unconditionally stable. In fact, under any mesh size h and time step k ,*

$$\|\mathbf{u}^{\varepsilon,n}\|_{l^2} = \|u_I^\varepsilon\|_{l^2}, \quad n = 1, 2, \dots,
 \tag{4.11}$$

and consequently:

$$\|u_{\text{Int}}^{\varepsilon,n}\| = \|u_{\text{Int}}^{\varepsilon,0}\|, \quad n = 1, 2, \dots
 \tag{4.12}$$

Here $u_{\text{Int}}^{\varepsilon,n}$ stands for the trigonometric polynomial interpolating $\{(x_0, u_0^{\varepsilon,n}), (x_1, u_1^{\varepsilon,n}), \dots, (x_M, u_M^{\varepsilon,n})\}$.

For the proof we refer to [5].

We now establish error estimates for (SP1).

We assume that the solution $u^\varepsilon = u^\varepsilon(x, t)$ of (4.1), (4.2) and the potential $V = V(x)$ in (4.1) are $(b - a)$ periodic and $C^\infty(\mathbb{R})$. Moreover, we assume that there are positive constants $C_m > 0$, $D_m > 0$, independent of ε , x , t , such that

$$(B) \quad \left\| \frac{\partial^{m_1+m_2}}{\partial x^{m_1} \partial t^{m_2}} u^\varepsilon \right\|_{C([0,T];L^2(a,b))} \leq \frac{C_{m_1+m_2}}{\varepsilon^{m_1+m_2}}, \quad \left\| \frac{d^m}{dx^m} V \right\|_{L^\infty(a,b)} \leq D_m, \quad (4.13)$$

for all $m, m_1, m_2 \in \mathbb{N} \cup \{0\}$.

Thus, we assume that the solution oscillates in space and time with wavelength ε .

Theorem 4.1. *Let $u^\varepsilon = u^\varepsilon(x, t)$ be the solution of (4.1), (4.2) and $\mathbf{u}^{\varepsilon,n}$ be the discrete approximation (SP1) given by (4.5). Under assumption (B), and assuming $\frac{k}{\varepsilon} = O(1)$, $\frac{h}{\varepsilon} = O(1)$, we have for all positive integers $m \geq 1$ and $t_n \in [0, T]$:*

$$\|u^\varepsilon(t_n) - u_{\text{Int}}^{\varepsilon,n}\| \leq G_m \frac{T}{k} \left(\frac{h}{\varepsilon(b-a)} \right)^m + \frac{CTk}{\varepsilon}, \quad (4.14)$$

where C is a positive constant independent of ε , h , k and m and G_m is independent of ε , h , k .

See [5] for the proof. A similar result can be established for (SP2).

Now, let $\delta > 0$ be the desired error bound such that

$$\|u^\varepsilon(t_n) - u_{\text{Int}}^{\varepsilon,n}\| \leq \delta \quad (4.15)$$

shall hold. Then the meshing strategy (on $O(1)$ -time and space intervals)

$$(a) \quad \frac{k}{\varepsilon} = O(\delta), \quad (b) \quad \frac{h}{\varepsilon} = O\left(\delta^{1/m} k^{1/m}\right) \quad (4.16)$$

is suggested by the Theorem, where $m \geq 1$ is an arbitrary integer, assuming that G_m does not increase too fast as $m \rightarrow \infty$.

This meshing, although more efficient than what is needed for finite differences, is even too restrictive for both (SP1) and (SP2) if only accurate quadratic observables are desired, cf. below.

Now let $u^\varepsilon(t)$ be the solution of the IVP (4.1), (4.2) and denote its Wigner transform by w^ε . Then w^ε satisfies the Wigner equation

$$w_t^\varepsilon + \xi \cdot \partial_x w^\varepsilon + \Theta^\varepsilon[V]w^\varepsilon = 0, \quad (x, \xi) \in \mathbb{R}_x \times \mathbb{R}_\xi, \quad t \in \mathbb{R}, \quad (4.17)$$

$$w^\varepsilon(t = 0) = w^\varepsilon[u_0^\varepsilon], \quad (4.18)$$

where $\Theta^\varepsilon[V]$ is the pseudo-differential operator:

$$\Theta^\varepsilon[V]w^\varepsilon(x, \xi, t) := \frac{i}{(2\pi)^d} \int_{\mathbb{R}^d} \frac{V(x + \frac{\varepsilon}{2}v) - V(x - \frac{\varepsilon}{2}v)}{\varepsilon} \hat{w}^\varepsilon(x, v, t) e^{iv \cdot \xi} dv, \tag{4.19}$$

here \hat{w}^ε stands for the Fourier-transform with respect to ξ .

Taking ε to 0 gives the Vlasov-equation (2.18) with $Q(x, \xi) = \frac{|\xi|^2}{2} + V(x)$:

$$w_t^0 + \xi \cdot \partial_x w^0 - \partial_x V(x) \cdot \partial_\xi w^0 = 0, \quad (x, \xi) \in \mathbb{R}_x \times \mathbb{R}_\xi, \quad t \in \mathbb{R} \tag{4.20}$$

$$w^0(t = 0) = w_I^0 := \lim_{\varepsilon \rightarrow 0} w^\varepsilon[u_I^\varepsilon], \tag{4.21}$$

Consider now the first-order time-splitting spectral method (SP1). To understand the splitting error we remark that the time-splitting (4.3), (4.4) corresponds to the time-splitting of the Wigner equation (4.17)

$$w_t^\varepsilon + \xi \cdot \partial_x w^\varepsilon = 0, \quad t \in [t_n, t_{n+1}] \tag{4.22}$$

followed by

$$w_t^\varepsilon + \Theta^\varepsilon[V]w^\varepsilon = 0, \quad t \in [t_n, t_{n+1}]. \tag{4.23}$$

Clearly, the limit $\varepsilon \rightarrow 0$ can be carried out in (4.23) leaving $k = \Delta t$ fixed and we obtain the corresponding time-splitting of the limiting Vlasov equation (4.20):

$$w_t^0 + \xi \cdot \partial_x w^0 = 0, \quad t \in [t_n, t_{n+1}] \tag{4.24}$$

followed by

$$w_t^0 - \partial_x V \cdot \partial_\xi w^0 = 0, \quad t \in [t_n, t_{n+1}]. \tag{4.25}$$

Note that **no** other error is introduced by the splitting (SP1) since the time-integrations are performed **exactly**.

These considerations, which can be made rigorous easily, show that a **uniform** (i.e. ε -**independent**) time-stepping control

$$k = O(\delta) \tag{4.26}$$

combined with the spectral mesh-size control (4.16)(b) gives an $O(\delta)$ -error uniformly as $\varepsilon \rightarrow 0$ in the Wigner-function and consequently in all observable mean-values. Essentially this implies that a fixed number of grid points in every spatial oscillation of wavelength ε combined with ε -independent time-stepping is sufficient, uniformly as $\varepsilon \rightarrow 0$, to guarantee accurate observables. This strategy is clearly superior to FD-schemes, which require $\frac{k}{\varepsilon} \rightarrow 0$ and $\frac{h}{\varepsilon} \rightarrow 0$ even for the approximation of observables.

We refer to [45] for the application of the time-splitting spectral method to the Zakharov system, to [44] for the numerical solution of the Dirac-Maxwell system and to [6] for numerical studies of nonlinear Schrödinger equations. Also we refer to [3], [4], [41] for numerical simulations of the cubically nonlinear Gross-Pitaevskii Schrödinger equation (Bose-Einstein condensation) using time-splitting spectral methods.

5. Highly Oscillatory Periodic Potentials

One of the main problems in solid state physics is to describe the motion of electrons under the action of the periodic potential generated by the ionic cores. This problem has been studied from a physical, as well as from a mathematical point of view in, e.g., [1, 12, 52, 61, 78], resulting in a profound theoretical understanding of the novel dynamical features. Indeed one of the most striking effect, known as *Peirl's substitution*, is a modification of the dispersion relation for Schrödinger's equation, where the classical energy relation $E_{\text{free}}(k) = \frac{1}{2}|k|^2$ has to be replaced by $E_m(k)$, $m \in \mathbb{N}$, the energy corresponding to the m th *Bloch band* [11]. The basic idea behind this replacement is a separation of scales. More precisely one recognizes that experimentally imposed external, electromagnetic fields typically vary on much *larger* spatial scales than the periodic potential generated by the cores. Moreover those external fields can be considered *weak* in comparison to the periodic fields of the cores [2].

To study this problem, consider the Schrödinger equation for the wavefunction $u = u^\varepsilon(x, t)$ of the electrons in a *semiclassical* asymptotic scaling [15, 61, 74], in $m = 1$ dimensions

$$\begin{cases} i\varepsilon\partial_t u^\varepsilon = -\frac{\varepsilon^2}{2}\partial_{xx}u^\varepsilon + V_\Gamma\left(\frac{x}{\varepsilon}\right)u^\varepsilon + U(x)u^\varepsilon, & x \in \mathbb{R}, t \in \mathbb{R}, \\ u|_{t=0} = u^\varepsilon_1(x). \end{cases} \quad (5.1)$$

Here, U is the external potential and the highly oscillating *lattice-potential* $V_\Gamma(y) \in \mathbb{R}$ is assumed to be *periodic* with respect to some *regular lattice* Γ . For definiteness we shall assume that

$$V_\Gamma(y + 2\pi) = V_\Gamma(y) \quad \forall y \in \mathbb{R}, \quad (5.2)$$

i.e. $\Gamma = 2\pi\mathbb{Z}$. For practical purposes we have to numerically solve (5.1) on a bounded computational domain \mathcal{D} , which we shall specify later on.

6. The Emergence of Bloch Bands

First, let us introduce some notations and recall some basic definitions used when dealing with periodic Schrödinger operators [2, 10, 74, 76].

With V_Γ obeying (5.2) we have:

- The fundamental domain of our lattice $\Gamma = 2\pi\mathbb{Z}$ is the interval $\mathcal{C} = (0, 2\pi)$.
- The *dual lattice* Γ^* can then be defined as the set of all wave numbers $k \in \mathbb{R}$, for which plane waves of the form $\exp(ikx)$ have the same periodicity as the potential V_Γ . This yields $\Gamma^* = \mathbb{Z}$ in our case.
- The fundamental domain of the dual lattice, e.g.the (first) *Brillouin zone*, $\mathcal{B} = \mathcal{C}^*$ is the set of all $k \in \mathbb{R}$ closer to zero than to any other dual lattice point. In our case $\mathcal{B} = \left(-\frac{1}{2}, \frac{1}{2}\right)$.

6.1. Recapitulation of Bloch’s decomposition method. One of our main points in what follows is that the dynamical behavior of (5.1) is mainly governed by the periodic part of the Hamiltonian, in particular for $\varepsilon \ll 1$. Thus it will be important to study its spectral properties. To this end consider the periodic *Hamiltonian* (where for the moment we set $y = x/\varepsilon$)

$$H = -\frac{1}{2} \partial_{yy} + V_\Gamma(y), \tag{6.1}$$

which we shall consider here only on $L^2(\mathcal{C})$. This is sufficient since the periodicity of V_Γ allows to cover all of \mathbb{R} by translations of \mathcal{C} . More precisely, for $k \in \overline{\mathcal{B}} = [-\frac{1}{2}, \frac{1}{2}]$ we equip the operator H with the following *quasi-periodic* boundary conditions

$$\begin{cases} u(y + 2\pi, t) = e^{2ik\pi} u(y, t) & \forall y \in \mathbb{R}, k \in \overline{\mathcal{B}}, \\ \partial_y u(y + 2\pi, t) = e^{2ik\pi} \partial_y u(y, t) & \forall y \in \mathbb{R}, k \in \overline{\mathcal{B}}. \end{cases} \tag{6.2}$$

It is well known [76] that under very mild conditions on V_Γ , the operator H admits a complete set of eigenfunction $\varphi_m(y, k), m \in \mathbb{N}$, providing, for each fixed $k \in \overline{\mathcal{B}}$, an orthonormal basis in $L^2(\mathcal{C})$. Correspondingly there exists a countable family of real-valued eigenvalues which can be ordered according to $E_1(k) \leq E_2(k) \leq \dots \leq E_m(k) \leq \dots, m \in \mathbb{N}$, where the respective multiplicities are accounted for in the ordering. The set $\{E_m(k) \mid k \in \mathcal{B}\} \subset \mathbb{R}$ is called the *m*th *energy band* of the operator H and the eigenfunction $\varphi_m(\cdot, k)$ is usually called *Bloch function*. (In the following the index $m \in \mathbb{N}$ will *always* denote the *band index*.) Concerning the dependence on $k \in \mathcal{B}$, it has been shown [76] that for any $m \in \mathbb{N}$ there exists a closed subset $\mathcal{A} \subset \mathcal{B}$ such that: $E_m(k)$ is analytic and $\varphi_m(\cdot, k)$ can be chosen to be a real analytic function for all $k \in \overline{\mathcal{B}} \setminus \mathcal{A}$. Moreover

$$E_{m-1} < E_m(k) < E_{m+1}(k) \quad \forall k \in \overline{\mathcal{B}} \setminus \mathcal{A}. \tag{6.3}$$

If this condition indeed holds for all $k \in \mathcal{B}$ then $E_m(k)$ is called an *isolated Bloch band* [74]. Moreover, it is known that

$$\text{meas } \mathcal{A} = \text{meas } \{k \in \overline{\mathcal{B}} \mid E_n(k) = E_m(k), n \neq m\} = 0. \tag{6.4}$$

This set of Lebesgue measure zero consists of the so called *band crossings*. Note that due to (6.2) we can rewrite $\varphi_m(y, k)$ as

$$\varphi_m(y, k) = e^{iky} \chi_m(y, k) \quad \forall m \in \mathbb{N}, \tag{6.5}$$

for a 2π -periodic function $\chi_m(\cdot, k)$. In terms of $\chi_m(y, k)$ the *Bloch eigenvalue problem* reads

$$\begin{cases} H(k)\chi_m(y, k) = E_m(k)\chi_m(y, k), \\ \chi_m(y + 2\pi, k) = \chi_m(y, k) \quad \forall k \in \mathcal{B}, \end{cases} \tag{6.6}$$

where $H(k)$ denotes the shifted Hamiltonian

$$H(k) := \frac{1}{2}(-i\partial_y + k)^2 + V_\Gamma(y). \tag{6.7}$$

Let us now introduce the so-called *Bloch transform* \mathcal{T} of the wave-function $u(t, \cdot) \in L^2(\mathbb{R})$, for any fixed $t \in \mathbb{R}$, as can be found in, e.g., [61, 74]. The Bloch transformation \mathcal{T} is the Fourier transform \mathcal{F} on $\ell^2(\Gamma)$ followed by a multiplication with e^{-iyk} , i.e.

$$(\mathcal{T}u)(k, y, t) := \sum_{\gamma \in \mathbb{Z}} u(y + 2\pi\gamma, t) e^{-ik(2\pi\gamma+y)}, \quad y \in \mathcal{C}, \quad k \in \mathcal{B}. \tag{6.8}$$

It is then easy to see that

$$\mathcal{T}H\mathcal{T}^{-1} = H(k). \tag{6.9}$$

which provides a link between the eigenvalue problem (6.6) and the periodic part of our Schrödinger equation acting on $u(t, \cdot)$.

Most importantly, the Bloch transformation allows to decompose the state space $\mathcal{H} = L^2(\mathbb{R})$ into a direct sum of so called *band spaces*, i.e.

$$L^2(\mathbb{R}) = \bigoplus_{m=1}^{\infty} \mathcal{H}_m, \quad \mathcal{H}_m := \left\{ u_m(y) = \int_{\mathcal{B}} f(k) \varphi_m(y, k) dk, \quad f \in L^2(\mathcal{B}) \right\}. \tag{6.10}$$

This is the well known *Bloch decomposition method*, which implies that for

$$u(\cdot, t) \in L^2(\mathbb{R}) : \quad u(y, t) = \sum_{m \in \mathbb{N}} u_m(y, t), \quad u_m \in \mathcal{H}_m. \tag{6.11}$$

The corresponding projection of $u(t)$ onto the m th band space is thereby given as

$$u_m(y, t) \equiv (\mathbb{P}_m u)(y, t) = \int_{\mathcal{B}} \left(\int_{\mathbb{R}} u(t, \zeta) \overline{\varphi}_m(\zeta, k) d\zeta \right) \varphi_m(y, k) dk \tag{6.12}$$

and we consequently denote by

$$C_m(k, t) := \int_{\mathbb{R}} u(\zeta, t) \overline{\varphi}_m(\zeta, k) d\zeta \tag{6.13}$$

the coefficients of the Bloch decomposition.

For a complete description and a rigorous mathematical proof of this decomposition we refer to, e.g., [64], chapter XI. Here it is only important to note that the Bloch transformation allows to obtain a spectral decomposition of periodic Hamiltonians H , upon solving the eigenvalue problem (6.6). Roughly speaking \mathcal{T} can be seen a Fourier-type transform adapted to the inclusion of periodic potentials.

This implies that, if $U \equiv 0$, we can indeed Bloch transform the evolution problem (5.1) and decompose it into the corresponding band spaces \mathcal{H}_m , i.e. we find a “diagonalization” of our evolution problem. In this case each $u_m(\cdot, t) \in \mathcal{H}_m$ then evolves according to the newly obtained PDE

$$\begin{cases} i\varepsilon \partial_t u_m = E_m(-i\partial_y)u_m, & y \in \mathbb{R}, t \in \mathbb{R}, \\ u_m|_{t=0} = (\mathbb{P}_m u_{\text{in}})(y). \end{cases} \tag{6.14}$$

Here $E_m(-i\partial_y)$ denotes the pseudo-differential operator corresponding to the symbol $E_m(k)$, cf. [31, 61, 74] and $u_{\text{in}}(y) = u_I^\varepsilon(\varepsilon y)$. The above given evolution equation comprises a rigorous justification of Peirl’s substitution. Moreover (6.14) is easily solved invoking the standard Fourier transformation \mathcal{F} on $L^2(\mathbb{R})$, which yields

$$u_m(y, t) = \mathcal{F}^{-1} \left(e^{-iE_m(k)t/\varepsilon} (\mathcal{F}(\mathbb{P}_m u_{\text{in}}))(k) \right). \tag{6.15}$$

Here the energy band $E_m(k)$ is understood to be periodically extended to all of \mathbb{R} . To this end, note that the following relation holds

$$\mathcal{F}(u_m)(k, t) = C_m(k, t) (\mathcal{F}\chi_m)(0, k), \tag{6.16}$$

as can be shown by a lengthy but straightforward calculation.

Of course if $U \not\equiv 0$ (the non-periodic part of the potential) the time evolution (5.1) in general *mixes* all band spaces \mathcal{H}_m , i.e. we can no longer diagonalize the full Hamiltonian operator (which now involves also non-periodic coefficients). On the other hand, since $U(x) = U(\varepsilon y)$ varies only slowly on the fast (periodic) scale $y = x/\varepsilon$, one might hope that even if $U \not\equiv 0$, the *effective Schrödinger type equation*

$$\begin{cases} i\varepsilon \partial_t u_m^{\text{eff}} = E_m(-i\partial_y)u_m^{\text{eff}} + U(\varepsilon y)u_m^{\text{eff}}, & y \in \mathbb{R}, t \in \mathbb{R}, \\ u_m^{\text{eff}}|_{t=0} = (\mathbb{P}_m u_{\text{in}})(y), \end{cases} \tag{6.17}$$

holds true, at least approximately for small $\varepsilon \ll 1$. In other words, the slowly varying external potential is almost constant on the lattice scale and thus causes only a small perturbation of the band structure determined via (6.1). Indeed this is the case as has been rigorously proven in [15, 36, 61], using different analytical approaches (for a broader overview, see [74] and the references given therein). To this end one has to assume that the m ’th energy band is *isolated* from the rest of the spectrum.

If this is not the case, energy transfer of order $\mathcal{O}(1)$ can occur at band crossings, the so-called Landau-Zener phenomena.

6.2. Numerical computation of the Bloch bands. As a preparatory step for our algorithm we shall first calculate the energy bands

$E_m(k)$ numerically as follows. Analogously to [33, 48], we consider the potential $V_\Gamma \in C^1(\mathbb{R})$ and expand it into its Fourier series, i.e.

$$V_\Gamma(y) = \sum_{\lambda \in \mathbb{Z}} \widehat{V}(\lambda) e^{i\lambda y}, \quad \widehat{V}(\lambda) = \frac{1}{2\pi} \int_0^{2\pi} V_\Gamma(y) e^{-i\lambda y} dy. \tag{6.18}$$

Likewise, we expand the Bloch eigenfunctions $\chi_m(\cdot, k)$ into their Fourier series

$$\chi_m(y, k) = \sum_{\lambda \in \mathbb{Z}} \widehat{\chi}_m(\lambda, k) e^{i\lambda y}, \quad \widehat{\chi}_m(\lambda, k) = \frac{1}{2\pi} \int_0^{2\pi} \chi_m(y, k) e^{-i\lambda y} dy. \tag{6.19}$$

If $V_\Gamma \in C^\infty(\mathbb{R})$, the corresponding Fourier coefficients $\widehat{V}(\lambda)$ decay faster than any power, as $\lambda \rightarrow \pm\infty$, and thus we only need to take into account a few coefficients.

For $\lambda \in \{-\Lambda, \dots, \Lambda - 1\} \subset \mathbb{Z}$, we consequently aim to approximate the Sturm-Liouville problem (6.6), by the following algebraic eigenvalue problem

$$\mathbf{H}(k) \begin{pmatrix} \widehat{\chi}_m(-\Lambda) \\ \widehat{\chi}_m(1 - \Lambda) \\ \vdots \\ \widehat{\chi}_m(\Lambda - 1) \end{pmatrix} = E_m(k) \begin{pmatrix} \widehat{\chi}_m(-\Lambda) \\ \widehat{\chi}_m(1 - \Lambda) \\ \vdots \\ \widehat{\chi}_m(\Lambda - 1) \end{pmatrix} \tag{6.20}$$

where the $2\Lambda \times 2\Lambda$ matrix $\mathbf{H}(k)$ is given by

$$\mathbf{H}(k) = \begin{pmatrix} \widehat{V}(0) + \frac{1}{2}(k - \Lambda)^2 & \widehat{V}(-1) & \cdots & \widehat{V}(1 - 2\Lambda) \\ \widehat{V}(1) & \widehat{V}(0) + \frac{1}{2}(k - \Lambda + 1)^2 & \cdots & \widehat{V}(2 - 2\Lambda) \\ \vdots & \vdots & \ddots & \vdots \\ \widehat{V}(2\Lambda - 1) & \widehat{V}(2\Lambda - 2) & \cdots & \widehat{V}(0) + \frac{1}{2}(k + \Lambda - 1)^2 \end{pmatrix} \tag{6.21}$$

The above matrix $\mathbf{H}(k)$ has 2Λ eigenvalues. Clearly, this number has to be large enough such that all the eigenvalues $E_m(k)$, which we shall need, are accounted for, i.e. we need $m \leq 2\Lambda$. The numerical cost of this algebraic problem is about $\mathcal{O}(\Lambda^3)$, cf. [39]. The number Λ is *independent* of the spatial grid, thus the numerical costs of this eigenvalue problem are often negligible compared to those of the evolutionary algorithms as detailed below. The approximate numerical computations of the Bloch bands $E_m(k)$ can be seen as a preprocessing, to be done only once.

Remark 6.1. *Accurate computations of the energy bands needed in practical applications, i.e. in more than one spatial dimensions and for different kinds of (composite) materials, become a highly nontrivial task. Nowadays though, there already exists a huge amount of numerical data detailing the energy band*

structure of the most important materials used in, e.g., the design of semiconductor devices, cf. [26, 47, 49]. In the context of photonic crystals the situation is similar [37]. Thus, relying on such data one can in principle avoid the above given eigenvalue-computations (and its generalizations to more dimensions) completely. To this end, one should also note that, given the energy bands $E_m(k)$, we do not need any knowledge about V_Γ in order to solve (5.1) numerically, cf. the algorithm described below. Also we remark that it was shown in [42] that the Bloch decomposition-based time splitting method is remarkably stable with respect to perturbations of the spectral data.

7. Bloch Decomposition Based Algorithm

For practical reasons we shall now introduce, for any fixed $t \in \mathbb{R}$, a new unitary transformation of $u(\cdot, t) \in L^2(\mathbb{R})$:

$$\tilde{u}(y, k, t) := \sum_{\gamma \in \mathbb{Z}} u(\varepsilon(y + 2\pi\gamma), t) e^{-i2\pi k\gamma}, \quad y \in \mathcal{C}, \quad k \in \mathcal{B}, \quad (7.1)$$

which has the properties that \tilde{u} is quasi-periodic w.r.t $y \in \Gamma$ and periodic w.r.t. $k \in \Gamma^*$, i.e.

$$\tilde{u}(y + 2\pi, k, t) = e^{i2\pi k} \tilde{u}(y, k, t), \quad \tilde{u}(y, k + 1, t) = \tilde{u}(y, k, t). \quad (7.2)$$

One should note that \tilde{u} is *not* the standard Bloch transformation \mathcal{T} , as defined in (6.8), but it is indeed closely related to it via

$$(\mathcal{T}u)(y, k, t) = \tilde{u}(y, k, t) e^{-iyk}, \quad k \in \mathcal{B}, \quad (7.3)$$

for $\varepsilon = 1$. Furthermore, we have the following inversion formula

$$u(\varepsilon(y + 2\pi\gamma), t) = \int_{\mathcal{B}} \tilde{u}(y, k, t) e^{i2\pi k\gamma} dk, \quad (7.4)$$

which is again very similar to the one of the standard Bloch transformation [74]. The main advantage in using \tilde{u} , instead of $\mathcal{T}u$ itself, is that we can rely on a standard fast Fourier transform (FFT) in the numerical algorithm below. If one aims to use $\mathcal{T}u$ directly one would be forced to modify a given FFT code accordingly. A straightforward computation then shows that

$$C_m(k, t) = \int_{\mathcal{C}} \tilde{u}(\zeta, k, t) \overline{\varphi_m(\zeta, k)} d\zeta, \quad (7.5)$$

where $C_m(t, k)$ is the Bloch coefficient, defined in (6.13).

In what follows let the time step be $\Delta t = T/N$, for some $N \in \mathbb{N}$, $T > 0$. Suppose that there are $L \in \mathbb{N}$ lattice cells (of length $2\pi\varepsilon$) within the computational domain \mathcal{D} , which we fix as the interval $(0, 2\pi)$ for the following, i.e.

$L = \frac{1}{\epsilon}$. In this domain, the wave function u is numerically computed at $L \times R$ grid points, for some $R \in \mathbb{N}$. In other words we assume that there are R grid points in each lattice cell, which yields the following discretization

$$\begin{cases} k_\ell = -\frac{1}{2} + \frac{\ell - 1}{L}, & \text{where } \ell = \{1, \dots, L\} \subset \mathbb{N}, \\ y_r = \frac{2\pi(r - 1)}{R}, & \text{where } r = \{1, \dots, R\} \subset \mathbb{N}, \end{cases} \tag{7.6}$$

and thus we finally evaluate $u^n = u(t_n)$ at the grid points $x = \epsilon(2\pi\gamma + y)$, i.e.

$$x_{\ell,r} = \epsilon(2\pi(\ell - 1) + y_r). \tag{7.7}$$

We remark that in our numerical computations we can use $R \ll L$, whenever $\epsilon \ll 1$, i.e. we only use a few grid points within each cell. Now we shall describe precisely the Bloch decomposition based algorithm used to solve (5.1) in detail.

Suppose that at the time t_n we are given $u^\epsilon(x_{\ell,r}, t_n) \approx u_{\ell,r}^n$. Then $u_{\ell,r}^{n+1}$, i.e. the solution at the (next) time step $t_{n+1} = t_n + \Delta t$, is obtained as follows:

Step 1. First, we solve the equation

$$i\epsilon\partial_t u = -\frac{\epsilon^2}{2} \partial_{xx} u + V_\Gamma\left(\frac{x}{\epsilon}\right) u, \tag{7.8}$$

on the time-interval (t_n, t_{n+1}) of length Δt . To this end we shall use the Bloch-decomposition method, as detailed below.

Step 2. In a second step, solve the ordinary differential equation

$$i\epsilon\partial_t u = U(x)u, \tag{7.9}$$

on the same time-interval, where the solution obtained in Step 1 serves as initial condition for Step 2. We obtain the exact solution of this linear ODE by

$$u(x, t) = u(x, t_n) e^{-iU(x)\frac{t-t_n}{\epsilon}}. \tag{7.10}$$

Remark 7.1. *Clearly, the algorithm given above is first order in time. We can easily obtain a second order scheme by the Strang splitting method. Note that in both cases the schemes conserve the particle density $\rho(x, t) := |u(x, t)|^2$ on the fully discrete level.*

Indeed Step 1 consists of several intermediate steps which we shall present in what follows:

Step 1.1. We first compute \tilde{u} at time t^n by

$$\tilde{u}_{\ell,r}^n = \sum_{j=1}^L u_{j,r}^n e^{-ik_\ell \cdot x_{j,1}}. \tag{7.11}$$

Step 1.2. Next, we compute the m th band Bloch coefficient $C_m(k, t)$, at time t^n , via (7.5), i.e.

$$\begin{aligned}
 C_m(k_\ell, t_n) &\approx C_{m,\ell}^n = \frac{2\pi}{R} \sum_{r=1}^R \tilde{u}_{\ell,r}^n \overline{\chi_m}(y_r, k_\ell) e^{-ik_\ell y_r} \\
 &\approx \frac{2\pi}{R} \sum_{r=1}^R \tilde{u}_{\ell,r}^n \sum_{\lambda=-R/2}^{R/2-1} \widehat{\chi}_m(\lambda, k_\ell) e^{-i(k_\ell+\lambda)y_r},
 \end{aligned}
 \tag{7.12}$$

where for the second line we simply inserted the Fourier expansion of χ_m , given in (6.19). Note that in total we have R Fourier coefficients for χ_m . Clearly this implies that we need $\Lambda > R/2$, where Λ is the number of Fourier modes required in the numerical approximation of the Bloch eigenvalue problem as discussed above. Here we only take the R lowest frequency Fourier coefficients.

Step 1.3. The obtained Bloch coefficients are then evolved up to time t^{n+1} , according to the explicit solution formula (6.15), taking into account (6.16). This yields

$$C_{m,\ell}^{n+1} = C_{m,\ell}^n e^{-iE_m(k_\ell)\Delta t/\varepsilon}.
 \tag{7.13}$$

Step 1.4. We consequently compute \tilde{u} at time t^{n+1} by summing up all band contributions and using the analytical formulas (6.12) and (6.13), i.e.

$$\tilde{u}_{\ell,r}^{n+1} = \sum_{m=1}^M (\mathbb{P}_m \tilde{u})_{\ell,r}^{n+1} \approx \sum_{m=1}^M C_{m,\ell}^{n+1} \sum_{\lambda=-R/2}^{R/2-1} \widehat{\chi}_m(\lambda, k_\ell) e^{i(k_\ell+\lambda)y_r}.
 \tag{7.14}$$

Step 1.5. Finally we numerically perform the inverse transformation to (7.1), i.e. we compute $u_{\ell,r}^{n+1}$ from $\tilde{u}_{\ell,r}^{n+1}$. Thus from (7.4), we obtain

$$u_{\ell,r}^{n+1} = \frac{1}{L} \sum_{j=1}^L \tilde{u}_{j,r}^{n+1} e^{ik_j x_{\ell,1}}.
 \tag{7.15}$$

Note that in this algorithm, the main numerical costs are incurred via the FFT in Steps 1.1 and 1.5. This also implies that on the same spatial grid, the numerical costs of our Bloch transform based algorithm is of the same order as the classical time-splitting spectral method. Moreover, we stress the fact that if there is no external potential, i.e. $U(x) \equiv 0$, then the above given algorithm numerically computes the *exact* solution of the evolutionary problem (5.1). In particular this fact allows us to solve the Schrödinger equation (5.1) for very large time steps, even if ε is small (see the results given below). One should also note that a possible lack of regularity in V_Γ only requires numerical care when approximating (6.6) by the algebraic problem (6.20). In particular, V_Γ itself does not enter in the time-evolution but only $E_m(k)$.

Ignoring for a moment the additional structure provided by the periodic potential V_Γ , one might straight forwardly apply the time-splitting spectral

methods (SP1) or (SP2) of Section 4. It is clear however that, due to the inclusion of $V_\Gamma\left(\frac{x}{\varepsilon}\right)$, the exact solution of the splitting step involving both potentials, namely

$$u(x, t) = u(x, t_n) e^{-i(V_\Gamma(x/\varepsilon) + U(x))(t - t_n)/\varepsilon}, \quad (7.16)$$

features a coupling of high frequency oscillations in x and t , in contrast to (7.10), where only t/ε -oscillations are present.

Remark 7.2. *In our (BD) algorithm, we compute the dominant effects from dispersion and periodic lattice potential in one step, and treat the non-periodic potential as a perturbation. Because the split-step commutator error between the periodic and non-periodic parts is relatively small, the step size can be chosen considerably larger than for the (SP) algorithms.*

Remark 7.3. *For the (BD) algorithm, the computational complexities of Step 1.1 and 1.5 are $\mathcal{O}(RL \log(L))$, the complexities of Step 1.2 and 1.4 are $\mathcal{O}(MLR \log(R))$, and for Step 1.3 we have $\mathcal{O}(ML)$. Also the complexity of the eigenvalue problem (6.20) is $\mathcal{O}(\Lambda^3)$. However, since Λ (and R) is independent of ε and since we only need to solve the eigenvalue problem (6.20) once in a preparatory step, the computation costs for this problem are negligible. On the other hand, for the time-splitting spectral algorithm, the complexities of Step 1 and 2 are $\mathcal{O}(RL \log(RL))$ and $\mathcal{O}(RL)$ respectively. As M and R are independent of ε , we can use $R \ll L$ and $M \ll L$, whenever $\varepsilon \ll 1$. Thus the complexities of both algorithms in each time step are comparable.*

We refer to [42] for the application of the (BD) algorithm to wave propagation problems and to [43] for simulation of nonlinear Gross-Pitaevskii equations with periodic potentials. Finally, we point out that recently another interesting numerical technique for semiclassical Schrödinger equations, based on so called Gaussian beams, has been developed. We refer to [70], [68], [69].

Acknowledgements

This publication is based on work supported by Award No. KUK-I1-007-43, made by the King Abdullah University of Science and Technology (KAUST) and the Royal Society through the Wolfson Research Merit Award of Peter Markowich.

References

- [1] J. Asch and A. Knauf, *Motion in periodic potentials*, Nonlinearity, Vol. 11, 175–200, 1998
- [2] N. W. Ashcroft and N. D. Mermin, *Solid state physics*, Saunders New York, 1976

- [3] W.Z. Bao, D. Jaksch, P.A. Markowich, *Numerical solution of the Gross-Pitaevskii equation for Bose-Einstein condensation*, J. Comput. Phys., Vol. 187, pp. 318–342, 2003
- [4] W.Z. Bao, D. Jaksch, P.A. Markowich, *Three dimensional simulation of jet formation in collapsing condensates*, J. Phys. B: At. Mol. Opt. Phys., Vol. 37, No. 2, 329–343, 2004.
- [5] W. Z. Bao, S. Jin, and P. Markowich, *On time-splitting spectral approximations for the Schrödinger equation in the semiclassical regime*, J. Comp. Phys., Vol. 175, 487–524, 2002
- [6] W.Z. Bao, S. Jin and P.A. Markowich, *Numerical Study of Time-Splitting Spectral Discretizations of Nonlinear Schrödinger Equations in the Semi-Classical Regimes*, SIAM J. Sci. Comp., Vol. 25, 27–64, 2003
- [7] J.C. Bronski and D. W. McLaughlin, *Semiclassical behavior in the NLS equation: optical shocks - focusing instabilities*, *Singular Limits of Dispersive Waves*, Plenum Press, New York and London, 1994
- [8] P. Bechouche, N. Mauser, and F. Poupaud, *Semiclassical limit for the Schrödinger-Poisson equation in a crystal*, Comm. Pure Appl. Math., Vol. 54, no. 7, 851–890, 2001
- [9] P. Bechouche, F. Poupaud, *Semi-classical limit of a Schrödinger equation for a stratified material*, Monatsh. Math., Vol. 129, no. 4, 281–301, 2000
- [10] A. Bensoussan, J. L. Lions, and G. Papanicolaou, *Asymptotic Analysis for Periodic Structures*, North-Holland Pub. Co., 1978
- [11] F. Bloch, *Über die Quantenmechanik der Elektronen in Kristallgittern*, Z. Phys., Vol. 52, 555–600, 1928
- [12] E. I. Blount, *Formalisms of band theory*, Solid State Physics, Vol. 13, 305–373, 1962
- [13] K. Busch, *Photonic band structure theory: assessment and perspectives*, Comptes Rendus Physique, Vol. 3, 53–66, 2002
- [14] R. Carles, *WKB analysis for nonlinear Schrödinger equations with a potential*, Comm. Math. Phys., Vol. 269, 95–221, 2007
- [15] R. Carles, P. A. Markowich and C. Sparber, *Semiclassical asymptotics for weakly nonlinear Bloch waves*, J. Stat. Phys., Vol. 117, 369–401, 2004
- [16] F. Castella, G. Dujardin, *Propagation of Gevrey regularity over long times for the fully discrete Lie Trotter splitting scheme applied to the linear Schrödinger equation*, M2AN, Vol. 43, 651–676, 2009
- [17] T.F. Chan, D. Lee, L. Shen, *Stable explicit schemes for equations of the Schrödinger type*, SIAM J. Numer. Anal., Vol. 23, 274–281, 1986
- [18] T.F. Chan, L. Shen, *Stability analysis of difference scheme for variable coefficient Schrödinger type equations*, SIAM J. Numer. Anal., Vol. 24, 336–349, 1987.
- [19] C. Conca, R. Orive, M. Vanninathan, *Bloch approximation in homogenization on bounded domains*, Asymptot. Anal., Vol. 41, no. 1, 71–91, 2005
- [20] C. Conca, N. Srinivasan, M. Vanninathan, *Numerical solution of elliptic partial differential equations by Bloch waves method*, Congress on Differential Equations and Applications/VII CMA (Salamanca, 2001), 63–83, 2001.

-
- [21] C. Conca, M. Vanninathan, *Homogenization of periodic structures via Bloch decomposition*, SIAM J. Appl. Math., Vol. 57, no. 6, 1639–1659, 1997
- [22] W. Dörfler, *A time- and space adaptive algorithm for the linear time dependent Schrödinger equation*. Numer. Math., Vol 73, pp 419–448, 1998
- [23] Delfour, M., Fortin, M., Payre, G. *Finite-Difference Solutions of a Nonlinear Schrödinger Equation*. Journal of Computational Physics, Vol. 44, 277–288, 1981
- [24] G. Dujardin, E. Faou, *Long time behavior of splitting methods applied to the linear Schrödinger equation*, C.R. Acad. Sci. Paris, Sér. I, Vol. 344, 89–92, 2007
- [25] A. Fannjiang, S. Jin, G. Papanicolaou, *High frequency behavior of the focusing nonlinear Schrödinger equation with random inhomogeneities*, SIAM J. Appl. Math., Vol. 63, 1328–1358, 2008.
- [26] M. V. Fischetti and S. E. Laux, *Monte Carlo analysis of electron transport in small semiconductor devices including band-structure and space-charge effects*, Phys. Rev. B, Vol. 38, 9721–9745, 1998
- [27] I. Gasser, P.A. Markowich, *Quantum hydrodynamics, Wigner transforms and the classical limit*, Asymptotic Analysis, Vol. 14, 97–116, 1997.
- [28] L. Gauckler, C. Lubich, *Splitting integrators for nonlinear Schrödinger equations over long times*, Found. Comput. Math., to appear.
- [29] P. Gérard. *Mesures semi-classiques et ondes de Bloch.*, Sémin. Ecole Polytechnique XVI: 1-19, 1990–1991
- [30] P. Gérard, *Microlocal defect measures*, Comm. PDE., Vol. 16, 1761–1794, 1991.
- [31] P. Gérard, P.A. Markowich, N.J. Mauser, F. Poupaud, *Homogenization limits and Wigner transforms*, Comm. Pure Appl. Math., Vol. 50, 321–377, 1997
- [32] L. Gosse, *Multiphase semiclassical approximation of an electron in a one-dimensional crystalline lattice. II. Impurities, confinement and Bloch oscillations*, J. Comput. Phys., Vol. 201, no. 1, 344–375, 2004
- [33] L. Gosse and P. A. Markowich, *Multiphase semiclassical approximation of an electron in a one-dimensional crystalline lattice - I. Homogeneous problems*, J. Comput Phys., Vol. 197, 387–417, 2004
- [34] L. Gosse and N. Mauser, *Multiphase semiclassical approximation of an electron in a one-dimensional crystalline lattice. III. From ab initio models to WKB for Schrödinger-Poisson*, J. Comput. Phys., Vol. 211, no. 1, 326–346, 2006
- [35] D. Gottlieb, S.A. Orszag, *Numerical Analysis of Spectral Methods*, SIAM, Philadelphia, 1977
- [36] J. C. Guillot, J. Ralston, and E. Trubowitz, *Semiclassical asymptotics in solid-state physics*, Comm. Math. Phys., Vol. 116, 401–415, 1998
- [37] D. Hermann, M. Frank, K. Busch, and P. Wölfle, *Photonic band structure computations*, Optics Express, Vol. 8, 167–173, 2001
- [38] L. Hörmander, *The Analysis of Linear Partial Differential Operators III*. Springer, 1985
- [39] R. Horn, C. Johnson, *Matrix analysis*, Cambridge University Press, Cambridge, 1985.

- [40] Z. Huang, S. Jin, P.A. Markowich and C. Sparber, *A Bloch decomposition based split-step pseudo spectral method for quantum dynamics with periodic potentials*, SIAM J. Sci. Comput., Vol. 29, issue 2, 515–538, 2007
- [41] W.Bao and J. Shen, *A fourth-order time-splitting Laguerre-Hermite pseudo-spectral method for Bose-Einstein condensates*, SIAM J. Sci. Comput., Vol. 26, No. 6, pp. 2010–2028, 2005
- [42] Z.Y. Huang, S. Jin, P.A. Markowich, C. Sparber, *On the Bloch decomposition based spectral method for wave propagation in periodic media*, Wave Motion, Vol. 46, 15–28, 2009
- [43] Z.Y. Huang, S. Jin, P.A. Markowich, C. Sparber, *Numerical simulation of the nonlinear Schrödinger equation with multi-dimensional periodic potentials*, SIAM Multiscale Modeling and Simulation, Vol. 7, 539–564, 2008
- [44] Z.Y. Huang, S. Jin, P.A. Markowich, C. Sparber, C.X. Zheng, *A time-splitting spectral scheme for the Maxwell-Dirac system*, J. Comp. Phys., Vol. 208, 761–789, 2005
- [45] S. Jin, P.A. Markowich, C.X. Zheng, *Numerical simulation of a generalized Zakharov system*, J. Comp. Phys., Vol. 201, 376–395, 2004
- [46] S. Jin, Z. Xin, *Numerical passage from systems of conservation laws to Hamilton-Jacobi equations, and a relaxation scheme*, SIAM J. Num. Anal., Vol. 35, 2385–2404, 1998
- [47] J. D. Joannopoulos, M. L. Cohen, *Theory of Short Range Order and Disorder in Tetrahedrally Bonded Semiconductors*, Solid State Physics, Vol. 31, 1976
- [48] H. J. Korsch, M. Glück, *Computing quantum eigenvalues made easy*, Eur. J. Phys., Vol. 23, 413–425, 2002
- [49] S. E. Laux, M. V. Fischetti, D. J. Frank, *Monte Carlo analysis of semiconductor devices: the DAMOCLES program*, IBM Journal of Research and Development, Vol. 34, 466–494, 1990
- [50] B. Desjardins, C.K. Lin, T.C. Tso, *Semiclassical limit of the derivative nonlinear Schrödinger equation*, M³AS, Vol. 10, 261–285, 2000.
- [51] L.D. Landau, E.M. Lifschitz, *Lehrbuch der Theoretischen Physik III: Quantenmechanik*. Akademie-Verlag, 1985
- [52] J.M. Luttinger, *The effect of a magnetic field on electrons in a periodic potential*, Phys. Rev., Vol. 84, 814–817, 1951
- [53] P.L. Lions, T. Paul, *Sur les Mesures de Wigner*. Revista Mat. Iberoamericana., Vol. 9, 553–618, 1993
- [54] P.A. Markowich, T. Paul, C. Sparber, *Bohmian measures and their classical limits*. submitted 2009
- [55] P.A. Markowich, F. Poupaud, *The Pseudo-Differential Approach to Finite Differences Revisited*. Calcolo, Vol. 36, no. 3, 161–186, 1999
- [56] P.A. Markowich, N.J. Mauser, F. Poupaud, *A Wigner function approach to semiclassical limits: electrons in a periodic potential.*, J. Math. Phys., Vol. 35, 1066–1094, 1994

- [57] P.A. Markowich, P. Pietra, C. Pohl, *Weak Limits of Finite Difference Schemes for Schrödinger-Type Equations*, Publ. Ian 1035: 1–57, 1997
- [58] P.A. Markowich, P. Pietra and C. Pohl, *Numerical approximation of quadratic observables of Schrödinger-type equations in the semi-classical limit*, Numer. Math., Vol. 81, 595–630, 1999
- [59] P.A. Markowich, P. Pietra, C. Pohl and H.P. Stimming, *A Wigner-Measure Analysis of the Dufort-Frankel scheme for the Schrödinger equation*, SIAM J. Num. Anal., Vol. 40, 1281–1310, 2002
- [60] P.D. Miller, S. Kamvissis, *On the semiclassical limit of the focusing nonlinear Schrödinger equation*, Phys. Letters A, Vol. 247, 75–86, 1998
- [61] G. Panati, H. Spohn, S. Teufel, *Effective dynamics for Bloch electrons: Peierls substitution and beyond*, Comm. Math. Phys., Vol. 242, 547–578, 2003
- [62] D. Pathria, J.L. Morris, *Pseudo-spectral solution of nonlinear Schrödinger equation*, J. Comput. Phys., Vol. 87, 108–125, 1990.
- [63] J.E. Pasciak, *Spectral and pseudo-spectral methods for advection equations*, Math. Comp., Vol. 35, 1081–1092, 1980
- [64] M. Reed, B. Simon, *Methods of modern mathematical physics IV. Analysis of operators*, Academic Press, 1978
- [65] J.M. Sanz-Serna, V.S. Manoranjan, *A Method for the Integration in Time of Certain Partial Differential Equations*, Journal of Computational Physics, Vol. 52, 273–289, 1983
- [66] S. Jin, C.D. Levermore and D. W. McLaughlin, *The semiclassical limit of the defocusing NLS hierarchy*, Comm. Pure Appl. Math. LII, 613–654, 1999
- [67] S. Jin, C.D. Levermore, D. W. McLaughlin, *The behavior of solutions of the NLS equation in the semiclassical limit*, *Singular Limits of Dispersive Waves*, Plenum Press, New York and London, 1994
- [68] S. Jin, H. Wu, X. Yang, *A numerical study of the Gaussian beam methods for one-dimensional Schrödinger-Poisson equations*, J. Comp. Math., to appear.
- [69] S. Jin, H. Wu, X. Yang, *Semi-Eulerian and High Order Gaussian Beam Methods for the Schrödinger Equation in the Semiclassical Regime*, preprint.
- [70] S. Jin, H. Wu, X. Yang, Z. Huang, *Bloch Decomposition-Based Gaussian Beam Method for the Schrödinger equation with Periodic Potentials*, preprint.
- [71] J.C. Strikwerda. *Finite Difference Schemes and Partial Differential Equations*. Wadsworth & Brooks/Cole Advanced Books & Software. Pacific Grove, California. 1989
- [72] L. Tartar, *H-measures: a new approach for studying homogenization, oscillations and concentration effects in partial differential equations*, Proc. Roy. Soc. Edinburgh Sect. A, Vol. 115, 193–230, 1990
- [73] T.R. Taha, M.J. Ablowitz, *Analytical and Numerical Aspects of Certain Nonlinear Evolution Equations*. II. Numerical, Nonlinear Schrödinger Equations, Journal of Computational Physics, Vol. 55, 203–230, 1984
- [74] S. Teufel, *Adiabatic perturbation theory in quantum dynamics*, Lecture Notes in Mathematics 1821, Springer, 2003

- [75] E. Wigner, *On the Quantum Correction for the Thermodynamic Equilibrium.*, Phys. Rev., Vol. 40, 1932
- [76] C. H. Wilcox, *Theory of Bloch waves*, J. Anal. Math., Vol. 33, 146–167, 1978
- [77] L. Wu, *Dufort-Frankel-Type Methods for Linear and Nonlinear Schrödinger Equations*. SIAM J. Numer. Anal., Vol. 33 (4), 1526–1533, 1996
- [78] J. Zak, *Dynamics of electrons in solids in external fields*, Phys. Rev., Vol. 168, 686–695, 1968
- [79] A. Zettel, *Spectral theory and computational methods for Sturm-Liouville problems*, in D. Hinton and P. W. Schäfer (eds.). Lecture Notes in Pure and Applied Math., Vol. 191, Dekker 1997.

Why Adaptive Finite Element Methods Outperform Classical Ones

Ricardo H. Nochetto*

Abstract

Adaptive finite element methods (AFEM) are a fundamental numerical tool in science and engineering. They are known to outperform classical FEM in practice and deliver optimal convergence rates when the latter cannot. This paper surveys recent progress in the theory of AFEM which explains their success and provides a solid mathematical framework for further developments.

Mathematics Subject Classification (2010). Primary 65N30, 65N50, 65N15; Secondary 41A25.

Keywords. Finite element methods, a posteriori error estimates, adaptivity, contraction, approximation class, nonlinear approximation, convergence rates.

1. Introduction

Mathematically sound adaptive finite element methods (AFEM) have been the subject of intense research since the late 70's, starting with the pioneering work of Babuška [4, 3]. It is known to practitioners that AFEM can achieve optimal performance, measured as error vs degrees of freedom, in situations when classical FEM cannot. However, it took about 30 years to develop a theory for the energy norm that explains this behavior and provides solid mathematical foundations for further development. This paper presents this theory [10, 33], and its connection to nonlinear approximation [17], for the *model elliptic PDE*

$$-\operatorname{div}(\mathbf{A}\nabla u) = f \quad \text{in } \Omega, \tag{1}$$

with Ω a polyhedral domain of \mathbb{R}^d ($d \geq 2$), homogeneous Dirichlet boundary condition on $\partial\Omega$, and \mathbf{A} symmetric, bounded, and uniformly positive definite.

*Partially supported by NSF grant DMS-0807811.

Department of Mathematics and Institute of Physical Science and Technology, University of Maryland, College Park, MD 20742. E-mail: rhn@math.umd.edu.

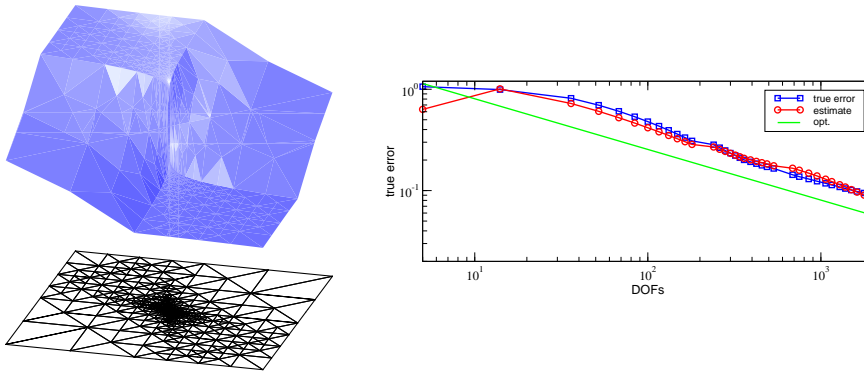


Figure 1. Discontinuous coefficients in checkerboard pattern: (a) graph of the discrete solution u , which is $u \approx r^{0.1}$, and underlying strongly graded grid \mathcal{T} towards the origin (notice the steep gradient of u at the origin); (b) estimate and true error in terms of $\#\mathcal{T}$ (the optimal decay for piecewise linear elements in 2d is indicated by the straight line with slope $-1/2$).

We start with a simple yet quite demanding example with discontinuous coefficients for $d = 2$ due to Kellogg [20], and used by Morin, Nochetto, and Siebert [25, 26] as a benchmark for AFEM. We consider $\Omega = (-1, 1)^2$, $\mathbf{A} = a_1 \mathbf{I}$ in the first and third quadrants, and $\mathbf{A} = a_2 \mathbf{I}$ in the second and fourth quadrants. This checkerboard pattern is the worst for the regularity of the solution u at the origin. For $f = 0$, a function of the form $u(r, \theta) = r^\gamma \mu(\theta)$ in polar coordinates solves (1) with nonvanishing Dirichlet condition for suitable $0 < \gamma < 1$ and μ [25, 26, 28]. We choose $\gamma = 0.1$, which leads to $u \in H^s(\Omega)$ for $s < 1.1$ and piecewise in W_p^1 for some $p > 1$. This corresponds to diffusion coefficients $a_1 \cong 161.44$ and $a_2 = 1$, which can be computed via Newton's method; the closer γ is to 0, the larger is the ratio a_1/a_2 . The solution u and a sample mesh are depicted in Figure 1(a).

Figure 1(b) documents the optimal performance of AFEM: both the energy error and estimator exhibit optimal decay $(\#\mathcal{T})^{-1/2}$ in terms of the cardinality $\#\mathcal{T}$ of the underlying mesh \mathcal{T} for piecewise linear finite elements. On the other hand, Figure 2 displays a strongly graded mesh \mathcal{T} towards the origin generated by AFEM using bisection, and three zooms which reveal a selfsimilar structure. It is worth stressing that the meshsize is of order 10^{-10} at the origin and $\#\mathcal{T} \approx 2 \times 10^3$, whereas to reach a similar resolution with a uniform mesh \mathcal{T} we would need $\#\mathcal{T} \approx 10^{20}$. This example clearly reveals that adaptivity can restore optimal performance even with modest computational resources.

Classical FEM with quasi-uniform meshes \mathcal{T} require regularity $u \in H^2(\Omega)$ to deliver an optimal convergence rate $(\#\mathcal{T})^{-1/2}$. Since $u \notin H^s(\Omega)$ for any $s > 1.1$, this is not possible for the example above. However, the problem is not quite the lack of second derivatives, but rather the fact that they are not

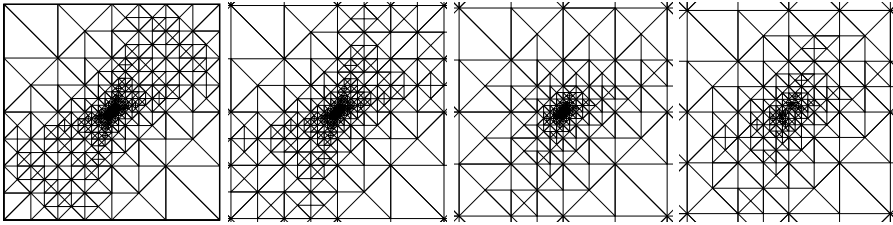


Figure 2. Discontinuous coefficients in checkerboard pattern: (a) final grid \mathcal{T} highly graded towards the origin with $\#\mathcal{T} \approx 2000$; (b) zoom to $(-10^{-3}, 10^{-3})^2$; (c) zoom to $(-10^{-6}, 10^{-6})^2$; (d) zoom to $(-10^{-9}, 10^{-9})^2$. For a similar resolution, a uniform grid \mathcal{T} would require $\#\mathcal{T} \approx 10^{20}$.

square integrable. In fact, the function u is in W_p^2 for $p > 1$ in each quadrant, and so over the initial mesh \mathcal{T}_0 , namely $u \in W_p^2(\Omega; \mathcal{T}_0)$.

To measure the performance of AFEM we introduce an approximation class \mathcal{A}_s for $s > 0$. Given an initial grid \mathcal{T}_0 , and the set \mathbb{T}_N of all conforming refinements \mathcal{T}_0 by *bisection* with at most N elements more than \mathcal{T}_0 , we consider the best error

$$\sigma_N(u) := \inf_{\mathcal{T} \in \mathbb{T}_N} \inf_{V \in \mathbb{V}(\mathcal{T})} \|u - V\|_\Omega \tag{2}$$

in the *energy norm* $\|\cdot\|_\Omega = \|\mathbf{A}^{1/2} \nabla \cdot\|_{L^2(\Omega)}$, where $\mathbb{V}(\mathcal{T}) \subset H_0^1(\Omega)$ is the conforming finite element space of piecewise polynomials of degree $\leq n$ with $n \geq 1$ over \mathcal{T} . We say that $u \in \mathcal{A}_s$ if

$$\sigma_N(u) \lesssim N^{-s}. \tag{3}$$

We wonder whether or not AFEM is able to deliver this asymptotic error decay. If we have access to the local energy error, we give a constructive proof in §3 of the fact that for $d = 2$

$$u \in W_p^2(\Omega; \mathcal{T}_0) \cap H_0^1(\Omega) \quad \Rightarrow \quad u \in \mathcal{A}_{1/2}. \tag{4}$$

This shows that piecewise linear finite element approximations can deliver optimal error decay. However, we only have indirect access to the solution u of (1) via the error estimators, so it is highly nontrivial whether a similar result holds for the Galerkin solution given by AFEM. The answer to this question requires two steps:

- *Contraction property*: we show in §5.1 that the energy error contracts provided the data is piecewise constant (so that the oscillation vanishes) and the interior node property holds. Otherwise, we identify in §5.2 a novel contractive quantity for general data, the so-called *quasi-error*, and prove that AFEM contracts it.
- *Convergence rate*: we show in §6.3 that the class \mathcal{A}_s is adequate provided the oscillation vanishes. However, the concept of approximation class for AFEM

is generally more involved than just \mathcal{A}_s because it entails dealing with the *total error*, namely the sum of energy error and oscillation. We discuss this issue in §6.1 and §6.2, and next prove that AFEM delivers a convergence rate similar to (3) up a multiplicative constant in §6.4.

It is worth stressing that AFEM learns about the decay rate $s > 0$ via the estimator. In fact, this exponent is never used in the design of AFEM. We discuss the basic modules of AFEM along with their key properties in §4, and the properties of bisection in §2.

2. The Bisection Method

We briefly discuss the *bisection* method, the most elegant and successful technique for subdividing Ω in any dimension into a conforming mesh made of simplices. We mention the recursive algorithms by Mitchell [24] for $d = 2$ and Kossaczky [21] for $d = 3$. We focus on the special case $d = 2$, and follow Binev, Dahmen, and DeVore [5] and Nochetto and Veerer [29], but the key Theorem 2 holds for any $d \geq 2$ as shown by Stevenson [34]. We refer to Nochetto, Siebert, and Veerer [28] for a rather complete discussion for $d \geq 2$.

2.1. Definition and Properties of Bisection. Let \mathcal{T} denote a *mesh* (triangulation or grid) made of simplices T , and let \mathcal{T} be *conforming* (edge-to-edge). Each element is labeled, namely it has an edge $E(T)$ assigned for refinement (and an opposite vertex $v(T)$ for $d = 2$); see Figure 3.

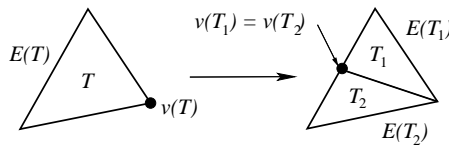


Figure 3. Triangle $T \in \mathcal{T}$ with vertex $v(T)$ and opposite refinement edge $E(T)$. The bisection rule for $d = 2$ consists of connecting $v(T)$ with the midpoint of $E(T)$, thereby giving rise to children T_1, T_2 with common vertex $v(T_1) = v(T_2)$, the newly created vertex, and opposite refinement edges $E(T_1), E(T_2)$.

The bisection method consists of a suitable *labeling* of the initial mesh \mathcal{T}_0 and a rule to assign the refinement edge to the two children. For $d = 2$ we follow Mitchell [24] and consider the *newest vertex bisection* as depicted in Figure 3. For $d > 2$ the situation is more complicated and one needs the concepts of type and vertex order [21, 28, 34].

Let \mathbb{T} be the set of all conforming bisection refinements of \mathcal{T}_0 . If $\mathcal{T}_* \in \mathbb{T}$ is a conforming refinement of $\mathcal{T} \in \mathbb{T}$, we write $\mathcal{T}_* \geq \mathcal{T}$. For instance, Figure 4 displays a sequence $\{\mathcal{T}_k\}_{k=0}^2$ with $\mathcal{T}_0 = \{T_i\}_{i=1}^4$ and $\mathcal{T}_k \geq \mathcal{T}_{k-1}$ obtained by bisecting the longest edge.

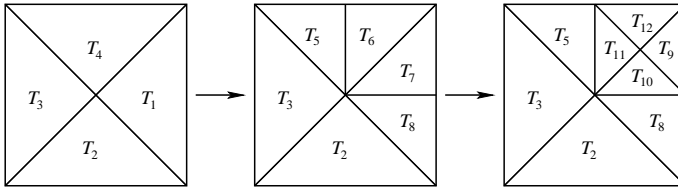


Figure 4. Sequence of bisection meshes $\{\mathcal{T}_k\}_{k=0}^2$ starting from the initial mesh $\mathcal{T}_0 = \{T_i\}_{i=1}^4$ with longest edges labeled for bisection. Mesh \mathcal{T}_1 is created from \mathcal{T}_0 upon bisecting T_1 and T_4 , whereas mesh \mathcal{T}_2 arises from \mathcal{T}_1 upon refining T_6 and T_7 . The bisection rule is described in Figure 3.

The following assertion about element shape is valid for $d \geq 2$ but we state it for $d = 2$.

Lemma 1 (Shape regularity). *The partitions $\mathcal{T} \in \mathbb{T}$ generated by newest vertex bisection satisfy a uniform minimal angle condition, or equivalently the maximal ratio of element diameter over diameter of largest inscribed ball for all $T \in \mathcal{T}$ is uniformly bounded, only depending on the initial partition \mathcal{T}_0 .*

We define the *generation* $g(T)$ of an element $T \in \mathcal{T}$ as the number of bisections needed to create T from its ancestor $T_0 \in \mathcal{T}_0$. Since bisection splits an element into two children with equal measure, we realize that

$$h_T = |T|^{1/2} = 2^{-g(T)/2} h_{T_0} \quad \text{for all } T \in \mathcal{T}. \tag{5}$$

Whether the recursive application of bisection does not lead to inconsistencies depends on a suitable initial *labeling of edges* and a *bisection rule*. For $d = 2$ they are simple to state [5], but for $d > 2$ we refer to Condition (b) of Section 4 of [34]. Given $T \in \mathcal{T}$ with generation $g(T) = i$, we assign the label $(i + 1, i + 1, i)$ to T with i corresponding to the refinement edge $E(T)$. The following rule dictates how the labeling changes with refinement: the side i is bisected and both new sides as well as the bisector are labeled $i + 2$ whereas the remaining labels do not change. To guarantee that the label of an edge is independent of the elements sharing this edge, we need a special labeling for \mathcal{T}_0 [5]:

$$\begin{aligned} &\text{edges of } \mathcal{T}_0 \text{ have labels 0 or 1 and all elements } T \in \mathcal{T} \text{ have} \\ &\text{exactly two edges with label 1 and one with label 0.} \end{aligned} \tag{6}$$

It is not obvious that such a labeling exists, but if it does then all elements of \mathcal{T}_0 can be split into pairs of compatibly divisible elements. We refer to Figure 5 for an example of initial labeling of \mathcal{T}_0 satisfying (6) and the way it evolves for two successive refinements $\mathcal{T}_2 \geq \mathcal{T}_1 \geq \mathcal{T}_0$ corresponding to Figure 4. Condition (6) can be enforced for $d = 2$ upon bisecting twice each element of \mathcal{T}_0 and labeling 0 the two newest edges [5]. For $d > 2$ the construction is much trickier [34].

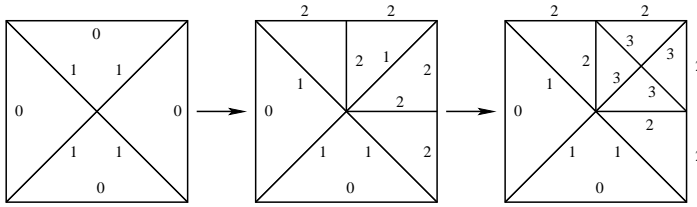


Figure 5. Initial labeling and its evolution for the sequence of conforming refinements of Figure 4.

2.2. Complexity of Bisection. Given $\mathcal{T} \in \mathbb{T}$ and a subset $\mathcal{M} \subset \mathcal{T}$ of marked elements to be refined, the procedure

$$\mathcal{T}_* = \text{REFINE}(\mathcal{T}, \mathcal{M})$$

creates a new conforming refinement \mathcal{T}_* of \mathcal{T} by bisecting all elements of \mathcal{M} at least once and perhaps additional elements to keep conformity.

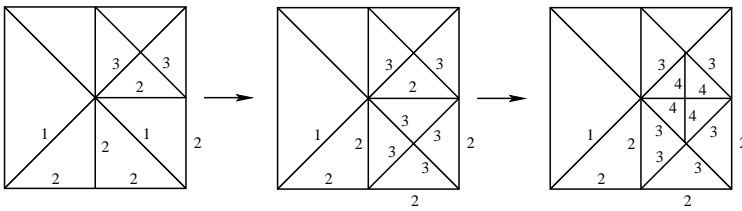


Figure 6. Recursive refinement of $T_{10} \in \mathcal{T}$ in Figures 4 and 5. This entails refining the chain $\{T_{10}, T_8, T_2\}$, starting from the last element $T_2 \in \mathcal{T}$, which form alone a compatible bisection patch because its refinement edge is on the boundary, and continuing with $T_8 \in \mathcal{T}$ and finally $T_{10} \in \mathcal{T}$. Note that the successive meshes are always conforming, that each element in the chain is bisected twice before getting back to T_{10} , and that $\#\{T_{10}, T_8, T_2\} = g(T_{10}) = 3$.

Conformity is a constraint in the refinement procedure that prevents it from being completely local. The propagation of refinement beyond the set of marked elements \mathcal{M} is a rather delicate matter. Figure 6 shows that a naive estimate of the form

$$\#\mathcal{T}_* - \#\mathcal{T} \leq \Lambda_0 \#\mathcal{M}$$

is *not* valid with an absolute constant Λ_0 independent of the refinement level because the constant may be as large as $g(T)$ with $T \in \mathcal{M}$.

This can be repaired upon considering the cumulative effect for a sequence of conforming bisection meshes $\{\mathcal{T}_k\}_{k=0}^\infty$. This is expressed in the following crucial complexity result due to Binev, Dahmen, and DeVore [5] for $d = 2$ and Stevenson [34] for $d > 2$. We refer to Nochetto, Siebert and Veeger [28] for a complete discussion for $d \geq 2$.

Theorem 2 (Complexity of REFINE). *If \mathcal{T}_0 satisfies the initial labeling (6) for $d = 2$, or that in [34, Section 4] for $d > 2$, then there exists a constant $\Lambda_0 > 0$ only depending on \mathcal{T}_0 and d such that for all $k \geq 1$*

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \leq \Lambda_0 \sum_{j=0}^{k-1} \#\mathcal{M}_j.$$

If elements $T \in \mathcal{M}$ are to be bisected $b \geq 1$ times, then the procedure REFINE can be applied recursively, and Theorem 2 remains valid with Λ_0 also depending on b .

3. Piecewise Polynomial Interpolation

3.1. Quasi-interpolation. If $v \in C^0(\bar{\Omega})$ we define the *Lagrange interpolant* $I_{\mathcal{T}}v$ of v as follows:

$$I_{\mathcal{T}}v(x) = \sum_{z \in \mathcal{N}} v(z)\phi_z(x).$$

For functions without point values, such as those in $H^1(\Omega)$ for $d > 1$, we need to determine nodal values by averaging. For any conforming refinement $\mathcal{T} \geq \mathcal{T}_0$ of \mathcal{T}_0 , the averaging process extends beyond nodes and so gives rise to the discrete neighborhood

$$N_{\mathcal{T}}(T) := \{T' \in \mathcal{T} \mid T' \cap T \neq \emptyset\} \quad \text{for all } T \in \mathcal{T}$$

which satisfies $\max_{T \in \mathcal{T}} \#N_{\mathcal{T}}(T) \leq C(\mathcal{T}_0)$ and $\max_{T' \in N_{\mathcal{T}}(T)} \frac{|T'|}{|T|} \leq C(\mathcal{T}_0)$ with $C(\mathcal{T}_0)$ depending only on the shape coefficient of \mathcal{T}_0 . We consider now the *quasi-interpolation operator* $I_{\mathcal{T}} : W_1^1(\Omega) \rightarrow \mathbb{V}(\mathcal{T})$ due to Scott and Zhang [9, 30]. For $n = 1$ it reads

$$I_{\mathcal{T}}v = \sum_{z \in \mathcal{N}(\mathcal{T})} \langle v, \phi_z^* \rangle \phi_z,$$

where $\{\phi_z^*\}_{z \in \mathcal{N}(\mathcal{T})}$ is a suitable set of dual functions for each node z so that $I_{\mathcal{T}}v = 0$ on $\partial\Omega$ provided $v = 0$ on $\partial\Omega$. We recall the notion of *Sobolev number*: $\text{sob}(W_p^s) = s - d/p$.

Proposition 3 (Local interpolation error). *Let s, t be regularity indices with $0 \leq t \leq s \leq n + 1$, and $1 \leq p, q \leq \infty$ be integrability indices so that $\text{sob}(W_p^s) > \text{sob}(W_q^t)$. The quasi-interpolation operator $I_{\mathcal{T}}$ is invariant in $\mathbb{V}(\mathcal{T})$ and satisfies for $s \geq 1$*

$$\|D^t(v - I_{\mathcal{T}}v)\|_{L^q(T)} \lesssim h_T^{\text{sob}(W_p^s) - \text{sob}(W_q^t)} \|D^s v\|_{L^p(N_{\mathcal{T}}(T))} \quad \text{for all } T \in \mathcal{T}, \quad (7)$$

provided \mathcal{T} is shape regular. Moreover, if $\text{sob}(W_p^2) > 0$, then v is continuous and (7) remains valid with $I_{\mathcal{T}}$ replaced by the Lagrange interpolation operator and $N_{\mathcal{T}}(T)$ by T .

3.2. Principle of Error Equidistribution. We investigate the relation between local meshsize and regularity for the design of graded meshes adapted to a given function $v \in H^1(\Omega)$ for $d = 2$. We formulate this as an optimization problem:

Given a function $v \in C^2(\Omega) \cap W_p^2(\Omega)$ and an integer $N > 0$ find conditions for a shape regular mesh \mathcal{T} to minimize the error $|v - I_{\mathcal{T}}v|_{H^1(\Omega)}$ subject to the constraint that the number of degrees of freedom $\#\mathcal{T} \leq N$.

We first convert this *discrete* optimization problem into a *continuous model*, following Babuška and Rheinboldt [4]. Let

$$\#\mathcal{T} = \int_{\Omega} \frac{dx}{h(x)^2}$$

be the number of elements of \mathcal{T} and let the Lagrange interpolation error

$$\|\nabla(v - I_{\mathcal{T}}v)\|_{L^2(\Omega)}^p = \int_{\Omega} h(x)^{2(p-1)} |D^2v(x)|^p dx$$

be dictated by (7) with $s = 2$ and $1 < p \leq 2$; note that $r = \text{sob}(W_p^2) - \text{sob}(H^1) = 2 - 2/p$ whence $rp = 2(p - 1)$ is the exponent of $h(x)$. We next propose the Lagrangian

$$\mathcal{L}[h, \lambda] = \int_{\Omega} \left(h(x)^{2(p-1)} |D^2v(x)|^p - \frac{\lambda}{h(x)^2} \right) dx$$

with Lagrange multiplier $\lambda \in \mathbb{R}$. The optimality condition reads $h(x)^{2(p-1)+2} |D^2v(x)|^p = \Lambda$, where $\Lambda > 0$ is a constant. To interpret this expression, we compute the interpolation error E_T incurred in element $T \in \mathcal{T}$. According to Proposition 3, E_T is given by

$$E_T^p \approx h_T^{2(p-1)} \int_T |D^2v(x)|^p \approx \Lambda$$

provided $D^2v(x)$ is about constant in T . Therefore we reach the heuristic, but insightful, conclusion that E_T is about constant, or equivalently

A graded mesh is quasi-optimal if the local error is equidistributed. (8)

Meshes satisfying (8) have been constructed by Babuška et al [2] for *corner singularities* and $d = 2$; see also [19]. If $0 < \gamma < 1$ and the function v behaves like $v(x) \approx r(x)^\gamma$, where $r(x)$ is the distance from $x \in \Omega$ to a reentrant corner of Ω , then

$$h(x) = \Lambda^{\frac{1}{2p}} r(x)^{-\frac{1}{2}(\gamma-2)}$$

is the optimal mesh grading. This in turn implies

$$\#\mathcal{T} = \int_{\Omega} h(x)^{-2} dx \approx \Lambda^{-\frac{1}{p}} \int_0^{\text{diam}(\Omega)} r^{\gamma-1} dr \approx \Lambda^{-\frac{1}{p}}.$$

This crucial relation is valid for any $\gamma > 0$ and $p > 1$; in fact the only condition on p is that $r = 2 - 2/p > 0$, or equivalently $\text{sob}(W_p^2) > \text{sob}(H^1)$. Therefore,

$$\|\nabla(v - I_{\mathcal{T}}v)\|_{L^2(\Omega)}^2 = \sum_{T \in \mathcal{T}} E_T^2 = \Lambda^{\frac{2}{p}}(\#\mathcal{T}) \approx (\#\mathcal{T})^{-1} \tag{9}$$

gives the optimal decay rate for $d = 2, n = 1$. What this argument does not address is whether such meshes \mathcal{T} exist in general and, more importantly, whether they can actually be constructed upon bisecting the initial mesh \mathcal{T}_0 so that $\mathcal{T} \in \mathbb{T}$.

3.3. Thresholding. We now construct graded bisection meshes \mathcal{T} for $n = 1, d = 2$ that achieve the optimal decay rate $(\#\mathcal{T})^{-1/2}$ under the global regularity assumption

$$v \in W_p^2(\Omega; \mathcal{T}_0) \cap H_0^1(\Omega), \quad p > 1. \tag{10}$$

Following Binev, Dahmen, DeVore and Petrushev [6], we use a thresholding algorithm that is based on the knowledge of the element errors and on bisection. The algorithm hinges on (8): if $\delta > 0$ is a given tolerance, the element error is equidistributed, that is $E_T \approx \delta^2$, and the global error decays with maximum rate $(\#\mathcal{T})^{-1/2}$, then

$$\delta^4 \#\mathcal{T} \approx \sum_{T \in \mathcal{T}} E_T^2 = |v - I_{\mathcal{T}}v|_{H^1(\Omega)}^2 \lesssim (\#\mathcal{T})^{-1}$$

that is $\#\mathcal{T} \lesssim \delta^{-2}$. With this in mind, we impose $E_T \leq \delta^2$ as a common threshold to stop refining and expect $\#\mathcal{T} \lesssim \delta^{-2}$. The following algorithm implements this idea.

Thresholding Algorithm. Given a tolerance $\delta > 0$ and a conforming mesh \mathcal{T}_0 , the procedure THRESHOLD finds a conforming refinement $\mathcal{T} \geq \mathcal{T}_0$ of \mathcal{T}_0 by bisection such that $E_T \leq \delta^2$ for all $T \in \mathcal{T}$: let $\mathcal{T} = \mathcal{T}_0$ and

```

THRESHOLD( $\mathcal{T}, \delta$ )
while  $\mathcal{M} := \{T \in \mathcal{T} \mid E_T > \delta^2\} \neq \emptyset$ 
     $\mathcal{T} := \text{REFINE}(\mathcal{T}, \mathcal{M})$ 
end while
return( $\mathcal{T}$ )
    
```

Since $W_p^2(\Omega; \mathcal{T}_0) \cap H_0^1(\Omega) \subset C^0(\bar{\Omega})$, because $p > 1$, we can use the Lagrange interpolant and local estimate (7) with $r = \text{sob}(W_p^2) - \text{sob}(H^1) = 2 - 2/p > 0$ and $N_{\mathcal{T}}(T) = T$:

$$E_T \lesssim h_T^r \|D^2v\|_{L^p(T)}. \tag{11}$$

Hence THRESHOLD *terminates* because h_T decreases monotonically to 0 with bisection. The quality of the resulting mesh is assessed next.

Theorem 4 (Thresholding). *If v verifies (10), then the output $\mathcal{T} \in \mathbb{T}$ of THRESHOLD satisfies*

$$|v - I_{\mathcal{T}}v|_{H^1(\Omega)} \leq \delta^2(\#\mathcal{T})^{1/2}, \quad \#\mathcal{T} - \#\mathcal{T}_0 \lesssim \delta^{-2} |\Omega|^{1-1/p} \|D^2v\|_{L^p(\Omega; \mathcal{T}_0)}.$$

Proof. Let $k \geq 1$ be the number of iterations of THRESHOLD before termination. Let $\mathcal{M} = \mathcal{M}_0 \cup \dots \cup \mathcal{M}_{k-1}$ be the set of marked elements. We organize the elements in \mathcal{M} by size in such a way that allows for a counting argument. Let \mathcal{P}_j be the set of elements T of \mathcal{M} with size

$$2^{-(j+1)} \leq |T| < 2^{-j} \quad \Rightarrow \quad 2^{-(j+1)/2} \leq h_T < h_T^{-j/2}.$$

We proceed in several steps.

1 We first observe that all T 's in \mathcal{P}_j are *disjoint*. This is because if $T_1, T_2 \in \mathcal{P}_j$ and $\overset{\circ}{T}_1 \cap \overset{\circ}{T}_2 \neq \emptyset$, then one of them is contained in the other, say $T_1 \subset T_2$, due to the bisection procedure. Thus $|T_1| \leq \frac{1}{2}|T_2|$, contradicting the definition of \mathcal{P}_j . This implies

$$2^{-(j+1)} \#\mathcal{P}_j \leq |\Omega| \quad \Rightarrow \quad \#\mathcal{P}_j \leq |\Omega| 2^{j+1}. \tag{12}$$

2 In light of (11), we have for $T \in \mathcal{P}_j$

$$\delta^2 \leq E_T \lesssim 2^{-(j/2)r} \|D^2v\|_{L^p(T)}.$$

Therefore

$$\delta^{2p} \#\mathcal{P}_j \lesssim 2^{-(j/2)rp} \sum_{T \in \mathcal{P}_j} \|D^2v\|_{L^p(T)}^p \leq 2^{-(j/2)rp} \|D^2v\|_{L^p(\Omega; \mathcal{T}_0)}^p$$

whence

$$\#\mathcal{P}_j \lesssim \delta^{-2p} 2^{-(j/2)rp} \|D^2v\|_{L^p(\Omega; \mathcal{T}_0)}^p. \tag{13}$$

3 The two bounds for $\#\mathcal{P}$ in (12) and (13) are complementary. The first is good for j small whereas the second is suitable for j large (think of $\delta \ll 1$). The crossover takes place for j_0 such that

$$2^{j_0+1}|\Omega| = \delta^{-2p} 2^{-j_0(rp/2)} \|D^2v\|_{L^p(\Omega; \mathcal{T}_0)}^p \quad \Rightarrow \quad 2^{j_0} \approx \delta^{-2} \frac{\|D^2v\|_{L^p(\Omega; \mathcal{T}_0)}}{|\Omega|^{1/p}}.$$

4 We now compute

$$\#\mathcal{M} = \sum_j \#\mathcal{P}_j \lesssim \sum_{j \leq j_0} 2^j |\Omega| + \delta^{-2p} \|D^2v\|_{L^p(\Omega; \mathcal{T}_0)}^p \sum_{j > j_0} (2^{-rp/2})^j.$$

Since

$$\sum_{j \leq j_0} 2^j \approx 2^{j_0}, \quad \sum_{j > j_0} (2^{-rp/2})^j \lesssim 2^{-(rp/2)j_0} = 2^{-(p-1)j_0}$$

we can write

$$\#\mathcal{M} \lesssim (\delta^{-2} + \delta^{-2p} \delta^{2(p-1)}) |\Omega|^{1-1/p} \|D^2 v\|_{L^p(\Omega; \mathcal{T}_0)} \approx \delta^{-2} |\Omega|^{1-1/p} \|D^2 v\|_{L^p(\Omega; \mathcal{T}_0)}.$$

We finally apply Theorem 2 to arrive at

$$\#\mathcal{T} - \#\mathcal{T}_0 \lesssim \#\mathcal{M} \lesssim \delta^{-2} |\Omega|^{1-1/p} \|D^2 v\|_{L^p(\Omega; \mathcal{T}_0)}.$$

5 It remains to estimate the energy error. We have, upon termination of THRESHOLD, that $E_T \leq \delta^2$ for all $T \in \mathcal{T}$. Then

$$|v - I_{\mathcal{T}} v|_{H^1(\Omega)}^2 = \sum_{T \in \mathcal{T}} E_T^2 \leq \delta^4 \#\mathcal{T}.$$

This concludes the Theorem. □

Upon relating the threshold δ and the number of elements N , we obtain a convergence rate. In particular, this implies (4): $\sigma_N(v) \lesssim \|D^2 v\|_{L^p(\Omega; \mathcal{T}_0)} N^{-1/2}$ for all $N \geq \#\mathcal{T}_0$.

Corollary 5 (Convergence rate). *Let v satisfy (10). Then for $N > \#\mathcal{T}_0$ integer there exists $\mathcal{T} \in \mathbb{T}$ such that*

$$|v - I_{\mathcal{T}} v|_{H^1(\Omega)} \lesssim |\Omega|^{1-1/p} \|D^2 v\|_{L^p(\Omega; \mathcal{T}_0)} N^{-1/2}, \quad \#\mathcal{T} - \#\mathcal{T}_0 \lesssim N.$$

Proof. Choose $\delta^2 = |\Omega|^{1-1/p} \|D^2 v\|_{L^p(\Omega; \mathcal{T}_0)} N^{-1}$ in Theorem 4. Then, there exists $\mathcal{T} \in \mathbb{T}$ such that $\#\mathcal{T} - \#\mathcal{T}_0 \lesssim N$ and

$$|v - I_{\mathcal{T}} v|_{H^1(\Omega)} \lesssim |\Omega|^{1-1/p} \|D^2 v\|_{L^p(\Omega; \mathcal{T}_0)} N^{-1} (\#\mathcal{T})^{1/2} \lesssim |\Omega|^{1-1/p} \|D^2 v\|_{L^p(\Omega; \mathcal{T}_0)} N^{-1/2},$$

because $\#\mathcal{T} \lesssim N$. This finishes the Corollary. □

Remark 6 (Case $p < 1$). We consider now polynomial degree $n \geq 1$. The integrability p corresponding to differentiability $n + 1$ results from equating Sobolev numbers:

$$n + 1 - \frac{d}{p} = \text{sob}(H^1) = 1 - \frac{d}{2} \quad \Rightarrow \quad p = \frac{2d}{2n + d}.$$

Depending on $d \geq 2$ and $n \geq 1$, this may lead to $0 < p < 1$, in which case $W_p^{n+1}(\Omega)$ is to be replaced by the Besov space $B_{p,p}^{n+1}(\Omega)$ [17]. The argument of Theorem 4 works provided we replace (11) by a modulus of regularity [6].

Remark 7 (Isotropic elements). Corollary 5 shows that isotropic graded meshes can always deal with geometric singularities for $d = 2$. This is no longer true for $d > 2$ due to *edge singularities*: if $d = 3$ and $v(x) \approx r(x)^\gamma$ near an edge, then $n = 1$ requires $\gamma > \frac{1}{3}$ whereas $n = 2$ needs $\gamma > \frac{2}{3}$. The latter corresponds to a dihedral angle $\omega < \frac{3\pi}{2}$.

4. Adaptive Finite Element Methods (AFEM)

We now present the four basic modules of AFEM for (1) and discuss their main properties.

4.1. Modules of AFEM. They are SOLVE, ESTIMATE, MARK, and REFINE.

Module SOLVE. If $\mathcal{T} \in \mathbb{T}$ is a conforming refinement of \mathcal{T}_0 and $\mathbb{V} = \mathbb{V}(\mathcal{T})$ is the finite element space of C^0 piecewise polynomials of degree $\leq n$, then

$$U = \text{SOLVE}(\mathcal{T})$$

determines the Galerkin solution *exactly*, namely,

$$U \in \mathbb{V} : \int_{\Omega} \mathbf{A} \nabla U \cdot \nabla V = \int_{\Omega} f V \quad \text{for all } V \in \mathbb{V}. \quad (14)$$

Module ESTIMATE. Given a conforming mesh $\mathcal{T} \in \mathbb{T}$ and the Galerkin solution $U \in \mathbb{V}(\mathcal{T})$, the output $\{\mathcal{E}_{\mathcal{T}}(U, T)\}_{T \in \mathcal{T}}$ of

$$\{\mathcal{E}_{\mathcal{T}}(U, T)\}_{T \in \mathcal{T}} = \text{ESTIMATE}(U, \mathcal{T})$$

are the element indicators defined as follows: for any $V \in \mathbb{V}$

$$\mathcal{E}_{\mathcal{T}}^2(V, T) = h_T^2 \|r(V)\|_T^2 + h_T \|j(V)\|_{\partial T}^2 \quad \text{for all } T \in \mathcal{T}, \quad (15)$$

where the *interior* and *jump residuals* are given by

$$\begin{aligned} r(V)|_T &= f + \text{div}(\mathbf{A} \nabla V) && \text{for all } T \in \mathcal{T} \\ j(V)|_S &= [\mathbf{A} \nabla V] \cdot \nu|_S && \text{for all } S \in \mathcal{S} \quad (\text{internal sides of } \mathcal{T}), \end{aligned}$$

and $j(V)|_S = 0$ for boundary sides $S \in \mathcal{S}$. We denote $\mathcal{E}_{\mathcal{T}}^2(V, \mathcal{P}) = \sum_{T \in \mathcal{P}} \mathcal{E}_{\mathcal{T}}^2(V, T)$ for any subset \mathcal{P} of \mathcal{T} and $\mathcal{E}_{\mathcal{T}}(V) = \mathcal{E}_{\mathcal{T}}(V, \mathcal{T})$.

Module MARK. Given $\mathcal{T} \in \mathbb{T}$, the Galerkin solution $U \in \mathbb{V}(\mathcal{T})$, and element indicators $\{\mathcal{E}_{\mathcal{T}}(U, T)\}_{T \in \mathcal{T}}$, the module MARK selects elements for refinement using *Dörfler Marking* (or bulk chasing) [18], i. e., using a fixed parameter $\theta \in (0, 1]$ the output \mathcal{M} of

$$\mathcal{M} = \text{MARK}(\{\mathcal{E}_{\mathcal{T}}(U, T)\}_{T \in \mathcal{T}}, \mathcal{T})$$

satisfies

$$\mathcal{E}_{\mathcal{T}}(U, \mathcal{M}) \geq \theta \mathcal{E}_{\mathcal{T}}(U, \mathcal{T}). \quad (16)$$

This marking guarantees that \mathcal{M} contains a substantial part of the total (or bulk), thus its name. The choice of \mathcal{M} does not have to be minimal at this

stage, that is, the marked elements $T \in \mathcal{M}$ do not necessarily must be those with largest indicators.

Module REFINE. Let $b \in \mathbb{N}$ be the number of desired bisections per marked element. Given $\mathcal{T} \in \mathbb{T}$ and a subset \mathcal{M} of marked elements, the output $\mathcal{T}_* \in \mathbb{T}$ of

$$\mathcal{T}_* = \text{REFINE}(\mathcal{T}, \mathcal{M})$$

is the smallest refinement \mathcal{T}_* of \mathcal{T} so that all elements of \mathcal{M} are at least bisected b times. Therefore, the piecewise constant meshsize functions satisfy $h_{\mathcal{T}_*} \leq h_{\mathcal{T}}$ and the strict reduction property

$$h_{\mathcal{T}_*}|_T \leq 2^{-b/d} h_{\mathcal{T}}|_T \quad \text{for all } T \in \mathcal{M}. \tag{17}$$

We finally let $\mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_*}$ be the subset of refined elements of \mathcal{T} and note that $\mathcal{M} \subset \mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_*}$.

AFEM. Given an initial grid \mathcal{T}_0 , set $k = 0$ and iterate

$$\begin{aligned} U_k &= \text{SOLVE}(\mathcal{T}_k); \\ \{\mathcal{E}_k(U_k, T)\}_{T \in \mathcal{T}_k} &= \text{ESTIMATE}(U_k, \mathcal{T}_k); \\ \mathcal{M}_k &= \text{MARK}(\{\mathcal{E}_k(U_k, T)\}_{T \in \mathcal{T}_k}, \mathcal{T}_k); \\ \mathcal{T}_{k+1} &= \text{REFINE}(\mathcal{T}_k, \mathcal{M}_k); \quad k \leftarrow k + 1. \end{aligned}$$

4.2. Basic Properties of AFEM. We next follow Cascón, Kreuzer, Nochetto, and Siebert [10] and summarize some basic properties of AFEM that emanate from the symmetry of the differential operator (i.e. of \mathbf{A}) and features of the modules. In doing this, any explicit constant or hidden constant in \lesssim will only depend on the uniform shape-regularity of \mathbb{T} , the dimension d , the polynomial degree n , and the (global) eigenvalues of \mathbf{A} , but not on a specific grid $\mathcal{T} \in \mathbb{T}$, except if explicitly stated. Furthermore, u will always be the weak solution of (1).

The following property relies on the fact that the underlying bilinear form is coercive and symmetric, and so induces a scalar product in \mathbb{V} equivalent to the H_0^1 -scalar product.

Lemma 8 (Pythagoras). *Let $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$ be such that $\mathcal{T} \leq \mathcal{T}_*$. The corresponding Galerkin solutions $U \in \mathbb{V}(\mathcal{T})$ and $U_* \in \mathbb{V}(\mathcal{T}_*)$ satisfy the following orthogonality property*

$$\|u - U\|_{\Omega}^2 = \|u - U_*\|_{\Omega}^2 + \|U_* - U\|_{\Omega}^2. \tag{18}$$

Property (18) is valid for (1) for the energy norm exclusively. This restricts the subsequent analysis to the energy norm, or equivalent norms, but does not extend to other, perhaps more practical, norms such as the maximum norm. This is an important open problem and a serious limitation of this theory.

We now continue with the concept of *oscillation*. We denote by $\text{osc}_{\mathcal{T}}(V, T)$ the *element oscillation* for any $V \in \mathbb{V}$

$$\text{osc}_{\mathcal{T}}(V, T) = \|h(r(V) - \overline{r(V)})\|_{L^2(T)} + \|h^{1/2}(j(V) - \overline{j(V)})\|_{L^2(\partial T \cap \Omega)}, \quad (19)$$

where $\overline{r(V)} = P_{2n-2}r(V)$ and $\overline{j(V)} = P_{2n-1}j(V)$ stand for L^2 -projections of the residuals $r(V)$ and $j(V)$ onto the polynomials $\mathbb{P}_{2n-2}(T)$ and $\mathbb{P}_{2n-1}(S)$ defined on the element T or side $S \subset \partial T$, respectively. For variable \mathbf{A} , $\text{osc}_{\mathcal{T}}(V, T)$ depends on the discrete function $V \in \mathbb{V}$, and its study is more involved than for piecewise constant \mathbf{A} . In the latter case, $\text{osc}_{\mathcal{T}}(V, T) = \|h(f - \bar{f})\|_{L^2(T)}$ is called *data oscillation* [25, 26].

Proposition 9 (A posteriori error estimates). *There exist constants $0 < C_2 \leq C_1$, such that for any $\mathcal{T} \in \mathbb{T}$ and the corresponding Galerkin solution $U \in \mathbb{V}(\mathcal{T})$ there holds*

$$\|u - U\|_{\Omega}^2 \leq C_1 \mathcal{E}_{\mathcal{T}}^2(U) \quad (20a)$$

$$C_2 \mathcal{E}_{\mathcal{T}}^2(U) \leq \|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U). \quad (20b)$$

This Proposition is essentially due to Babuška and Miller [3]; see also [1, 8, 28, 35]. The constants C_1 and C_2 depend on the smallest and largest global eigenvalues of \mathbf{A} as well as interpolation estimates. The definitions of $\overline{r(V)}$ and $\overline{j(V)}$, as well as the lower bound (20b), are immaterial for deriving a contraction property of §5 but are important for proving convergence rates in §6; we refer to [28] for a discussion of oscillation.

One serious difficulty in dealing with AFEM is that one has access to the energy error $\|u - U\|_{\Omega}$ only through the estimator $\mathcal{E}_{\mathcal{T}}(U)$. The latter, however, fails to be monotone because it depends on the discrete solution $U \in \mathbb{V}(\mathcal{T})$ that changes with the mesh. This is tackled in the next two lemmas [10, 27].

Lemma 10 (Reduction of $\mathcal{E}_{\mathcal{T}}(V)$ with respect to \mathcal{T}). *If $\lambda = 1 - 2^{-b/d}$, then*

$$\mathcal{E}_{\mathcal{T}_*}^2(V, \mathcal{T}_*) \leq \mathcal{E}_{\mathcal{T}}^2(V, \mathcal{T}) - \lambda \mathcal{E}_{\mathcal{T}}^2(V, \mathcal{M}) \quad \text{for all } V \in \mathbb{V}(\mathcal{T}). \quad (21)$$

Lemma 11 (Lipschitz property of $\mathcal{E}_{\mathcal{T}}(V)$ with respect to V). *Let $\text{div } \mathbf{A}$ be the divergence of \mathbf{A} computed by rows, and $\eta_{\mathcal{T}}(\mathbf{A}) := \max_{\mathcal{T}} (h_T \|\text{div } \mathbf{A}\|_{L^\infty(T)} + \|\mathbf{A}\|_{L^\infty(T)})$. Then the following estimate is valid*

$$|\mathcal{E}_{\mathcal{T}}(V) - \mathcal{E}_{\mathcal{T}}(W)| \lesssim \eta_{\mathcal{T}}(\mathbf{A}) \|V - W\|_{\Omega} \quad \text{for all } V, W \in \mathbb{V}(\mathcal{T}).$$

Upon combining Lemmas 10 and 11 we obtain the following crucial property.

Proposition 12 (Estimator reduction). *Given $\mathcal{T} \in \mathbb{T}$ and a subset $\mathcal{M} \subset \mathcal{T}$ of marked elements, let $\mathcal{T}_* = \text{REFINE}(\mathcal{T}, \mathcal{M})$. Then there exists a constant $\Lambda > 0$, such that for all $V \in \mathbb{V}(\mathcal{T})$, $V_* \in \mathbb{V}_*(\mathcal{T}_*)$ and any $\delta > 0$ we have*

$$\mathcal{E}_{\mathcal{T}_*}^2(V_*, \mathcal{T}_*) \leq (1 + \delta)(\mathcal{E}_{\mathcal{T}}^2(V, \mathcal{T}) - \lambda \mathcal{E}_{\mathcal{T}}^2(V, \mathcal{M})) + (1 + \delta^{-1}) \Lambda \eta_{\mathcal{T}}^2(\mathbf{A}) \|V_* - V\|_{\Omega}^2.$$

5. Contraction Property of AFEM

A key question to ask is what is (are) the quantity(ies) that AFEM may contract. In light of (18), an obvious candidate is the energy error $\|u - U_k\|_\Omega$; see Dörfler [18]. We first show in §5.1, in the simplest scenario of piecewise constant data \mathbf{A} and f , that this is in fact the case provided an interior node property holds. However, the energy error may not contract in general unless REFINE enforces several levels of refinement. We discuss this in §5.2, and present an approach that eliminates the interior node property at the expense of a more complicated contractive quantity, the quasi-error; see Theorem 16.

5.1. Piecewise Constant Data. We now assume that both f and \mathbf{A} are piecewise constant in the initial mesh \mathcal{T}_0 , so that $\text{osc}_k(U_k) = 0$ for all $k \geq 0$. The following property was introduced by Morin, Nochetto, and Siebert [25].

Definition 13 (Interior node property). *The refinement $\mathcal{T}_{k+1} \geq \mathcal{T}_k$ satisfies an interior node property with respect to \mathcal{T}_k if each element $T \in \mathcal{M}_k$ contains at least one node of \mathcal{T}_{k+1} in the interiors of T and of each side of T .*

This property is valid upon enforcing a fixed number b_* of bisections ($b_* = 3, 6$ for $d = 2, 3$). An immediate consequence of this property, proved in [25, 26], is the following *discrete* lower a posteriori bound:

$$C_2 \mathcal{E}_k^2(U_k, \mathcal{M}_k) \leq \|U_k - U_{k+1}\|_\Omega^2 + \text{osc}_k^2(U_k). \tag{22}$$

Lemma 14 (Contraction property for piecewise constant data). *If \mathcal{T}_{k+1} satisfies an interior node property with respect to \mathcal{T}_k and $\text{osc}_k(U_k) = 0$, then for $\alpha := (1 - \theta^2 \frac{C_2}{C_1})^{1/2} < 1$*

$$\|u - U_{k+1}\|_\Omega \leq \alpha \|u - U_k\|_\Omega, \tag{23}$$

where $0 < \theta < 1$ is the parameter in (16) and $C_1 \geq C_2$ are the constants in (20).

Proof. For convenience, we use the notation

$$e_k = \|u - U_k\|_\Omega, \quad E_k = \|U_{k+1} - U_k\|_\Omega, \quad \mathcal{E}_k = \mathcal{E}_k(U_k, \mathcal{T}_k), \quad \mathcal{E}_k(\mathcal{M}_k) = \mathcal{E}_k(U_k, \mathcal{M}_k).$$

The key idea is to use the Pythagoras equality (18), namely $e_{k+1}^2 = e_k^2 - E_k^2$, and show that E_k is a significant portion of e_k . Since (22) together with $\text{osc}_k(U_k) = 0$ imply $C_2 \mathcal{E}_k^2(\mathcal{M}_k) \leq E_k^2$, applying Dörfler marking (16) and the upper bound (20a), we deduce

$$E_k^2 \geq C_2 \theta^2 \mathcal{E}_k^2 \geq \frac{C_2}{C_1} \theta^2 e_k^2.$$

This is the desired property of E_k and leads to (23). □

We wonder whether or not the interior node property is necessary for (23). We present an example, introduced in [25, 26] to justify such a property for constant data and $n = 1$.

Example 15 (Lack of strict monotonicity). Let $\Omega = (0, 1)^2$, $\mathbf{A} = \mathbf{I}$, $f = 1$ (constant data), and consider the following sequences of meshes depicted in Figure 7. If ϕ_0 denotes the basis function associated with the only interior node of the initial mesh \mathcal{T}_0 , then $U_0 = U_1 = \frac{1}{12} \phi_0$ and $U_2 \neq U_1$.

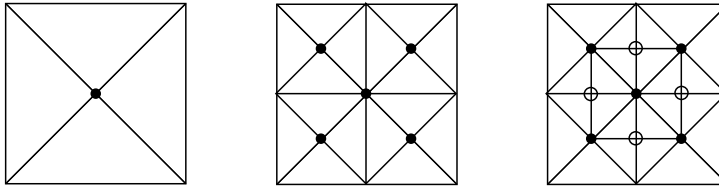


Figure 7. Grids \mathcal{T}_0 , \mathcal{T}_1 , and \mathcal{T}_2 of Example 15. The mesh \mathcal{T}_1 has nodes in the middle of sides of \mathcal{T}_0 , but only \mathcal{T}_2 has nodes in the interior of elements of \mathcal{T}_0 . Hence, \mathcal{T}_2 satisfies the interior node property of Definition 13 with respect to \mathcal{T}_0 whereas \mathcal{T}_1 does not.

The mesh $\mathcal{T}_1 \geq \mathcal{T}_0$ is produced by a standard 2-step bisection ($b = 2$) in $2d$. Since $U_0 = U_1$ we conclude that the energy error does not change $\|u - U_0\|_\Omega = \|u - U_1\|_\Omega$ between two consecutive steps of AFEM for $b = d = 2$. This is no longer true provided an interior node in each marked element is created, because then Lemma 14 holds.

5.2. General Data. If $\text{osc}_k(U_k) \neq 0$, then the contraction property of AFEM becomes trickier because the energy error and estimator are no longer equivalent regardless of the interior node property. The first question to ask is what quantity replaces the energy error in the analysis. We explore this next and remove the interior node property.

Heuristics. According to (18), the energy error is monotone $\|u - U_{k+1}\|_\Omega \leq \|u - U_k\|_\Omega$, but the previous Example shows that strict inequality may fail. However, if $U_{k+1} = U_k$, estimate (21) reveals a strict estimator reduction $\mathcal{E}_{k+1}(U_k) < \mathcal{E}_k(U_k)$. We thus expect that, for a suitable scaling factor $\gamma > 0$, the so-called *quasi error*

$$\|u - U_k\|_\Omega^2 + \gamma \mathcal{E}_k^2(U_k) \tag{24}$$

may be contractive. This heuristics illustrates a distinct aspect of AFEM theory, the interplay between continuous quantities such the energy error $\|u - U_k\|_\Omega$ and discrete ones such as the estimator $\mathcal{E}_k(U_k)$: no one alone has the requisite properties to yield a contraction between consecutive adaptive steps.

Theorem 16 (Contraction property). *Let $\theta \in (0, 1]$ be the Dörfler Marking parameter, and $\{\mathcal{T}_k, \mathbb{V}_k, U_k\}_{k=0}^\infty$ be a sequence of conforming meshes, finite element spaces and discrete solutions created by AFEM for the model problem (1). Then there exist constants $\gamma > 0$ and $0 < \alpha < 1$, additionally depending on the number $b \geq 1$ of bisections and θ , such that for all $k \geq 0$*

$$\|u - U_{k+1}\|_\Omega^2 + \gamma \mathcal{E}_{k+1}^2(U_{k+1}) \leq \alpha^2 \left(\|u - U_k\|_\Omega^2 + \gamma \mathcal{E}_k^2(U_k) \right). \tag{25}$$

Proof. We split the proof into four steps and use the notation in Lemma 14.

1 The error orthogonality (18) reads

$$e_{k+1}^2 = e_k^2 - E_k^2. \tag{26}$$

Employing Proposition 12 with $\mathcal{T} = \mathcal{T}_k$, $\mathcal{T}_* = \mathcal{T}_{k+1}$, $V = U_k$ and $V_* = U_{k+1}$ gives

$$\mathcal{E}_{k+1}^2 \leq (1 + \delta)(\mathcal{E}_k^2 - \lambda \mathcal{E}_k^2(\mathcal{M}_k)) + (1 + \delta^{-1}) \Lambda_0 E_k^2, \tag{27}$$

where $\Lambda_0 = \Lambda \eta_{\mathcal{T}_0}^2(\mathbf{A}) \geq \Lambda \eta_{\mathcal{T}_k}^2(\mathbf{A})$. After multiplying (27) by $\gamma > 0$, to be determined later, we add (26) and (27) to obtain

$$e_{k+1}^2 + \gamma \mathcal{E}_{k+1}^2 \leq e_k^2 + (\gamma(1 + \delta^{-1}) \Lambda_0 - 1) E_k^2 + \gamma(1 + \delta) (\mathcal{E}_k^2 - \lambda \mathcal{E}_k^2(\mathcal{M}_k)).$$

2 We now choose the parameters δ, γ , the former so that

$$(1 + \delta)(1 - \lambda\theta^2) = 1 - \frac{\lambda\theta^2}{2},$$

and the latter to verify

$$\gamma(1 + \delta^{-1}) \Lambda_0 = 1.$$

Note that this choice of γ yields

$$e_{k+1}^2 + \gamma \mathcal{E}_{k+1}^2 \leq e_k^2 + \gamma(1 + \delta) (\mathcal{E}_k^2 - \lambda \mathcal{E}_k^2(\mathcal{M}_k)).$$

3 We next employ Dörfler Marking, namely $\mathcal{E}_k(\mathcal{M}_k) \geq \theta \mathcal{E}_k$, to deduce

$$e_{k+1}^2 + \gamma \mathcal{E}_{k+1}^2 \leq e_k^2 + \gamma(1 + \delta)(1 - \lambda\theta^2) \mathcal{E}_k^2$$

which, in conjunction with the choice of δ , gives

$$e_{k+1}^2 + \gamma \mathcal{E}_{k+1}^2 \leq e_k^2 + \gamma \left(1 - \frac{\lambda\theta^2}{2}\right) \mathcal{E}_k^2 = e_k^2 - \frac{\gamma\lambda\theta^2}{4} \mathcal{E}_k^2 + \gamma \left(1 - \frac{\lambda\theta^2}{4}\right) \mathcal{E}_k^2.$$

4 Finally, the upper bound (20a), namely $e_k^2 \leq C_1 \mathcal{E}_k^2$, implies that

$$e_{k+1}^2 + \gamma \mathcal{E}_{k+1}^2 \leq \left(1 - \frac{\gamma\lambda\theta^2}{4C_1}\right) e_k^2 + \gamma \left(1 - \frac{\lambda\theta^2}{4}\right) \mathcal{E}_k^2.$$

This in turn leads to

$$e_{k+1}^2 + \gamma \mathcal{E}_{k+1}^2 \leq \alpha^2 (e_k^2 + \gamma \mathcal{E}_k^2),$$

with $\alpha^2 := \max \left\{1 - \frac{\gamma\lambda\theta^2}{4C_1}, 1 - \frac{\lambda\theta^2}{4}\right\} < 1$, and thus concludes the theorem. \square

Remark 17 (Basic ingredients). This proof solely uses Dörfler marking, Pythagoras identity (18), the a posteriori upper bound (20a), and the estimator reduction property (Proposition 12). The proof does not use the lower bound (20b).

Remark 18 (Separate marking). MARK is driven by \mathcal{E}_k exclusively, as it happens in all practical AFEM. Previous proofs in [14, 23, 25, 26] require separate marking by estimator and oscillation. It is shown in [10] that separate marking may lead to suboptimal convergence rates. On the other hand, we will prove in §6 that the present AFEM yields quasi-optimal convergence rates.

6. Convergence Rates of AFEM

A crucial insight for the simplest scenario, the Laplacian and piecewise constant forcing f , is due to Stevenson [33]:

any marking strategy that reduces the energy error relative to the current value must contain a substantial portion of $\mathcal{E}_{\mathcal{T}}(U)$, and so it can be related to Dörfler Marking. (28)

This allows one to compare meshes produced by AFEM with optimal ones and to conclude a quasi-optimal error decay. We discuss this issue in §6.3. However, this is not enough to handle the model problem (1) with variable data \mathbf{A} and f .

The objective of this section is to study (1) for general data \mathbf{A} and f . This study hinges on the total error and its relation with the quasi-error, which is contracted by AFEM. This approach allows us to improve upon and extend Stevenson [33] to variable data. In doing so, we follow closely Cascón, Kreuzer, Nochetto, and Siebert [10]. The present theory, however, does not extend to noncoercive problems and marking strategies other than Dörfler's. These remain important open questions.

As in §5, u will always be the weak solution of (1) and, except when stated otherwise, any explicit constant or hidden constant in \lesssim may depend on the uniform shape-regularity of \mathbb{T} , the dimension d , the polynomial degree n , the (global) eigenvalues of \mathbf{A} , and the oscillation $\text{osc}_{\mathcal{T}_0}(\mathbf{A})$ of \mathbf{A} on the initial mesh \mathcal{T}_0 , but not on a specific grid $\mathcal{T} \in \mathbb{T}$.

6.1. The Total Error. We first present the concept of *total error* for the Galerkin function $U \in \mathbb{V}(\mathcal{T})$, introduced by Mekchay and Nochetto [23],

$$\|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U), \quad (29)$$

and next assert its equivalence to the quasi error (24). In fact, in view of the upper and lower a posteriori error bounds (20), and $\text{osc}_{\mathcal{T}}^2(U) \leq \mathcal{E}_{\mathcal{T}}^2(U)$, we have

$$C_2 \mathcal{E}_{\mathcal{T}}^2(U) \leq \|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U) \leq \|u - U\|_{\Omega}^2 + \mathcal{E}_{\mathcal{T}}^2(U) \leq (1 + C_1) \mathcal{E}_{\mathcal{T}}^2(U),$$

whence

$$\mathcal{E}_{\mathcal{T}}^2(U) \approx \|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U). \quad (30)$$

Since AFEM selects elements for refinement based on information extracted exclusively from the error indicators $\{\mathcal{E}_{\mathcal{T}}(U, T)\}_{T \in \mathcal{T}}$, we realize that the decay

rate of AFEM must be characterized by the total error. Moreover, on invoking the upper bound (20a) again, we also see that the total error is equivalent to the quasi error

$$\|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U) \approx \|u - U\|_{\Omega}^2 + \mathcal{E}_{\mathcal{T}}^2(U).$$

The latter is the quantity being strictly reduced by AFEM (Theorem 16). Finally, the total error satisfies the following Cea’s type-lemma, or equivalently AFEM is quasi-optimal regarding the total error [10].

Lemma 19 (Quasi-optimality of total error). *Let $\Lambda_1 = 2\Lambda$ with Λ the constant in Proposition 12, and let $C_3 := \Lambda_1 \text{osc}_{\mathcal{T}_0}^2(\mathbf{A})$ and $\Lambda_2 := \max\{2, 1 + C_3\}$. Then, for any $\mathcal{T} \in \mathbb{T}$ and the corresponding Galerkin solution $U \in \mathbb{V}(\mathcal{T})$, there holds*

$$\|u - U\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(U) \leq \Lambda_2 \inf_{V \in \mathbb{V}(\mathcal{T})} \left(\|u - V\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(V) \right).$$

6.2. Approximation Classes. In view of (30) and Lemma 19, the definition of approximation class \mathbb{A}_s depends on the triple (u, f, \mathbf{A}) , not just u , and hinges on the concept of best total error for meshes \mathcal{T} with N elements more than \mathcal{T}_0 , namely $\mathcal{T} \in \mathbb{T}_N$:

$$\sigma_N(u, f, \mathbf{A}) := \inf_{\mathcal{T} \in \mathbb{T}_N} \inf_{V \in \mathbb{V}(\mathcal{T})} \left(\|u - V\|_{\Omega}^2 + \text{osc}_{\mathcal{T}}^2(V) \right)^{1/2}.$$

We say that $(u, f, \mathbf{A}) \in \mathbb{A}_s$ for $s > 0$ if and only if $\sigma_N(u, f, \mathbf{A}) \lesssim N^{-s}$, and denote $|u, f, \mathbf{A}|_s := \sup_{N > 0} (N^s \sigma_N(u, f, \mathbf{A}))$. We point out the upper bound $s \leq n/d$ for polynomial degree $n \geq 1$; this can be seen with full regularity $H^{n+1}(\Omega)$ and uniform refinement. Note that if $(u, f, \mathbf{A}) \in \mathbb{A}_s$ then for all $\varepsilon > 0$ there exist $\mathcal{T}_{\varepsilon} \geq \mathcal{T}_0$ conforming and $V_{\varepsilon} \in \mathbb{V}(\mathcal{T}_{\varepsilon})$ such that

$$\|v - V_{\varepsilon}\|_{\Omega}^2 + \text{osc}_{\mathcal{T}_{\varepsilon}}^2(V_{\varepsilon}) \leq \varepsilon^2 \quad \text{and} \quad \#\mathcal{T}_{\varepsilon} - \#\mathcal{T}_0 \leq |v, f, \mathbf{A}|_s^{1/s} \varepsilon^{-1/s}. \quad (31)$$

Mesh Overlay. For the subsequent discussion it will be convenient to merge (or superpose) two conforming meshes $\mathcal{T}_1, \mathcal{T}_2 \in \mathbb{T}$, thereby giving rise to the so-called *overlay* $\mathcal{T}_1 \oplus \mathcal{T}_2$. This operation corresponds to the union in the sense of trees [10, 33]. We next bound the cardinality of $\mathcal{T}_1 \oplus \mathcal{T}_2$ in terms of that of \mathcal{T}_1 and \mathcal{T}_2 ; see [10, 33].

Lemma 20 (Overlay). *The overlay $\mathcal{T} = \mathcal{T}_1 \oplus \mathcal{T}_2$ is conforming and*

$$\#\mathcal{T} \leq \#\mathcal{T}_1 + \#\mathcal{T}_2 - \#\mathcal{T}_0. \quad (32)$$

Discussion of \mathbb{A}_s . We now would like to show a few examples of membership in \mathbb{A}_s and highlight some important open questions. We first investigate the class \mathbb{A}_s for \mathbf{A} piecewise polynomial of degree $\leq n$ over \mathcal{T}_0 . In this simplified scenario, the oscillation $\text{osc}_{\mathcal{T}}(U)$ of (19) reduces to *data oscillation* $\text{osc}_{\mathcal{T}}(f) :=$

$\|h(f - P_{2n-2}f)\|_{L^2(\Omega)}$. We then have the following characterization of \mathbb{A}_s in terms of the approximation class \mathcal{A}_s and [5, 6, 33]:

$$\mathcal{B}_s := \left\{ g \in L^2(\Omega) \mid |g|_{\mathcal{B}_s} := \sup_{N>0} \left(N^s \inf_{\mathcal{T} \in \mathbb{T}_N} \text{osc}_{\mathcal{T}}(g) \right) < \infty \right\}.$$

Lemma 21 (Equivalence of classes). *Let \mathbf{A} be piecewise polynomial of degree $\leq n$ over \mathcal{T}_0 . Then $(u, f, \mathbf{A}) \in \mathbb{A}_s$ if and only if $(u, f) \in \mathcal{A}_s \times \mathcal{B}_s$ and*

$$|u, f, \mathbf{A}|_s \approx |u|_{\mathcal{A}_s} + |f|_{\mathcal{B}_s}. \tag{33}$$

Corollary 22 (Membership in $\mathbb{A}_{1/2}$ with piecewise constant \mathbf{A}). *Let $d = 2$, $n = 1$, $p > 1$. If $f \in L^2(\Omega)$, \mathbf{A} is piecewise constant over \mathcal{T}_0 , and the solution $u \in W_p^2(\Omega; \mathcal{T}_0) \cap H_0^1(\Omega)$ of (1) is piecewise W_p^2 over the initial grid \mathcal{T}_0 , then $(u, f, \mathbf{A}) \in \mathbb{A}_{1/2}$ and*

$$|u, f, \mathbf{A}|_{1/2} \lesssim \|D^2u\|_{L^p(\Omega; \mathcal{T}_0)} + \|f\|_{L^2(\Omega)}.$$

Proof. Since $f \in L^2(\Omega)$, we realize that for all quasi-uniform refinements $\mathcal{T} \in \mathbb{T}$

$$\text{osc}_{\mathcal{T}}(f) = \|h(f - P_0f)\|_{L^2(\Omega)} \leq h_{\max}(\mathcal{T})\|f\|_{L^2(\Omega)} \lesssim (\#\mathcal{T})^{-1/2}\|f\|_{L^2(\Omega)}.$$

This implies $f \in \mathcal{B}_{1/2}$ with $|f|_{\mathcal{B}_{1/2}} \lesssim \|f\|_{L^2(\Omega)}$. On the other hand, for $u \in W_p^2(\Omega; \mathcal{T}_0)$ we learn from Corollary 5 that $u \in \mathcal{A}_{1/2}$ and $|u|_{\mathcal{A}_{1/2}} \lesssim \|D^2u\|_{L^2(\Omega; \mathcal{T}_0)}$. The assertion then follows from Lemma 21. \square

Corollary 23 (Membership in $\mathbb{A}_{1/2}$ with variable \mathbf{A}). *Let $d = 2$, $n = 1$, $p > 1$. If $f \in L^2(\Omega)$, $\mathbf{A} \in W_\infty^1(\Omega, \mathcal{T}_0)$ is piecewise Lipschitz over \mathcal{T}_0 , and $u \in W_p^2(\Omega; \mathcal{T}_0) \cap H_0^1(\Omega)$ is piecewise W_p^2 over \mathcal{T}_0 , then $(u, f, \mathbf{A}) \in \mathbb{A}_{1/2}$ and*

$$|u, f, \mathbf{A}|_{1/2} \lesssim \|D^2u\|_{L^p(\Omega; \mathcal{T}_0)} + \|f\|_{L^2(\Omega)} + \|\mathbf{A}\|_{W_\infty^1(\Omega; \mathcal{T}_0)}.$$

6.3. Quasi-Optimal Cardinality: Vanishing Oscillation. In this section we follow the ideas of Stevenson [33] for the simplest scenario with vanishing oscillation $\text{osc}_{\mathcal{T}}(U) = 0$, and thereby explore the insight (28). We recall that in this case the a posteriori error estimates (20) become

$$C_2 \mathcal{E}_{\mathcal{T}}^2(U) \leq \|u - U\|_{\Omega}^2 \leq C_1 \mathcal{E}_{\mathcal{T}}^2(U). \tag{34}$$

It is then evident that the ratio $C_2/C_1 \leq 1$, between the *reliability* constant C_1 and the *efficiency* constant C_2 , is a quality measure of the estimator $\mathcal{E}_{\mathcal{T}}(U)$: the closer to 1 the better! This ratio is usually closer to 1 for non-residual estimators for which this theory extends [12, 22].

Assumptions for Optimal Decay Rate. The following are further restrictions on AFEM to achieve optimal error decay, as predicted by the approximation class \mathcal{A}_s .

Assumption 24 (Marking parameter: vanishing oscillation). *The parameter θ of Dörfler marking satisfies $\theta \in (0, \theta_*)$ with $\theta_* := \sqrt{C_2/C_1}$.*

Assumption 25 (Cardinality of \mathcal{M}). *MARK selects a set \mathcal{M} with minimal cardinality.*

Assumption 26 (Initial labeling). *The labeling of the initial mesh \mathcal{T}_0 satisfies (6) for $d = 2$ [24, 5] or its multidimensional counterpart for $d > 2$ [33, 28].*

A few comments about these assumptions are now in order.

Remark 27 (Threshold $\theta_* < 1$). It is reasonable to be cautious in making marking decisions if the constants C_1 and C_2 are very disparate, and thus the ratio C_2/C_1 is far from 1. This justifies the upper bound $\theta_* < 1$ in Assumption 24.

Remark 28 (Minimal \mathcal{M}). According to the equidistribution principle (8) and the local lower bound $C_2 \mathcal{E}_{\mathcal{T}}(U, T) \leq \|u - U\|_{N_{\mathcal{T}}(T)}$ without oscillation, it is natural to mark elements with largest error indicators. This leads to a minimal set \mathcal{M} , as stated in Assumption 25, and turns out to be crucial to link AFEM with optimal meshes.

Remark 29 (Initial triangulation). Assumption 26 guarantees the complexity estimate of module REFINED stated in Theorem 2: $\#\mathcal{T}_k - \#\mathcal{T}_0 \leq \Lambda_0 \sum_{j=0}^{k-1} \#\mathcal{M}_j$.

Even though we cannot expect local upper bounds between the continuous and discrete solution, the following crucial result shows that this is not the case between discrete solutions on nested meshes $\mathcal{T}_* \geq \mathcal{T}$: what matters is the set of elements of \mathcal{T} which are no longer in \mathcal{T}_* [33, 10, 28].

Lemma 30 (Localized upper bound). *Let $\mathcal{T}, \mathcal{T}_* \in \mathbb{T}$ satisfy $\mathcal{T}_* \geq \mathcal{T}$ and let $\mathcal{R} := \mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_*}$ be the refined set. If $U \in \mathbb{V}$, $U_* \in \mathbb{V}_*$ are the corresponding Galerkin solutions, then*

$$\|U_* - U\|_{\Omega}^2 \leq C_1 \mathcal{E}_{\mathcal{T}}^2(U, \mathcal{R}). \tag{35}$$

We are now ready to explore Stevenson’s insight (28) for the simplest scenario with vanishing oscillation $\text{osc}_{\mathcal{T}}(U) = 0$.

Lemma 31 (Dörfler marking: vanishing oscillation). *Let θ satisfy Assumption 24 and set $\mu := 1 - \theta^2/\theta_*^2 > 0$. Let $\mathcal{T}_* \geq \mathcal{T}$ and $U_* \in \mathbb{V}(\mathcal{T}_*)$ satisfy*

$$\|u - U_*\|_{\Omega}^2 \leq \mu \|u - U\|_{\Omega}^2. \tag{36}$$

Then the refined set $\mathcal{R} = \mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_}$ satisfies the Dörfler property*

$$\mathcal{E}_{\mathcal{T}}(U, \mathcal{R}) \geq \theta \mathcal{E}_{\mathcal{T}}(U, \mathcal{T}). \tag{37}$$

Proof. Since $\mu < 1$ we use the lower bound in (34), in conjunction with (36) and Pythagoras equality (18), to derive

$$(1 - \mu)C_2\mathcal{E}_{\mathcal{T}}^2(U, \mathcal{T}) \leq (1 - \mu)\|u - U\|_{\Omega}^2 \leq \|u - U\|_{\Omega}^2 - \|u - U_*\|_{\Omega}^2 = \|U - U_*\|_{\Omega}^2.$$

In view of Lemma 30, we thus deduce

$$(1 - \mu)C_2\mathcal{E}_{\mathcal{T}}^2(U, \mathcal{T}) \leq C_1\mathcal{E}_{\mathcal{T}}^2(U, \mathcal{R}),$$

which is the assertion in disguise. □

To examine the cardinality of \mathcal{M}_k in terms of $\|u - U_k\|_{\Omega}$ we must relate AFEM with the approximation class \mathcal{A}_s . Even though this might appear like an undoable task, the key to unravel this connection is given by Lemma 31. We show this now.

Lemma 32 (Cardinality of \mathcal{M}_k). *Let Assumptions 24 and 25 hold. If $u \in \mathcal{A}_s$ then*

$$\#\mathcal{M}_k \lesssim |u|_s^{1/s} \|u - U_k\|_{\Omega}^{-1/s} \quad \text{for all } k \geq 0. \tag{38}$$

Proof. We invoke that $u \in \mathcal{A}_s$ and (31) with $\varepsilon^2 = \mu \|u - U_k\|_{\Omega}^2$ to find a mesh $\mathcal{T}_{\varepsilon} \in \mathbb{T}$ and the Galerkin solution $U_{\varepsilon} \in \mathbb{V}(\mathcal{T}_{\varepsilon})$ so that

$$\|u - U_{\varepsilon}\|_{\Omega}^2 \leq \varepsilon^2, \quad \#\mathcal{T}_{\varepsilon} - \#\mathcal{T}_0 \lesssim |u|_s^{\frac{1}{s}} \varepsilon^{-\frac{1}{s}}.$$

Since $\mathcal{T}_{\varepsilon}$ may be totally unrelated to \mathcal{T}_k , we introduce the overlay $\mathcal{T}_* = \mathcal{T}_{\varepsilon} \oplus \mathcal{T}_k$. We exploit the property $\mathcal{T}_* \geq \mathcal{T}_{\varepsilon}$ to conclude that the Galerkin solution $U_* \in \mathbb{V}(\mathcal{T}_*)$ satisfies

$$\|u - U_*\|_{\Omega}^2 \leq \|u - U_{\varepsilon}\|_{\Omega}^2 \leq \varepsilon^2 = \mu \|u - U\|_{\Omega}^2.$$

Therefore, Lemma 31 implies that the refined set $\mathcal{R} = \mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_*}$ satisfies a Dörfler marking with parameter $\theta < \theta_*$. But MARK delivers a minimal set \mathcal{M}_k with this property, according to Assumption 25, whence

$$\#\mathcal{M}_k \leq \#\mathcal{R} \leq \#\mathcal{T}_* - \#\mathcal{T}_k \leq \#\mathcal{T}_{\varepsilon} - \#\mathcal{T}_0 \lesssim |u|_s^{\frac{1}{s}} \varepsilon^{-\frac{1}{s}},$$

where we use Lemma 20 to account for the overlay. The proof is complete. □

Proposition 33 (Quasi-optimality: vanishing oscillation). *Let Assumptions 24-26 hold. If $u \in \mathcal{A}_s$, then AFEM gives rise to a sequence $(\mathcal{T}_k, \mathbb{V}_k, U_k)_{k=0}^{\infty}$ such that*

$$\|u - U_k\|_{\Omega} \lesssim |u|_s (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s} \quad \text{for all } k \geq 1.$$

Proof. We make use of Assumption 26, along with Theorem 2, to infer that

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \leq \Lambda_0 \sum_{j=0}^{k-1} \#\mathcal{M}_j \lesssim |u|_s^{\frac{1}{s}} \sum_{j=0}^{k-1} \|u - U_j\|_{\Omega}^{-\frac{1}{s}}.$$

We now use the contraction property $\|u - U_k\|_\Omega \leq \alpha^{k-j} \|u - U_j\|_\Omega$ of Lemma 14 to replace the sum above by

$$\sum_{j=0}^{k-1} \|u - U_j\|_\Omega^{-\frac{1}{s}} \leq \|u - U_k\|_\Omega^{-\frac{1}{s}} \sum_{j=0}^{k-1} \alpha^{\frac{k-j}{s}} < \frac{\alpha^{\frac{1}{s}}}{1 - \alpha^{\frac{1}{s}}} \|u - U_k\|_\Omega^{-\frac{1}{s}},$$

because $\alpha < 1$ and the series is summable. This completes the proof. □

6.4. Quasi-Optimal Cardinality: General Data. In this section we remove the restriction $\text{osc}_\mathcal{T}(U) = 0$, and thereby make use of the basic ingredients developed in §6.1 and §6.2. Therefore, we replace the energy error by the total error and the linear approximation class \mathcal{A}_s for u by the nonlinear class \mathbb{A}_s for the triple (u, f, \mathbf{A}) ; see (31) for the definition of \mathbb{A}_s . To account for the presence of general data f and \mathbf{A} , we need to make an even more stringent assumption on the threshold θ_* .

Assumption 34 (Marking parameter: general data). *Let $C_3 = \Lambda_1 \text{osc}_{\mathcal{T}_0}^2(\mathbf{A})$ be the constant in Lemma 19. The marking parameter θ satisfies $\theta \in (0, \theta_*)$ with*

$$\theta_* = \sqrt{\frac{C_2}{1 + C_1(1 + C_3)}}.$$

We now proceed along the same lines as those of §6.3.

Lemma 35 (Dörfler marking: general data). *Let Assumption 34 hold and set $\mu := \frac{1}{2}(1 - \frac{\theta^2}{\theta_*^2}) > 0$. If $\mathcal{T}_* \geq \mathcal{T}$ and $U_* \in \mathbb{V}(\mathcal{T}_*)$ satisfy*

$$\|u - U_*\|_\Omega^2 + \text{osc}_{\mathcal{T}_*}^2(U_*) \leq \mu(\|u - U\|_\Omega^2 + \text{osc}_\mathcal{T}^2(U)), \tag{39}$$

then the refined set $\mathcal{R} = \mathcal{R}_{\mathcal{T} \rightarrow \mathcal{T}_*}$ satisfies the Dörfler property

$$\mathcal{E}_\mathcal{T}(U, \mathcal{R}) \geq \theta \mathcal{E}_\mathcal{T}(U, \mathcal{T}). \tag{40}$$

Proof. We split the proof into four steps.

□ In view of the global lower bound (20b) and (39), we can write

$$\begin{aligned} (1 - 2\mu) C_2 \mathcal{E}_\mathcal{T}^2(U) &\leq (1 - 2\mu) (\|u - U\|_\Omega^2 + \text{osc}_\mathcal{T}^2(U)) \\ &\leq (\|u - U\|_\Omega^2 - 2\|u - U_*\|_\Omega^2) + (\text{osc}_\mathcal{T}^2(U) - 2\text{osc}_{\mathcal{T}_*}^2(U_*)). \end{aligned}$$

□ Combining the Pythagoras orthogonality relation (18)

$$\|u - U\|_\Omega^2 - \|u - U_*\|_\Omega^2 = \|U - U_*\|_\Omega^2.$$

with the localized upper bound (35) yields

$$\|u - U\|_\Omega^2 - 2\|u - U_*\|_\Omega^2 \leq \|U - U_*\|_\Omega^2 \leq C_1 \mathcal{E}_\mathcal{T}^2(U, \mathcal{R}).$$

[3] To deal with oscillation we decompose the elements of \mathcal{T} into two disjoint sets: \mathcal{R} and $\mathcal{T} \setminus \mathcal{R}$. In the former case, we have

$$\text{osc}_{\mathcal{T}}^2(U, \mathcal{R}) - 2 \text{osc}_{\mathcal{T}_*}^2(U_*, \mathcal{R}) \leq \text{osc}_{\mathcal{T}}^2(U, \mathcal{R}) \leq \mathcal{E}_{\mathcal{T}}^2(U, \mathcal{R}),$$

because $\text{osc}_{\mathcal{T}}(U, T) \leq \mathcal{E}_{\mathcal{T}}(U, T)$ for all $T \in \mathcal{T}$. On the other hand, we use that $\mathcal{T} \setminus \mathcal{R} = \mathcal{T} \cap \mathcal{T}_*$ and apply a variant of Lemma 11 for $\text{osc}_{\mathcal{T}}(U)$ together with Lemma 30, to get

$$\text{osc}_{\mathcal{T}}^2(U, \mathcal{T} \setminus \mathcal{R}) - 2 \text{osc}_{\mathcal{T}_*}^2(U_*, \mathcal{T} \setminus \mathcal{R}) \leq C_3 \|U - U_*\|_{\Omega}^2 \leq C_1 C_3 \mathcal{E}_{\mathcal{T}}^2(U, \mathcal{R}).$$

Adding these two estimates gives

$$\text{osc}_{\mathcal{T}}^2(U) - 2 \text{osc}_{\mathcal{T}_*}^2(U_*) \leq (1 + C_1 C_3) \mathcal{E}_{\mathcal{T}}^2(U, \mathcal{R}).$$

[4] Returning to [1] we realize that

$$(1 - 2\mu) C_2 \mathcal{E}_{\mathcal{T}}^2(U, \mathcal{T}) \leq (1 + C_1(1 + C_3)) \mathcal{E}_{\mathcal{T}}^2(U, \mathcal{R}),$$

which is the asserted estimate (40) in disguise. □

Lemma 36 (Cardinality of \mathcal{M}_k : general data). *Let Assumptions 25 and 34 hold. If the triple $(u, f, \mathbf{A}) \in \mathbb{A}_s$, then*

$$\#\mathcal{M}_k \lesssim |u, f, \mathbf{A}|_s^{1/s} (\|u - U_k\|_{\Omega} + \text{osc}_k(U_k))^{-1/s} \quad \text{for all } k \geq 0. \quad (41)$$

Proof. We split the proof into three steps.

[1] We set $\varepsilon^2 := \mu \Lambda_2^{-1} (\|u - U_k\|_{\Omega}^2 + \text{osc}_k^2(U_k))$ with $\mu = \frac{1}{2} (1 - \frac{\theta^2}{\theta_*^2}) > 0$ as in Lemma 35 and Λ_2 given Lemma 19. Since $(u, f, \mathbf{A}) \in \mathbb{A}_s$, in view of (31) there exists $\mathcal{T}_{\varepsilon} \in \mathbb{T}$ and $U_{\varepsilon} \in \mathbb{V}(\mathcal{T}_{\varepsilon})$ such that

$$\|u - U_{\varepsilon}\|_{\Omega}^2 + \text{osc}_{\varepsilon}^2(U_{\varepsilon}) \leq \varepsilon^2 \quad \text{and} \quad \#\mathcal{T}_{\varepsilon} - \#\mathcal{T}_0 \lesssim |u, f, \mathbf{A}|_s^{1/2} \varepsilon^{-1/s}.$$

Since $\mathcal{T}_{\varepsilon}$ may be totally unrelated to \mathcal{T}_k we introduce the overlay $\mathcal{T}_* = \mathcal{T}_k \oplus \mathcal{T}_{\varepsilon}$.

[2] We claim that the total error over \mathcal{T}_* reduces by a factor μ relative to that one over \mathcal{T}_k . In fact, since $\mathcal{T}_* \geq \mathcal{T}_{\varepsilon}$ and so $\mathbb{V}(\mathcal{T}_*) \supset \mathbb{V}(\mathcal{T}_{\varepsilon})$, we use Lemma 19 to obtain

$$\begin{aligned} \|u - U_*\|_{\Omega}^2 + \text{osc}_{\mathcal{T}_*}^2(U_*) &\leq \Lambda_2 \left(\|u - U_{\varepsilon}\|_{\Omega}^2 + \text{osc}_{\varepsilon}^2(U_{\varepsilon}) \right) \\ &\leq \Lambda_2 \varepsilon^2 = \mu (\|u - U_k\|_{\Omega}^2 + \text{osc}_k^2(U_k)). \end{aligned}$$

Upon applying Lemma 35 we conclude that the set $\mathcal{R} = \mathcal{R}_{\mathcal{T}_k \rightarrow \mathcal{T}_*}$ of refined elements satisfies a Dörfler marking (40) with parameter $\theta < \theta_*$.

[3] According to Assumption 25, MARK selects a minimal set \mathcal{M}_k satisfying this property. Therefore, employing Lemma 20 to account for the cardinality of the overlay, we deduce

$$\#\mathcal{M}_k \leq \#\mathcal{R} \leq \#\mathcal{T}_* - \#\mathcal{T}_k \leq \#\mathcal{T}_\varepsilon - \#\mathcal{T}_0 \lesssim |u, f, \mathbf{A}|_s^{1/s} \varepsilon^{-1/s}.$$

Finally, recalling the definition of ε we end up with the asserted estimate (41). □

We are ready to prove the main result of this section, which combines Theorem 16 and Lemma 36.

Theorem 37 (Quasi-optimality: general data). *Let Assumptions 25, 26 and 34 hold. If $(u, f, \mathbf{A}) \in \mathbb{A}_s$, then AFEM gives rise to a sequence $(\mathcal{T}_k, \mathbb{V}_k, U_k)_{k=0}^\infty$ such that*

$$\|u - U_k\|_\Omega + \text{osc}_k(U_k) \lesssim |u, f, \mathbf{A}|_s (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-s} \quad \text{for all } k \geq 1.$$

Proof. Since no confusion arises, we use the notation $\text{osc}_j = \text{osc}_j(U_j)$ and $\mathcal{E}_j = \mathcal{E}_j(U_j)$.

[1] In light of Assumption 26, which yields Theorem 2, and (41) we have

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \lesssim \sum_{j=0}^{k-1} \#\mathcal{M}_j \lesssim |u, f, \mathbf{A}|_s^{1/s} \sum_{j=0}^{k-1} (\|u - U_j\|_\Omega^2 + \text{osc}_j^2)^{-1/(2s)}.$$

[2] Let $\gamma > 0$ be the scaling factor in the (contraction) Theorem 16. The lower bound (20b) along with $\text{osc}_j \leq \mathcal{E}_j$ implies

$$\|u - U_j\|_\Omega^2 + \gamma \text{osc}_j^2 \leq \|u - U_j\|_\Omega^2 + \gamma \mathcal{E}_j^2 \leq \left(1 + \frac{\gamma}{C_2}\right) (\|u - U_j\|_\Omega^2 + \text{osc}_j^2).$$

[3] Theorem 16 yields for $0 \leq j < k$

$$\|u - U_k\|_\Omega^2 + \gamma \mathcal{E}_k^2 \leq \alpha^{2(k-j)} (\|u - U_j\|_\Omega^2 + \gamma \mathcal{E}_j^2),$$

whence

$$\#\mathcal{T}_k - \#\mathcal{T}_0 \lesssim |u, f, \mathbf{A}|_s^{1/s} (\|u - U_k\|_\Omega^2 + \gamma \mathcal{E}_k^2)^{-1/(2s)} \sum_{j=0}^{k-1} \alpha^{(k-j)/s}.$$

Since $\sum_{j=0}^{k-1} \alpha^{(k-j)/s} < \sum_{j=1}^\infty \alpha^{j/s} < \infty$ because $\alpha < 1$, the assertion follows easily. □

We conclude this section with a couple of applications of Theorem 37. The first one is valid for the example of §1.

Corollary 38 (W_p^2 -regularity with piecewise constant \mathbf{A}). *Let $d = 2$, the polynomial degree be $n = 1$, $f \in L^2(\Omega)$, and let \mathbf{A} be piecewise constant over \mathcal{T}_0 . If $u \in W_p^2(\Omega; \mathcal{T}_0)$ for $p > 1$, then AFEM gives rise to a sequence $\{\mathcal{T}_k, \mathbb{V}_k, U_k\}_{k=0}^\infty$ satisfying $\text{osc}_k(U_k) = \|h_k(f - P_0 f)\|_{L^2(\Omega)}$ and for all $k \geq 1$*

$$\|u - U_k\|_\Omega + \text{osc}_k(U_k) \lesssim \left(\|D^2 u\|_{L^p(\Omega; \mathcal{T}_0)} + \|f\|_{L^2(\Omega)} \right) (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-1/2}.$$

Proof. Combine Corollary 22 with Theorem 37. \square

Corollary 39 (W_p^2 -regularity with variable \mathbf{A}). *Besides the assumptions of Corollary 38, let \mathbf{A} be piecewise Lipschitz over the initial grid \mathcal{T}_0 . Then AFEM gives rise to a sequence $\{\mathcal{T}_k, \mathbb{V}_k, U_k\}_{k=0}^\infty$ satisfying for all $k \geq 1$*

$$\begin{aligned} \|u - U_k\|_\Omega + \text{osc}_k(U_k) \\ \lesssim \left(\|D^2 u\|_{L^p(\Omega; \mathcal{T}_0)} + \|f\|_{L^2(\Omega)} + \|\mathbf{A}\|_{W_\infty^1(\Omega; \mathcal{T}_0)} \right) (\#\mathcal{T}_k - \#\mathcal{T}_0)^{-1/2}. \end{aligned}$$

Proof. Combine Corollary 23 with Theorem 37. \square

7. Extensions and Limitations

Nonconforming Meshes. Bonito and Nochetto [7] have shown that Theorem 2 extends to *admissible* nonconforming meshes for $d \geq 2$ (those with a fixed level of nonconformity), along with the theory of §5 and §6.

Discontinuous Galerkin Methods (dG). Bonito and Nochetto [7] have also shown that such theory extends to the interior penalty dG method for the model problem (1) and for $d \geq 2$. This relies on a result of independent interest:

the approximation classes for discontinuous and continuous elements of any degree $n \geq 1$ coincide.

Non-residual Estimators. Cascón and Nochetto [12] and Kreuzer and Siebert [22] have extended the above theory to non-residual estimators (hierarchical estimators, Zienkiewicz-Zhu and Braess-Schoerbel estimators, and those based on the solution of local problems).

Other Norms. The above theory is just for the energy norm. We refer to Demlow [15] for local energy norms and Demlow and Stevenson [16] for the L^2 -norm. The theory for more practical norms, such as L^∞ or W_∞^1 , is open.

Other Problems and Markings. The theory above relies strongly on the Pythagoras equality (18) and Dörfler marking (16), and extends to symmetric

problems in $H(\text{div})$ [11] and $H(\text{curl})$ [37] as well as to non-symmetric coercive problems [12]. For non-coercive problems, for which we just have an inf-sup condition, as well as markings other than Dörfler, the theory is mostly lacking except for mixed AFEM for (1) [13]. We refer to Morin, Siebert, and Vesser [27] and Siebert [31] for convergence results without rates.

Multilevel Methods on Graded Meshes. We refer to Xu, Chen, and Nochetto [36] for a theory of multilevel methods on graded meshes created by bisection. The analysis uses several geometric properties of bisection, discussed in [36], and is valid for any d and n .

References

- [1] M. AINSWORTH AND J.T. ODEN, *A Posteriori Error Estimation in Finite Element Analysis*, Pure and Applied Mathematics (New York), Wiley-Interscience, 2000.
- [2] I. BABUŠKA, R.B. KELLOGG, AND J. PITKÄRANTA, *Direct and inverse error estimates for finite elements with mesh refinements*, Numer. Math., 33 (1979), pp. 447–471.
- [3] I. BABUŠKA AND A. MILLER, *A feedback finite element method with a posteriori error estimation. I. The finite element method and some basic properties of the a posteriori error estimator*, Comput. Methods Appl. Mech. Engrg. 61 (1) (1987), pp. 1–40.
- [4] I. BABUŠKA AND W. RHEINBOLDT, *Error estimates for adaptive finite element computations* SIAM J. Numer. Anal., 15 (1978), pp. 736–754.
- [5] P. BINEV, W. DAHMEN, AND R. DEVORE, *Adaptive finite element methods with convergence rates*, Numer. Math., 97 (2004), pp. 219–268.
- [6] P. BINEV, W. DAHMEN, R. DEVORE, AND P. PETRUSHEV, *Approximation classes for adaptive methods*, Serdica Math. J., 28 (2002), pp. 391–416. Dedicated to the memory of Vassil Popov on the occasion of his 60th birthday.
- [7] A. BONITO AND R.H. NOCHETTO, *Quasi-optimal convergence rate for an adaptive discontinuous Galerkin method*, SIAM J. Numer. Anal. (to appear).
- [8] D. BRAESS, *Finite Elements. Theory, Fast Solvers, and Applications in Solid Mechanics*, 2nd edition edn. Cambridge University Press (2001).
- [9] S. BRENNER AND L.R. SCOTT, *The Mathematical Theory of Finite Element Methods*, Springer Texts in Applied Mathematics 15 (2008).
- [10] J. M. CASCÓN, C. KREUZER, R. H. NOCHETTO, AND K. G. SIEBERT, *Quasi-optimal convergence rate for an adaptive finite element method*, SIAM J. Numer. Anal., 46 (2008), pp. 2524–2550.
- [11] J. M. CASCÓN, L. CHEN, R. H. NOCHETTO, AND K. G. SIEBERT, *Quasi-optimal convergence rates for AFEM in $H(\text{div})$* , (in preparation).
- [12] J. M. CASCÓN AND R. H. NOCHETTO, *Convergence and quasi-optimality for AFEM based on non-residual a posteriori error estimators*, (in preparation).

- [13] L. CHEN, M. HOLST, AND J. XU, *Convergence and optimality of adaptive mixed finite element methods*, Math. Comp., 78 (2009), pp. 35–53.
- [14] Z. CHEN AND J. FENG, *An adaptive finite element algorithm with reliable and efficient error control for linear parabolic problems*, Math. Comp., 73 (2006), pp. 1167–1042.
- [15] A. DEMLOW, *Convergence of an adaptive finite element method for controlling local energy errors*, (submitted).
- [16] A. DEMLOW AND R. STEVENSON, *Convergence and quasi-optimality of an adaptive finite element method for controlling L_2 errors*, (submitted).
- [17] R.A. DEVORE, *Nonlinear approximation*, Acta Numerica, 7, (1998), pp. 51–150.
- [18] W. DÖRFLER, *A convergent adaptive algorithm for Poisson’s equation*, SIAM J. Numer. Anal., 33 (1996), pp. 1106–1124.
- [19] P. GRISVARD, *Elliptic Problems in Nonsmooth Domains, Monographs and Studies in Mathematics*, vol. 24. Pitman (Advanced Publishing Program), Boston, MA (1985).
- [20] R.B. KELLOGG, *On the Poisson equation with intersecting interfaces*, Applicable Anal., 4 (1974/75), 101–129.
- [21] I. KOSSACZKÝ, *A recursive approach to local mesh refinement in two and three dimensions*, J. Comput. Appl. Math., 55 (1994), pp. 275–288.
- [22] CH. KREUZER AND K.G. SIEBERT, *Decay rates of adaptive finite elements with Dörfler marking*, (submitted).
- [23] K. MEKCHAY AND R. H. NOCHETTO, *Convergence of adaptive finite element methods for general second order linear elliptic PDEs*, SIAM J. Numer. Anal., 43 (2005), pp. 1803–1827 (electronic).
- [24] W. F. MITCHELL, *A Comparison of adaptive refinement techniques for elliptic problems*, ACM Trans. Math. Softw., 15 (1989), pp. 326 - 347.
- [25] P. MORIN, R. H. NOCHETTO, AND K. G. SIEBERT, *Data oscillation and convergence of adaptive FEM*, SIAM J. Numer. Anal., 38 (2000), pp. 466–488.
- [26] P. MORIN, R. H. NOCHETTO, AND K. G. SIEBERT, *Convergence of adaptive finite element methods*, SIAM Review, 44 (2002), pp. 631–658.
- [27] P. MORIN, K. G. SIEBERT, AND A. VEESER, *A basic convergence result for conforming adaptive finite elements*, Math. Mod. Meth. Appl. Sci., 5 (2008), pp. 707–737.
- [28] R.H. NOCHETTO, K. G. SIEBERT, AND A. VEESER, *Theory of adaptive finite element methods: an introduction*, in *Multiscale, Nonlinear and Adaptive Approximation*, A. Kunoth and R. DeVore eds, Springer, 2009, pp. 409–542.
- [29] R. H. NOCHETTO AND A. VEESER, *Primer of adaptive finite element methods*, in *Multiscale and Adaptivity: Modeling, Numerics and Applications*, CIME-EMS Summer School in Applied Mathematics, G. Naldi and G. Russo eds., Springer (2010).
- [30] L. R. SCOTT AND S. ZHANG, *Finite element interpolation of nonsmooth functions satisfying boundary conditions*, Math. Comp., 54 (1990), pp. 483–493.

-
- [31] K. G. SIEBERT, *A convergence proof for adaptive finite elements without lower bound*, Preprint Universität Duisburg-Essen and Universität Freiburg No. 1/2009.
- [32] K. G. SIEBERT, *Mathematically founded design of adaptive finite element software*, in *Multiscale and Adaptivity: Modeling, Numerics and Applications*, CIME-EMS Summer School in Applied Mathematics, G. Naldi and G. Russo eds., Springer (2010).
- [33] R. STEVENSON, *Optimality of a standard adaptive finite element method*, *Found. Comput. Math.*, 7 (2007), pp. 245–269.
- [34] R. STEVENSON, *The completion of locally refined simplicial partitions created by bisection*, *Math. Comput.*, 77 (2008), pp. 227–241.
- [35] R. VERFÜRTH, *A Review of A Posteriori Error Estimation and Adaptive Mesh-Refinement Techniques*, *Adv. Numer. Math.* John Wiley, Chichester, UK (1996).
- [36] J. XU, L. CHEN, AND R. H. NOCHETTO, *Adaptive multilevel methods on graded bisection grids*, in *Multiscale, Nonlinear and Adaptive Approximation*, A. Kunothe and R. DeVore eds, Springer, 2009, pp. 599–659.
- [37] L. ZHONG, L. CHEN, S. SHU, G. WITTUM, AND J. XU, *Quasi-optimal convergence of adaptive edge finite element methods for three dimensional indefinite time-harmonic Maxwell's equations*, (submitted).

Wavelet Frames and Image Restorations

Zuowei Shen*

Abstract

One of the major driven forces in the area of applied and computational harmonic analysis over the last decade or longer is the development of redundant systems that have sparse approximations of various classes of functions. Such redundant systems include *framelet* (tight wavelet frame), *ridgelet*, *curvelet*, *shearlet* and so on. This paper mainly focuses on a special class of such redundant systems: tight wavelet frames, especially, those tight wavelet frames generated via a multiresolution analysis. In particular, we will survey the development of the unitary extension principle and its generalizations. A few examples of tight wavelet frame systems generated by the unitary extension principle are given. The unitary extension principle makes constructions of tight wavelet frame systems straightforward and painless which, in turn, makes a wide usage of the tight wavelet frames possible. Applications of wavelet frame, especially frame based image restorations, are also discussed in details.

Mathematics Subject Classification (2010). Primary 42C15; 42C40; 94A08
Secondary 42C30; 65T60; 90C90.

Keywords. Tight wavelet frames, Unitary extension principle, Image restorations.

1. Introduction

Since the publication of [35, 69] on compactly supported orthonormal wavelet generated by the multiresolution analysis (MRA), wavelet analysis and its applications lead the area of applied and computational harmonic analysis over the last two decades and wavelet methods become powerful tools in various applications in image and signal analysis and processing. One of the well known successful examples of applications of wavelets is image compression using orthonormal or bi-orthogonal wavelet bases generated by the MRA as given in [32, 35]. Another successful example of applications of wavelets is noise removal using redundant wavelet systems by [33, 44].

*Department of Mathematics, National University of Singapore, Singapore 119076.
E-mail: matzuows@nus.edu.sg.

Theory of frames, especially theory of the Gabor frames (see e.g. [36, 58, 70]) and wavelet frames (see e.g. [36, 70]), has a long history of the development even before the discovery of the multiresolution analysis of [69] and the systematic construction of compactly supported orthonormal wavelets of [35]. The concept of frame can be traced back to [47]. The wide scope of applications of frames can be found in the early literature on applications of Gabor and wavelet frames (see e.g. [36, 58, 70]). Such applications include time frequency analysis for signal processing, coherent state in quantum mechanics, filter bank design in electrical engineering, edge and singularity detection in image processing, and etc. It is not the goal of this paper to give a survey on all of these and the interested reader should consult [36, 58, 70, 71, 72] and references therein for details.

The publication of the unitary extension principle of [79] generates wide interests in tight wavelet frame systems derived by multiresolution analysis. One can find the rich literature by consulting [31, 40] and the references in these papers. Having tight wavelet frames with a multiresolution structure is very important in order to make any use of them in applications, since this guarantees the existence of the fast decomposition and reconstruction algorithms. Recently, tight wavelet frames derived by the multiresolution analysis are used to open a few new areas of applications of frames. The application of tight wavelet frames in image restorations is one of them that includes image inpainting, image denoising, image deblurring and blind deblurring, and image decompositions (see e.g. [8, 9, 10, 13, 14, 20, 23, 24, 25, 28]). In particular, the unitary extension principle is used in [8, 20, 23, 25, 28] to design a tight wavelet frame system adaptive to the real life problems in hand. Frame based algorithms for image and surface segmentation, 3D surface reconstruction, and CT image reconstruction are currently being explored.

In this paper, we start with a brief survey of the theory of tight wavelet frames. A characterization of the tight wavelet frame of [54, 59, 79] is given. We then focus on the tight wavelet frames and their constructions via the multiresolution analysis (MRA). In particular, the unitary extension principle of [79] and the construction of tight wavelet frame from it will be given. We will also give an overview of the generalizations of the unitary extension principle. The second part of this paper focuses on the recent applications of tight wavelet frames in image restorations. In particular, the balanced approach of [8, 9, 10, 20, 23, 24, 25, 28] and the corresponding algorithms for image denoising, deblurring, inpainting and decomposition will be discussed in details.

Finally we remark that there are a few redundant wavelet systems other than the tight wavelet systems discussed here that are developed fast and used widely in image and signal analysis and processing. Such redundant systems include, for example, bi-frames of [31, 40, 59, 80], ridgelets of [46], curvelets of [21, 22], and shearlets of [60, 67]. We forgo discussing all of these in this paper in order to have a well focus of this paper and the interested reader should consult the relevant references for the details.

2. Tight Wavelet Frame

We introduce the notion of tight wavelet frame in space $L_2(\mathbb{R})$, together with some other basic concepts and notations. The space $L_2(\mathbb{R})$ is the set of all the functions $f(x)$ satisfying $\|f\|_{L_2(\mathbb{R})} := (\int_{\mathbb{R}} |f(x)|^2 dx)^{\frac{1}{2}} < \infty$ and, similar, $\ell_2(\mathbb{Z})$ is the set of all sequences defined on \mathbb{Z} satisfying $\|h\|_{\ell_2(\mathbb{Z})} := (\sum_{k \in \mathbb{Z}} |h[k]|^2)^{\frac{1}{2}} < \infty$.

For any function $f \in L_2(\mathbb{R})$, the dyadic dilation operator \mathcal{D} is defined by $\mathcal{D}f(x) := \sqrt{2}f(2x)$ and the translation operator T is defined by $T_a f(x) := f(x - a)$ for $a \in \mathbb{R}$. Given $j \in \mathbb{Z}$, we have $T_a \mathcal{D}^j = \mathcal{D}^j T_{2^j a}$.

For given $\Psi := \{\psi_1, \dots, \psi_r\} \subset L_2(\mathbb{R})$, define the wavelet system

$$X(\Psi) := \{\psi_{\ell,j,k} : 1 \leq \ell \leq r; j, k \in \mathbb{Z}\},$$

where $\psi_{\ell,j,k} = \mathcal{D}^j T_k \psi_{\ell} = 2^{j/2} \psi_{\ell}(2^j \cdot -k)$. The system $X(\Psi) \subset L_2(\mathbb{R})$ is called a tight wavelet frame of $L_2(\mathbb{R})$ if

$$\|f\|_{L_2(\mathbb{R})}^2 = \sum_{g \in X(\Psi)} |\langle f, g \rangle|^2,$$

holds for all $f \in L_2(\mathbb{R})$, where $\langle \cdot, \cdot \rangle$ is the inner product in $L_2(\mathbb{R})$ and $\|\cdot\|_{L_2(\mathbb{R})} = \sqrt{\langle \cdot, \cdot \rangle}$. This is equivalent to $f = \sum_{g \in X(\Psi)} \langle f, g \rangle g$, for all $f \in L_2(\mathbb{R})$.

It is clear that an orthonormal basis is a tight frame. When $X(\Psi)$ forms an orthonormal basis of $L_2(\mathbb{R})$, then $X(\Psi)$ is called an orthonormal wavelet basis. When $X(\Psi)$ forms a tight frame of $L_2(\mathbb{R})$, then $X(\Psi)$ is called a tight wavelet frame. We note that in some literature, the definition of tight frame here is called the tight frame with bound one or Parseval frame.

Finally, the Fourier transform of a function $f \in L_1(\mathbb{R})$ is defined as usual by:

$$\widehat{f}(\omega) := \int_{\mathbb{R}} f(x) e^{-i\omega x} dx, \quad \omega \in \mathbb{R},$$

and its inverse is

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \widehat{f}(\omega) e^{i\omega x} d\omega, \quad x \in \mathbb{R}.$$

They can be extended to more general functions, e.g. the functions in $L_2(\mathbb{R})$. Similarly, we can define the Fourier series for a sequence $h \in \ell_2(\mathbb{Z})$ by

$$\widehat{h}(\omega) := \sum_{k \in \mathbb{Z}} h[k] e^{-ik\omega}, \quad \omega \in \mathbb{R}.$$

2.1. A characterization. To characterize the wavelet system $X(\Psi)$ to be a tight frame or even an orthonormal basis for $L_2(\mathbb{R})$ in terms of its generators Ψ , the dual Gramian analysis of [78] is used in [79].

The dual Gramian analysis identifies the frame operator corresponding to the wavelet system $X(\Psi)$ as the dual Gramian matrix with each entry being written in term of the Fourier transform of the generators Ψ . Recall that for a given system $X(\Psi)$, the corresponding frame operator is defined by

$$Sf = \sum_{g \in X(\Psi)} \langle f, g \rangle g, \quad f \in L_2(\mathbb{R}).$$

It is clear that $X(\Psi)$ is a tight frame of $L_2(\mathbb{R})$ if and only if S is the identity. The dual Gramian analysis decomposes the operator S into a collection of simpler operators which is called fibers in [78] in Fourier domain. The operator S is the identity if and only if each fiber operator is the identity. This leads to the conclusion that wavelet system $X(\Psi)$ forms a tight frame of $L_2(\mathbb{R})$ if and only if the dual Gramian matrix corresponding to the wavelet system $X(\Psi)$ is the identity almost everywhere. Writing each entry of the dual Gramian explicitly, one obtains the following theorem (see, e.g. Corollary 1.3 of [79]):

Theorem 1. *The wavelet system $X(\Psi)$ is a tight frame of $L_2(\mathbb{R})$ if and only if the identities*

$$\sum_{\psi \in \Psi} \sum_{k \in \mathbb{Z}} |\widehat{\psi}(2^k \omega)|^2 = 1; \quad \sum_{\psi \in \Psi} \sum_{k=0}^{\infty} \widehat{\psi}(2^k \omega) \overline{\widehat{\psi}(2^k(\omega + (2j+1)2\pi))} = 0 \quad j \in \mathbb{Z} \quad (1)$$

hold for a.e. $\omega \in \mathbb{R}$. Furthermore, $X(\Psi)$ is an orthonormal basis of $L_2(\mathbb{R})$ if and only if (1) holds and $\|\psi\| = 1$ for all $\psi \in \Psi$.

Note that the key part of this theorem is the tight frame part. The orthonormal basis part follows from the fact that a tight frame with each generator having norm one is an orthonormal basis. The details about the dual Gramian analysis can be found in [78, 81]. The dual Gramian analysis is also applied to the Gabor frame analysis in [82] to derive the duality principle for the Gabor frames.

There were many contributions, during the last two decades, to the study of the Bessel, frame and other related properties of wavelet systems. Examples of univariate wavelet frames can be found in [37]; necessary and sufficient conditions for mother wavelets to generate frames were discussed (implicitly) in [36, 71]. Characterizations of univariate orthonormal basis associated with integer dilation were established independently in [57] and [64], with the multivariate counterparts of these results appearing in [54] for the dyadic dilation. Characterization of bi-frame (tight frame is a special case) in multivariate case for an integer dilation matrix was given in [59]. Independently of all these, a general characterization of all wavelet frames whose dilation matrix is an integral (via dual Gramian analysis) were provided in [79] and derived from it a special characterization of tight wavelet frames in [79] and bi-frame in [80].

Although this theorem gives a complete characterization of the wavelet system $X(\Psi)$ being tight frame of $L_2(\mathbb{R})$, it provides little help for construction

of such wavelet system with compactly supported generators, although it may help to obtain tight wavelet frame systems with bandlimited generators. Furthermore, a tight wavelet frame from MRA is handy to use, since it has fast decomposition and reconstruction algorithms. This motivates the study of multiresolution analysis generated tight wavelet frames in [79] as we shall present next. The MRA based bandlimited tight wavelet frames are also constructed in [2].

2.2. Tight wavelet frame generated from MRA. The starting element of a multiresolution analysis is the concept of refinable function. Since we are interested here to construct compactly supported wavelets with finitely supported masks, for the simplicity, we start with a compactly supported refinable function ϕ , although the unitary extension principle can be stated for general refinable function in $L_2(\mathbb{R})$ (see [40, 79]). A compactly supported function $\phi \in L_2(\mathbb{R})$ is refinable if it satisfies the following refinement equation

$$\phi(x) = 2 \sum_{k \in \mathbb{Z}} h_0[k] \phi(2x - k), \quad (2)$$

for some finite supported sequence $h_0 \in \ell_2(\mathbb{Z})$. By taking the Fourier transform, equation (2) becomes

$$\widehat{\phi(2 \cdot)} = \widehat{h_0} \widehat{\phi}, \quad \text{a.e. } \omega \in \mathbb{R}.$$

We call the sequence h_0 the refinement mask of ϕ and $\widehat{h_0}(\omega)$ the refinement symbol of ϕ .

For a compactly supported refinable function $\phi \in L_2(\mathbb{R})$, let V_0 be the closed shift invariant space generated by $\{\phi(\cdot - k) : k \in \mathbb{Z}\}$ and $V_j := \{f(2^j \cdot) : f \in V_0\}$, $j \in \mathbb{Z}$. It is known that when ϕ is compactly supported, then $\{V_j\}_{j \in \mathbb{Z}}$ forms a multiresolution analysis (MRA). Here a multiresolution analysis is defined to be a family of closed subspaces $\{V_j\}_{j \in \mathbb{Z}}$ of $L_2(\mathbb{R})$ that satisfies: (i) $V_j \subset V_{j+1}$, (ii) $\bigcup_j V_j$ is dense in $L_2(\mathbb{R})$, and (iii) $\bigcap_j V_j = \{0\}$ (see [4, 66]). The unitary extension principle is a principle of construction of MRA based tight wavelet frame.

For a given ϕ , define the quasi-interpolatory operator as

$$P_j : f \mapsto \sum_{k \in \mathbb{Z}} \langle f, \phi_{j,k} \rangle \phi_{j,k}, \quad (3)$$

for an arbitrary $f \in L_2(\mathbb{R})$, where $\phi_{j,k} = 2^{j/2} \phi(2^j \cdot - k)$. It is clear that $P_j f \in V_j$. Since $\phi \in L_2(\mathbb{R})$ is refinable, one has

$$\sum_{k \in \mathbb{Z}} \phi(\cdot + k) = 1.$$

A standard proof from approximation theory shows that $\lim_{j \rightarrow \infty} P_j f = f$, (see e.g. [40]).

2.2.1. Unitary extension principle. Let V_j , $j \in \mathbb{Z}$ be the MRA generated by the refinable function ϕ and the refinement mask h_0 . Let $\Psi := \{\psi_1, \dots, \psi_r\} \subset V_1$ be of the form

$$\psi_\ell(x) = 2 \sum_{k \in \mathbb{Z}} h_\ell[k] \phi(2x - k). \quad (4)$$

The finitely supported sequences h_1, \dots, h_r are called wavelet masks, or the high pass filters of the system, and the refinement mask h_0 is called the low pass filter. In the Fourier domain, (4) can be written as

$$\widehat{\psi}_\ell(2\cdot) = \widehat{h}_\ell \widehat{\phi}, \quad \ell = 1, \dots, r, \quad (5)$$

where $\widehat{h}_1, \dots, \widehat{h}_r$ are 2π periodic functions and are called wavelet symbols.

The Unitary extension principle of [79] for this simple case can be stated as following. For the unitary extension principle in the most general setting, the interested reader should consult [40, 79] for the details.

Theorem 2 (Unitary Extension Principle, (UEP) [79]). *Let $\phi \in L_2(\mathbb{R})$ be the compactly supported refinable function with its finitely supported refinement mask h_0 satisfying $\widehat{h}_0(0) = 1$. Let (h_1, \dots, h_r) be a set of finitely supported sequences. Then the system $X(\Psi)$ where $\Psi = \{\psi_1, \dots, \psi_r\}$ defined in (4) forms a tight frame in $L_2(\mathbb{R})$ provided the equalities*

$$\sum_{\ell=0}^r |\widehat{h}_\ell(\xi)|^2 = 1 \quad \text{and} \quad \sum_{\ell=0}^r \widehat{h}_\ell(\xi) \overline{\widehat{h}_\ell(\xi + \pi)} = 0 \quad (6)$$

hold for almost all $\xi \in [-\pi, \pi]$. Furthermore, assuming $r = 1$ and $\|\phi\| = 1$, then $X(\Psi)$ is an orthonormal wavelet bases of $L_2(\mathbb{R})$.

Conditions in (6) can be written in terms of sequences h_0, \dots, h_r . The first condition becomes

$$\sum_{\ell=0}^r \sum_{k \in \mathbb{Z}} \overline{h_\ell[k]} h_\ell[k - p] = \delta_{0,p}, \quad p \in \mathbb{Z}, \quad (7)$$

where $\delta_{0,p} = 1$ when $p = 0$ and 0 otherwise and the second condition can be written as

$$\sum_{\ell=0}^r \sum_{k \in \mathbb{Z}} (-1)^{k-p} \overline{h_\ell[k]} h_\ell[k - p] = 0, \quad p \in \mathbb{Z}. \quad (8)$$

Proof. Let $\{V_j\}$, $j \in \mathbb{Z}$ be a given MRA with underlying refinable function ϕ ; P_j be the quasi-interpolatory operator defined in (3); and $\Psi = \{\psi_1, \dots, \psi_r\}$ be the set of corresponding tight framelets derived from the UEP. A simple calculation,

which is the standard decomposition and reconstruction algorithms for the tight wavelet frame given in [40], shows that condition (7) and (8) imply that

$$P_j f = P_{j-1} f + \sum_{\ell=1}^r \sum_{k \in \mathbb{Z}} \langle f, \psi_{\ell, j-1, k} \rangle \psi_{\ell, j-1, k}. \tag{9}$$

Iterating (9) and applying the fact $\lim_{j \rightarrow -\infty} V_j = \{0\}$ which follows from $\bigcap_{j \in \mathbb{Z}} V_j = \{0\}$, one derives that this quasi-interpolatory operator P_j is the same as *truncated representation*

$$Q_j : f \mapsto \sum_{\ell=1}^r \sum_{j' < j, k \in \mathbb{Z}} \langle f, \psi_{\ell, j', k} \rangle \psi_{\ell, j', k}, \tag{10}$$

i.e. $P_j f = Q_j f$. Since $\lim_{j \rightarrow \infty} P_j f = f$ for all $f \in L_2(\mathbb{R})$, one concludes that

$$f = \lim_j P_j f = \lim_{j \rightarrow \infty} Q_j f = \sum_{\ell=1}^r \sum_{j \in \mathbb{Z}, k \in \mathbb{Z}} \langle f, \psi_{\ell, j, k} \rangle \psi_{\ell, j, k}.$$

Hence, $X(\Psi)$ is a tight frame of $L_2(\mathbb{R})$. The orthonormal basis part follows from the fact that if $r = 1$ and $\|\phi\| = 1$, then the ψ constructed from the UEP has the norm one as well. ■

The generators Ψ via the UEP is called *framelet* in [40]. For the special case $r = 1$, the above theorem is given in [68]. The freedom of the choice of the number of the generators r in the UEP, makes the construction of tight framelets become painless. For example, one can construct tight framelets from spline easily. In fact, [79] gives a systematic construction of tight wavelet frame system from B-splines by using the UEP. Next, we give two examples of spline tight framelets of [79].

Example 1. Let $h_0 = [\frac{1}{4}, \frac{1}{2}, \frac{1}{4}]$ be the refinement mask of the piecewise linear function $\phi(x) = \max(1 - |x|, 0)$. Define $h_1 = [-\frac{1}{4}, \frac{1}{2}, -\frac{1}{4}]$ and $h_2 = [\frac{\sqrt{2}}{4}, 0, -\frac{\sqrt{2}}{4}]$. Then h_0, h_1 and h_2 satisfy (7) and (8). Hence, the system $X(\Psi)$ where $\Psi = \{\psi_1, \psi_2\}$ defined in (4) by using h_1, h_2 and ϕ is a tight frame of $L_2(\mathbb{R})$ (see Figure 1).

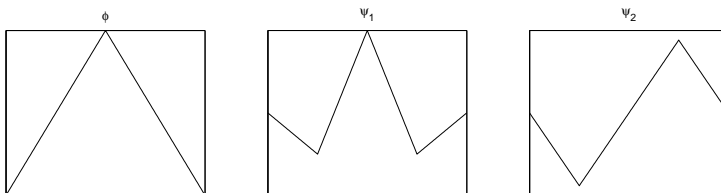


Figure 1. Piecewise linear refinable spline and corresponding framelets.

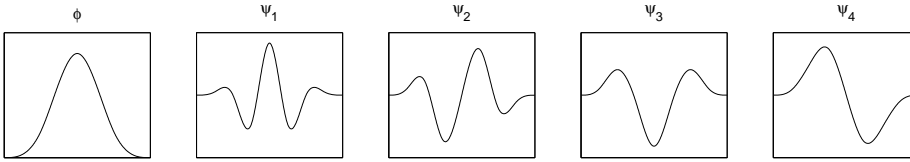


Figure 2. Piecewise cubic refinable B-spline and corresponding framelets.

Example 2. Let $h_0 = [\frac{1}{16}, \frac{1}{4}, \frac{3}{8}, \frac{1}{4}, \frac{1}{16}]$ be the refinement mask of ϕ . Then ϕ is the piecewise cubic B-spline. Define h_1, h_2, h_3, h_4 as follows:

$$h_1 = [\frac{1}{16}, -\frac{1}{4}, \frac{3}{8}, -\frac{1}{4}, \frac{1}{16}], \quad h_2 = [-\frac{1}{8}, \frac{1}{4}, 0, -\frac{1}{4}, \frac{1}{8}],$$

$$h_3 = [\frac{\sqrt{6}}{16}, 0, -\frac{\sqrt{6}}{8}, 0, \frac{\sqrt{6}}{16}], \quad h_4 = [-\frac{1}{8}, -\frac{1}{4}, 0, \frac{1}{4}, \frac{1}{8}].$$

Then h_0, h_1, h_2, h_3, h_4 satisfy (7) and (6) and hence the system $X(\Psi)$ where $\Psi = \{\psi_1, \psi_2, \psi_3, \psi_4\}$ defined in (4) by h_1, h_2, h_3, h_4 and ϕ is a tight frame of $L_2(\mathbb{R})$ (see Figure 2).

An advantage of the tight wavelet frames derived from the UEP is that those systems have fast decomposition and reconstruction algorithms, just as the orthonormal wavelet bases of [35]. The detailed discussions of decomposition and reconstruction algorithms are given in [40].

2.2.2. Pseudo-spline tight wavelet frames. The approximation order of the truncated tight wavelet frame generated by the UEP depends on the flatness of the Fourier transform of the underlying refinable function ϕ with refinement mask h_0 at the origin. More precisely, the order of the approximation of $Q_j f$, where the operator Q_j is the truncation operator defined in (10), to a sufficient smooth function f cannot exceed the order of the zero of $1 - |\widehat{\phi}|^2$ at the origin (see [40] for details) which is the same as the order of zeros of $1 - |\widehat{h}_0|^2$ at the origin.

Recall that the operator Q_j provides approximation order m_1 , if for all f in the Sobolev space $W_2^{m_1}(\mathbb{R})$

$$\|f - Q_j f\|_{L_2(\mathbb{R})} = O(2^{-nm_1}).$$

As shown in [40], the approximation order of $Q_j f$ depends on the order of the zero of $1 - |\widehat{h}_0|^2$ at the origin. In fact, if $1 - |\widehat{h}_0|^2 = O(|\cdot|^{m_2})$ at the origin, then $m_1 = \min\{m_0, m_2\}$ (see [40] for details), where m_0 is the order of Strang-Fix condition that ϕ satisfies. Recall that a function ϕ satisfies the Strang-Fix condition of order m_0 if $\widehat{\phi}(0) \neq 0, \widehat{\phi}^{(j)}(2\pi k) = 0, j = 0, 1, 2, \dots, m_0 - 1, k \in \mathbb{Z} \setminus \{0\}$.

Furthermore, it is easy to see from the UEP condition (6) and the definition of Ψ that there is at least one of framelets in Ψ that has the vanishing moment of the half of the order of zero of $1 - |\widehat{\phi}|^2$ at the origin. Recall that the order

of the vanishing moment of a function is the order of the zero of its Fourier transform at the origin. However, for an arbitrary refinable spline ϕ , the order of the zero of $1 - |\widehat{\phi}|^2$ at the origin cannot exceed 2. This means that $Q_j f$ cannot have approximation order more than 2. (In fact, the approximation order $Q_j f$ for any refinable spline is exact 2.) Furthermore, there is at least one framelet among Ψ constructed via the UEP from a refinable spline only has the vanishing moment of order 1. It is clear that the high order of the approximation of Q_j gives good approximations for smooth functions and the high order of vanishing moment of the framelets gives good sparse approximations for piecewise smooth functions. Hence, in order to have a good tight wavelet system, we need to have refinable functions whose Fourier transform are very flat at the origin. This leads to the introduction of the pseudo-splines in [40, 42]. The results given here are mainly from [42].

Pseudo-splines are defined in terms of their refinement masks. It starts with the simple identity, for given nonnegative integers l and m with $l \leq m - 1$,

$$1 = (\cos^2(\xi/2) + \sin^2(\xi/2))^{m+l}. \tag{11}$$

The refinement masks of pseudo-splines are defined by the summation of the first $l + 1$ terms of the binomial expansion of (11). In particular, the refinement mask of a *pseudo-spline of Type I with order (m, l)* is given by, for $\xi \in [-\pi, \pi]$,

$$|_1\widehat{a}(\xi)|^2 := |_1\widehat{a}_{(m,l)}(\xi)|^2 := \cos^{2m}(\xi/2) \sum_{j=0}^l \binom{m+l}{j} \sin^{2j}(\xi/2) \cos^{2(l-j)}(\xi/2) \tag{12}$$

and the refinement mask of a *pseudo-spline of Type II with order (m, l)* is given by, for $\xi \in [-\pi, \pi]$,

$${}_2\widehat{a}(\xi) := {}_2\widehat{a}_{(m,l)}(\xi) := \cos^{2m}(\xi/2) \sum_{j=0}^l \binom{m+l}{j} \sin^{2j}(\xi/2) \cos^{2(l-j)}(\xi/2). \tag{13}$$

We note that the mask of Type I is obtained by taking the square root of the mask of Type II using the Fejér-Riesz lemma (see e.g. [36]), i.e. ${}_2\widehat{a}(\xi) = |_1\widehat{a}(\xi)|^2$. Type I and Type II were introduced and used in [40] and [42] respectively in their constructions of tight framelets. Furthermore, it was shown in [40, 42]. (see e.g. Theorem 3.10 [42]), when ϕ is a pseudo-spline of an arbitrary type with order (m, l) the order of zero of $1 - |h|^2$ at the origin is $2l + 2$.

The corresponding pseudo-splines can be defined in terms of their Fourier transforms, i.e.

$${}_k\widehat{\phi}(\xi) := \prod_{j=1}^{\infty} {}_k\widehat{a}(2^{-j}\xi), \quad k = 1, 2. \tag{14}$$

The pseudo-splines with order $(m, 0)$ for both types are B-splines. Recall that a B-spline with order m and its refinement mask are defined by

$$\widehat{B}_m(\xi) = e^{-ij\frac{\xi}{2}} \left(\frac{\sin(\xi/2)}{\xi/2} \right)^m \quad \text{and} \quad \widehat{a}(\xi) = e^{-ij\frac{\xi}{2}} \cos^m(\xi/2),$$

where $j = 0$ when m is even, $j = 1$ when m is odd. The pseudo-splines of Type I with order $(m, m-1)$ are the refinable functions with orthonormal shifts (called orthogonal refinable functions) given in [35]. The key step of construction of orthonormal wavelet systems is to derive orthogonal refinable functions. The pseudo-splines of Type II with order $(m, m-1)$ are the interpolatory refinable functions (which were first introduced in [43] and a systematic construction was given in [35]). Recall that a continuous function $\phi \in L_2(\mathbb{R})$ is interpolatory if $\phi(j) = \delta(j)$, $j \in \mathbb{Z}$, i.e. $\phi(0) = 1$, and $\phi(j) = 0$, for $j \neq 0$ (see e.g. [43]). The other pseudo-splines fill in the gap between the B-splines and orthogonal or interpolatory refinable functions.

A complete regularity analysis of the pseudo-splines is given in [42] through the analysis of the decay of the Fourier transform of pseudo-splines. For fixed m , since the value of the mask $|\widehat{k_a}(\xi)|$, for $k = 1, 2$ and $\xi \in \mathbb{R}$, increases with l and the length of the mask k_a also increases with l , we conclude that the decay rate of the Fourier transform of a pseudo-spline decreases with l and the support of the corresponding pseudo-spline increases with l . In particular, for fixed m , the pseudo-spline with order $(m, 0)$ has the highest order of smoothness with the shortest support, the pseudo-spline with order $(m, m-1)$ has the lowest order of smoothness with the largest support in the family. As mentioned above, when we move from B-splines to orthogonal or interpolatory refinable functions, we sacrifice the smoothness and short support of the B-splines to gain some other desirable properties, such as orthogonality or interpolatory property. What do we get from the pseudo-splines of the other orders? When we move from B-splines to pseudo-splines, we gain the sparse approximation power of the corresponding tight wavelet frame $X(\Psi)$ derived by the UEP, since the Fourier transform of the corresponding refinable functions becomes flat at the origin.

Next, we give a genetic construction of tight wavelet frame system from pseudo-splines. The construction is from [42] which is motivated from [30] and one of the constructions of [40]. The construction can be applied to any refinable function whose mask is a trigonometric polynomial and satisfies

$$|\widehat{h_0}|^2 + |\widehat{h_0}(\cdot + \pi)|^2 \leq 1. \quad (15)$$

Note that when (6) holds, (15) must hold. Hence, (15) is the necessary condition to apply the UEP.

Let $\phi \in L_2(\mathbb{R})$ be a compactly supported refinable function with its trigonometric polynomial refinement mask $\widehat{h_0}$ satisfying $\widehat{h_0}(0) = 1$ and (15). Let

$$\mathcal{A} = \frac{1}{2} \sqrt{1 - |\widehat{h_0}|^2 - |\widehat{h_0}(\cdot + \pi)|^2}.$$

Here the square root is derived via the Fejér-Riesz lemma. Hence, \mathcal{A} is a trigonometric polynomial. Define

$$\widehat{h_1}(\xi) := e^{-i\xi} \overline{\widehat{h_0}(\xi + \pi)}, \quad \widehat{h_2}(\xi) := \mathcal{A}(\xi) + e^{-i\xi} \mathcal{A}(-\xi) \quad \text{and} \quad \widehat{h_3}(\xi) := e^{-i\xi} \overline{\widehat{h_2}(\xi + \pi)}.$$

Let $\Psi := \{\psi_1, \psi_2, \psi_3\}$, where

$$\widehat{\psi}_j(\xi) := \widehat{h}_j(\xi/2)\widehat{\phi}(\xi/2), \quad j = 1, 2, 3. \tag{16}$$

Then $X(\Psi)$ is a tight frame for $L_2(\mathbb{R})$. Each generator in Ψ is compactly supported. Moreover, if the refinement masks h_0 is symmetry to the origin, which leads to the refinable function ϕ is symmetric to the origin, ψ_1 is symmetric about $\frac{1}{2}$, ψ_2 is symmetric about $\frac{1}{4}$ and ψ_3 is antisymmetric about $\frac{1}{4}$. Furthermore, it was shown in [42, 61] that $X(\psi_1)$ forms a Riesz basis for $L_2(\mathbb{R})$ when ϕ is a pseudo-spline. On the other hand, since \widehat{h}_2 and \widehat{h}_3 have zeros at both 0 and π , one can check easily that neither the shifts of ψ_2 nor those of ψ_3 can form a Riesz system. Hence, $X(\psi_2)$ and $X(\psi_3)$ cannot form a Riesz basis for $L_2(\mathbb{R})$. In this case, the redundancy provided by the systems $X(\psi_2)$ and $X(\psi_3)$ moves Riesz system $X(\psi_1)$ to a self dual tight frame system. When ϕ is the pseudo-spline of Type I with order $(m, m-1)$, ϕ and its integer shifts form an orthonormal system and its masks satisfies $|\widehat{h}_0|^2 + |\widehat{h}_0(\cdot + \pi)|^2 = 1$. In this case, the above construction leads to that $\psi_2 = \psi_3 = 0$ and $X(\psi_1)$ is an orthonormal basis of $L_2(\mathbb{R})$ which is the compactly supported orthonormal wavelet construction of [35].

Since the order of zero of $1 - |h_0|^2$ at the origin is $2l + 2$ when ϕ is a pseudo-spline of an arbitrary type with order (m, l) , the approximation order of $Q_j f$ corresponding to $X(\Psi)$ constructed above is $2l + 2$ and the order of vanishing moments is $l + 1$. Next, we give an example from [42].

Example 3. Let \widehat{a} to be the mask of pseudo-spline of Type II with order $(3, 1)$ i.e.

$$\widehat{a}(\xi) = \cos^6(\xi/2)(1 + 3 \sin^2(\xi/2)).$$

We define

$$\widehat{b}_1(\xi) := e^{-i\xi} \overline{\widehat{a}(\xi + \pi)} = e^{-i\xi} \sin^6(\xi/2)(1 + 3 \cos^2(\xi/2)),$$

$$\widehat{b}_2(\xi) := \mathcal{A}(\xi) + e^{-i\xi} \mathcal{A}(-\xi) \quad \text{and} \quad \widehat{b}_3(\xi) := e^{-i\xi} \mathcal{A}(-\xi) - \mathcal{A}(\xi),$$

where

$$\begin{aligned} \mathcal{A} = \frac{1}{2} \bigg(& 0.00123930398199e^{-4i\xi} + 0.00139868605052e^{-2i\xi} - 0.22813823298962 \\ & + 0.44712319189971e^{2i\xi} - 0.22162294894260e^{4i\xi} \bigg). \end{aligned}$$

The graphs of Ψ are given by (b)-(d) in Figure 3. The tight frame system has approximation order 4.

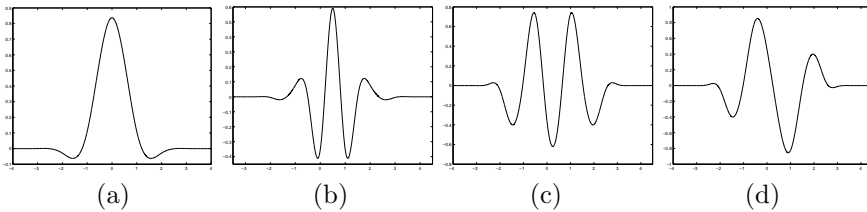


Figure 3. (a) is the pseudo-spline of Type II with order (3, 1) and (b)-(d) are the corresponding (anti)symmetric tight framelets.

2.3. Other extension principles. Since the publication of [79] in 1997, there are many generalizations of the unitary extension principle. Here, we briefly review some of them. The interested reader should consult the references mentioned below for the details.

We start with the oblique extension principle of [31, 40]. As mentioned before, when the unitary extension principle is applied to construct tight wavelet frames from refinable spline functions, the approximation order of the corresponding truncated wavelet system cannot exceed 2; and there is at least one framelet that has its vanishing moment to be 1. To obtain spline tight wavelet systems with better approximation power, the unitary extension principle was extended to oblique extension principle in [31, 40] by introducing a 2π periodic function Θ . The oblique extension principle says that in order to find tight wavelet frame system $X(\Psi)$ from a given refinable function with its refinement mask h_0 , one needs to find the 2π periodic function Θ which is non-negative, essentially bounded, continues at the origin with $\Theta(0) = 1$, and the wavelet masks h_1, \dots, h_r , such that the following two equalities hold a.e. $\omega \in \mathbb{R}$:

$$|h_0(\omega)|^2\Theta(2\omega) + \sum_{\ell=1}^r |\widehat{h}_\ell(\omega)|^2 = \Theta(\omega); \quad h_0(\omega)\overline{\widehat{h}_0(\omega + \pi)}\Theta(2\omega) + \sum_{\ell=1}^r \widehat{h}_\ell(\omega)\overline{\widehat{h}_\ell(\omega + \pi)} = 0. \tag{17}$$

We note that the unitary extension principle can be viewed as a special case of the oblique extension principle by taking Θ to be 1. When the Fourier transform of the refinable function used is not flat at the origin, one can chose a proper Θ which is flat at the origin, so that the resulting framelets have high order of the vanishing moment and the truncated tight wavelet system has a high approximation order. The detailed discussions can be found in [40]. This leads to many nice examples of spline tight wavelet frames with high order of vanishing moment and approximation power in [31, 39, 40].

In order to get an arbitrary high approximation order of the truncated tight frame system, one has to use the non-stationary wavelets, i.e. the masks used at the different level are different, as suggested by [62]. It starts with a non-stationary multiresolution analysis that has the different refinable function and refinement mask at the different level. The non-stationary version of the unitary extension principle is established and the corresponding wavelet masks

are obtained in [62]. The different level has a different set of wavelet masks, since the refinement mask at the different level is different. By a proper choice of the masks, symmetric C^∞ real-valued tight wavelet frames in $L_2(\mathbb{R})$ with compact support and the spectral frame approximation order are obtained in [62].

More recently, in order to get a fast flexible decomposition strategy adapted to the data that give a sparse approximation of the underlying function, a concept of an adaptive MRA (AMRA) structure which is a variant of the classical MRA structure is introduced in [60]. For this general case of affine systems, a unitary extension principle for filter design is derived and then applied to the directional representation system of shearlets. This, in turn, leads to the unitary extension principle for shearlets which further leads to a comprehensive theory for fast decomposition algorithms associated with 2D as well as 3D-shearlet systems which encompasses tight shearlet frame with spatially compactly supported generators within such an AMRA structure. Furthermore, shearlet-like systems associated with parabolic scaling and unimodular matrices optimally close to rotation are studied within the framework in [60].

Finally, both the unitary extension principle and the oblique extension principle can be generalized to a bi-frame setting which is called the mixed extension principle. The interested reader should consult [31, 40, 80] where the mixed extension principle is given in the multivariate setting with arbitrary integer dilation matrix.

Furthermore, the mixed extension principle for $L_2(\mathbb{R}^d)$ of [80] is generalized to a pair of dual Sobolev spaces $H^s(\mathbb{R}^d)$ and $H^{-s}(\mathbb{R}^d)$ in [63]. Here we briefly discuss the univariate case and encourage the reader to consult [63] for details of multivariate case. The mixed extension principle is given to ensure that a pair of systems $X^s(\phi; \psi_1, \dots, \psi_r)$ and $X^{-s}(\tilde{\phi}; \tilde{\psi}_1, \dots, \tilde{\psi}_r)$ forms a dual wavelet frame pair in the corresponding dual Sobolev spaces $H^s(\mathbb{R})$ and $H^{-s}(\mathbb{R})$. Recall that the system $X^s(\psi, \Psi) := X^s(\phi; \psi_1, \dots, \psi_r)$ is the homogenous wavelet system generated by ϕ and $\Psi := \{\psi_1, \dots, \psi_r\}$, i.e.,

$$X^s(\psi, \Psi) := \{\phi(\cdot - k) : k \in \mathbb{Z}^d\} \cup \{2^{j(d/2-s)}\psi_\ell(2^j \cdot -k) : j \in \mathbb{N}_0, k \in \mathbb{Z}^d, 1 \leq \ell \leq r\}.$$

In this general mixed extension principle, the regularity and vanishing moment are shared by two different systems in the dual pair separately instead of requiring both systems in the dual pair to have certain order of regularity and vanishing moment. For $s > 0$, the regularity of $\phi, \psi_1, \dots, \psi_r$, and the vanishing moments of $\tilde{\psi}_1, \dots, \tilde{\psi}_r$ are required, while allowing $\phi, \psi_1, \dots, \psi_r$ to be tempered distributions instead of in $L_2(\mathbb{R})$ and $\tilde{\psi}_1, \dots, \tilde{\psi}_r$ to have no vanishing moments. This implies that the systems $X^s(\phi; \psi_1, \dots, \psi_r)$ and $X^{-s}(\tilde{\phi}; \tilde{\psi}_1, \dots, \tilde{\psi}_r)$ are not necessary to be able to be normalized into a frame of $L_2(\mathbb{R})$. This leads to simple constructions of frames in an arbitrary given Sobolev space. For example, it was shown in [63] that $\{2^{j(1/2-s)}B_m(2^j \cdot -k) : j \in \mathbb{N}_0, k \in \mathbb{Z}\}$ is a wavelet frame in $H^s(\mathbb{R})$ for any $0 < s < m - 1/2$, where B_m is the B -spline of order m . This construction is also applied to multivariate box splines to obtain wavelet

frames with small supports while it is well known that it is hard to construct nonseparable multivariate wavelet frames with small supports if the traditional approach is taken, i.e. normalizing a frame in $L_2(\mathbb{R})$ to a frame in Sobolev space, since it is hard to construct small support wavelet frames in $L_2(\mathbb{R}^d)$ in general. This general mixed extension principle also naturally leads to a characterization of the Sobolev norm of a function in terms of weighted norm of its wavelet coefficient sequence (decomposition sequence) without requiring that dual wavelet frames should be in $L_2(\mathbb{R})$, which is quite different to other approaches in the literature (see e.g. [5, 6, 65, 71]). Furthermore, by applying this general mixed extension principle, a characterization for a pair of systems $X^s(\phi; \psi_1, \dots, \psi_r)$ and $X^{-s}(\check{\phi}; \check{\psi}_1, \dots, \check{\psi}_r)$ in Sobolev spaces $H^s(\mathbb{R})$ and $H^{-s}(\mathbb{R})$ that forms a pair of dual Riesz bases is obtained. This characterization, for example, leads to a proof of the fact that all interpolatory wavelet systems defined in [45] generated by an interpolatory refinable function $\phi \in H^s(\mathbb{R})$ with $s > 1/2$ are Riesz bases of the Sobolev space $H^s(\mathbb{R})$.

3. Frame Based Image Restoration

Image restoration is often formulated as an inverse problem. For the simplicity of the notation, we denote images as vectors in \mathbb{R}^n by concatenating their columns. The objective is to find the unknown true image $u \in \mathbb{R}^n$ from an observed image (or measurements) $b \in \mathbb{R}^\ell$ defined by

$$b = Au + \eta, \quad (18)$$

where η is a white Gaussian noise with variance σ^2 , and $A \in \mathbb{R}^{\ell \times n}$ is a linear operator, typically a convolution operator in image deconvolution, a projection in image inpainting and the identity in image denoising.

This section is devoted to frame based image restorations. The frame, especially tight wavelet frame, based image restoration has been developed very fast in the past decade, since the redundancy makes algorithms robust and stable.

Tight frames are redundant system in \mathbb{R}^n generated by tight wavelet frames. In particular, for given $W \in \mathbb{R}^{m \times n}$ (with $m \geq n$), the rows of W form a *tight frame* in \mathbb{R}^n if W satisfies $W^T W = I$, where I is the identity matrix. Thus, for every vector $u \in \mathbb{R}^n$,

$$u = W^T(Wu). \quad (19)$$

The components of the vector Wu are called the canonical coefficients representing u . The matrix W normally generated from the decomposition algorithm of a tight wavelet frame system by using the corresponding masks. The details in the construction of W from a given wavelet tight frame system can be found in, for example, [8, 9, 10, 11, 20, 23, 24, 25, 28].

Since tight frame systems are redundant systems, the mapping from the image u to its coefficients is not one-to-one, i.e., the representation of u in the frame domain is not unique. Therefore, there are three formulations for

the sparse approximation of the underlying images, namely analysis based approach, synthesis based approach and balanced approach. The analysis based approach was first proposed in [49, 85]. In that approach, we assume that the analyzed coefficient vector Wu can be sparsely approximated, and it is usually formulated as a minimization problem involving a penalty on the term $\|Wu\|_1$. The synthesis based approach was first introduced in [41, 50, 51, 52, 53]. In that approach, the underlying image u is assumed to be synthesized from a sparse coefficient vector α with $u = W^T\alpha$, and it is usually formulated as a minimization problem involving a penalty on the term $\|\alpha\|_1$. The balanced approach was first used in [24, 25] for high resolution image reconstruction. It was further developed for various image restoration in [8, 9, 10, 11, 20, 23, 28]. In that approach, the underlying image u is assumed to be synthesized from some sparse coefficient vector α with $u = W^T\alpha$, and it is usually formulated as a minimization problem involving a penalty on the term $\|\alpha\|_1$ and the distance of the α to the range of W . Although the synthesis based, analysis based and balanced approaches are developed independently in the literature, the balanced approach can be motivated from our desire to balance the analysis and synthesis based approaches.

Next, we give the exact models of the above three approaches. Before that, we set up some notation. For any $x \in \mathbb{R}^n$, $\|x\|_p = \left(\sum_{j=1}^n |x_j|^p\right)^{1/p}$, $1 \leq p < \infty$. For simplicity, we write $\|x\| = \|x\|_2$. Let $\|x\|_D$ denote the D -norm, where D is a symmetric positive definite matrix, defined by $\|x\|_D = \sqrt{x^T D x}$. For any real symmetric matrix H_1 , $\lambda_{\max}(H_1)$ denotes the maximum eigenvalue of H_1 . For any $m \times n$ real matrices A , $\|A\|_2 = \sqrt{\lambda_{\max}(A^T A)}$.

These three approaches can be formulated as the following minimization problem:

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{2} \|AW^T\alpha - b\|_D^2 + \frac{\kappa}{2} \|(I - WW^T)\alpha\|^2 + \|\text{diag}(\lambda)\alpha\|_1, \quad (20)$$

where $0 \leq \kappa \leq \infty$, λ is a given positively weighted vector, and D is a given symmetric positive definite matrix.

When $0 < \kappa < \infty$, the problem (20) is called balanced approach.

When $\kappa = 0$, the problem (20) is reduced to a synthesis based approach:

$$\min_{\alpha \in \mathbb{R}^m} \frac{1}{2} \|AW^T\alpha - b\|_D^2 + \|\text{diag}(\lambda)\alpha\|_1. \quad (21)$$

On the other extreme, when $\kappa = \infty$, the problem (20) is reduced to an analysis based approach. To see this, we note that the distance $\|(I - WW^T)\alpha\|$ must be 0 when $\kappa = \infty$. This implies that α is in the range of W , i.e., $\alpha = Wu$ for some $u \in \mathbb{R}^n$, so we can rewrite (20) as

$$\min_{\alpha \in \text{Range}(W)} \frac{1}{2} \|AW^T\alpha - b\|_D^2 + \|\text{diag}(\lambda)\alpha\|_1 = \min_{u \in \mathbb{R}^n} \frac{1}{2} \|Au - b\|_D^2 + \|\text{diag}(\lambda)Wu\|_1 \quad (22)$$

Problem (22) is the analysis based approach. It is clear that when $0 < \kappa < \infty$, (20) balances between (21) and (22), hence is called a balanced approach.

We note that when the rows of W form an orthonormal basis, instead of being a redundant tight frame, the above three approaches are exactly the same, since in this case, $WW^T = I$. However, for redundant tight frame system W , the analysis based, synthesis based and balanced approaches cannot be derived from one another. In fact, it was observed in, for example, [29, 48] that there is a gap between the analysis based and synthesis based approaches. Both of them have their own favorable data sets and applications. In general, it is hard to draw definitive conclusions on which approach is better without specifying the applications and data sets. We further note that the ℓ_1 -minimization problem arising from compressed sensing is akin to the synthesis based approach in nature. On the other hand, the TV-norm minimization problem in imaging restoration is, in many cases, an analysis based approach. For frame based image restoration, numerical simulation results in [15] show that the analysis based approach tends to generate smoother images. This is because the coefficient Wu is quite often linked to the smoothness of the underlying image [5, 6, 56, 63, 65]. However, the synthesis based approach tends to explore more on the sparse representation of the underlying solution in terms of the given frame system by utilizing the redundancy. This enhances and sharpens edges, although it may introduce some artifacts as shown in [10]. The balanced approach bridges the analysis based and synthesis based approaches in image restoration and it balances the smoothness and the sparsity provided by frames as shown in [8, 9, 10, 20, 23, 24, 25, 28].

For the synthesis based approach, the proximal forward and backward splitting algorithm was used in [38, 41, 50, 51, 52, 53]. The accelerated proximal gradient algorithms of [83] can be applied to get a fast algorithm for the synthesis based approach.

For the analysis approach, the coordinate descent method is used in [49, 85]. The split Bregman iteration is used to develop a fast algorithm for the analysis based approach in frame based image restoration in [15]. The numerical simulation shows that the split Bregman is efficient for image deblurring, decomposition, denoise, and inpainting. The split Bregman iteration was first proposed in [55] which was shown to be powerful in [55, 88] when it is applied to various PDE based image restoration approaches, e.g., ROF and nonlocal PDE models. The convergence analysis of the split Bregman was given in [15].

For the balanced approach in frame based image restoration, the model and algorithm were first developed in [23, 24, 25, 28]. The balanced approach was reformulated as the proximal forward-backward splitting algorithm in [8, 9, 10, 20]. The balanced approach gives satisfactory simulation results, as shown in [8, 9, 10, 20, 23, 24, 25, 28]. Recently, fast algorithms for the balanced approach in frame based image restoration whose convergence speeds are much faster than those of the proximal forward-backward splitting algorithm are developed in [83]. The accelerated proximal gradient algorithms proposed in [83] are based on

and extended from several variants of accelerated proximal gradient algorithms that were studied in [1, 73, 74, 75, 76, 86]. These accelerated proximal gradient algorithms have an attractive iteration complexity of $O(1/\sqrt{\epsilon})$ for achieving ϵ -optimality. Also these accelerated proximal gradient algorithms are simple and use only the soft-thresholding operator, just like algorithms such as the linearized Bregman iteration, the split Bregman iteration and the proximal forward-backward splitting algorithm.

Recently, the linearized Bregman iteration is applied to develop a fast algorithm for frame based image deblurring in [14], which converges to the minimizer of the follows minimization problem:

$$\min_{\alpha \in \mathbb{R}^m} \left\{ \frac{\kappa}{2} \|\alpha\|^2 + \|\text{diag}(\lambda)\alpha\|_1 : AW^T\alpha = b \right\}, \quad (23)$$

when A is invertible. Furthermore, it converges to the minimizer of $\min_{\alpha \in \mathbb{R}^m} \{\|\alpha\|_1 : AW^T\alpha = b\}$ as $\lambda \rightarrow \infty$, where, for the simplicity, the each entry of the vector λ is set to be the same (see [14]). Hence, linearized Bregman is used here to solve a variation of the synthesis based approach. The linearized Bregman iteration was first proposed to solve the ℓ_1 -minimization problems in compressed sensing by [87] and it was made efficient in [77]. The convergence analysis of linearized Bregman iteration was given in [12, 13]. It was then used in the nuclear norm minimization in matrix completion by [7]. The linearized Bregman can be re-formulated as the Uzawa's algorithm as shown in [7].

A simple computation of [77] shows that (23) is equivalent to

$$\min_{\alpha \in \mathbb{R}^m} \left\{ \frac{\kappa}{2} \|(I - WW^T)\alpha\|^2 + \|\text{diag}(\lambda)\alpha\|_1 : AW^T\alpha = b \right\},$$

when A is invertible. This looks like a variation of balanced approach. However, when the large parameter vector λ is chosen which happens when one applies linearized Bregman, it is more close to a variation of the synthesis based approach. For the case that A is not invertible, the detailed discussions also given in [77].

The advantages of Bregman iterations (either linearized Bregman or split Bregman iterations) in frame based image restorations are that big coefficients come back at first after few iterations and stay. This is, in particular, important in image deblurring, since the big wavelet frame coefficients contain information of edges and features of images. The main goal of deblurring is to restore the blurred edges and features. Although neither the synthesis nor analysis based approach is the focus of this paper, we still give an example of blind deblurring using analysis base approach with split Bregman iterations at the end of this paper. The frame based blind deblurring has been investigated extensively in [16, 17, 18, 19].

The formulation of (20) can also be extended to image restoration of two-layered images [15, 49, 85]. Real images usually have two layers, referring to cartoons (the piecewise smooth part of the image) and textures (the oscillating

pattern part of the image). Different layers usually have sparse approximations under different tight frame systems. Therefore, these two different layers should be considered separately. One natural idea is to use two tight frame systems that can sparsely represent cartoons and textures separately. The corresponding image restoration problem can be formulated as the following ℓ_1 -minimization problem:

$$\min_{\alpha_1, \alpha_2} \frac{1}{2} \|A \left(\sum_{i=1}^2 W_i^T \alpha_i \right) - b\|_D^2 + \sum_{i=1}^2 \frac{\kappa_i}{2} \|(I - W_i W_i^T) \alpha_i\|^2 + \sum_{i=1}^2 \|\text{diag}(\lambda)_i \alpha_i\|_1, \quad (24)$$

where, for $i = 1, 2$, $W_i^T W_i = I$, $\kappa_i > 0$, λ_i is a given positive weight vector, and D is a given symmetric positive definite matrix.

In the rest of the paper, we will focus on the balanced approach in frame based image restorations. For those who are interested in the synthesis and analysis based approach for frame based image restorations and the linearized Bregman and split Bregman iterations should consult the literature mentioned above for details.

3.1. Balanced approach for image inpainting. The balanced approach for frame based image restorations was first developed in [24, 25, 28] for the high resolution image reconstruction from a few low resolution images. The problem of high resolution image construction is converted to the problem of filling the miss wavelet frame coefficients, i.e. inpainting a wavelet frame transform domain, by designing a proper wavelet tight frame in [24, 25, 28]. The ideas of [24, 25] is used in [27] to develop balanced approach for frame based image inpainting (in pixel domain) whose complete analysis of convergence and optimal properties of the solution are given in [10]. Analysis of the convergence and optimal properties of the solutions of algorithms in [24, 25, 28] is given in [8, 20, 23]. In this section, we use image inpainting as an example to illustrate how the ideas of the balanced approach are formed and developed.

The mathematical model for image inpainting can be stated as follows. We will denote images as vectors in \mathbb{R}^n by concatenating their columns. Let the original image u be defined on the domain $\Omega = \{1, 2, \dots, n\}$ and the nonempty set $\Lambda \subsetneq \Omega$ be the given observed region. Then the observed (incomplete) image b is

$$b(i) = \begin{cases} u(i) + \eta(i), & i \in \Lambda, \\ \text{arbitrary}, & i \in \Omega \setminus \Lambda, \end{cases} \quad (25)$$

where $\eta(i)$ is the noise. The goal is to find u from b . When $\eta(i) = 0$ for all $i \in \Lambda$, we require that $u(i) = b(i)$ and u is just the solution of an interpolation problem. Otherwise, we seek a smooth solution u that satisfies $|u(i) - b(i)| \leq \eta(i)$ for all $i \in \Lambda$. In both cases, variational approaches will penalize some cost functionals (which normally are weighted function norms of the underlying solution) to control the roughness of the solution, see for instance [3, 26].

The image inpainting is to recover data by interpolation. There are many interpolation schemes available, e.g., spline interpolation, but majority of them are only good for smooth functions. Images are either piecewise smooth function or texture which do not have the required globe smoothness to provide a good approximation of underlying solutions. The major challenge in image inpainting is to keep the features, e.g. edges of images which many of those available interpolation algorithms cannot preserve. Furthermore, since images are contaminated by noises, the algorithms should have a building in denoising component.

The simple idea of the balanced approach for frame based image inpainting comes as follows: one may use any simple interpolation scheme to interpolate the given data that leads to an inpainted image. The edges might be blurred in this inpainted image. One of the simplest ways to sharpen the image is to throw out small coefficients under a tight wavelet frame transform. The deletion of small wavelet frame coefficients not only sharpens edges but also removes noises. When it is reconstruct back to image domain, it will not interpolate the data anymore, the simplest way to make it interpolate the given data is to put the given data back. One may iterate this process till convergence.

To be precise, let \mathcal{P}_Λ be the diagonal matrix with diagonal entries 1 for the indices in Λ and 0 otherwise. Starting with the initial guess u_0 , the iteration is

$$u_{k+1} = \mathcal{P}_\Lambda b + (\mathcal{I} - \mathcal{P}_\Lambda)W^T \mathcal{T}_\lambda(Wu_k).$$

Here

$$\mathcal{T}_\lambda([\beta_1, \beta_2, \dots, \beta_m]^T) \equiv [t_{\lambda_1}(\beta_1), t_{\lambda_2}(\beta_2), \dots, t_{\lambda_m}(\beta_m)]^T \quad (26)$$

with $\lambda = [\lambda_1, \lambda_2, \dots, \lambda_m]^T$, and $t_{\lambda_i}(\cdot)$ is the soft-thresholding function [44]:

$$t_{\lambda_i}(\beta_i) \equiv \begin{cases} \text{sgn}(\beta_i)(|\beta_i| - \lambda_i), & \text{if } |\beta_i| > \lambda_i, \\ 0, & \text{if } |\beta_i| \leq \lambda_i. \end{cases} \quad (27)$$

Note that by using the soft-thresholding instead of the hard-thresholding which is traditionally used to sharpen edges, we reduces artifacts and obtain the desire minimization property in each iteration. Besides, the thresholding operator \mathcal{T}_λ also plays two other important roles, namely, removing noises in the image and perturbing the frame coefficients Wu_n so that information contained in the given region can permeate into the missing region.

Let the thresholding parameters be

$$\lambda = [\lambda_1, \lambda_2, \dots, \lambda_m]^T \quad (28)$$

where $\lambda_i > 0$ for $i = 1, \dots, m$. The whole algorithm is given as follows:

Algorithm 1.

- (i) Set an initial guess u_0 .
- (ii) Iterate on n until convergence:

$$u_{k+1} = \mathcal{P}_\Lambda b + (\mathcal{I} - \mathcal{P}_\Lambda)W^T \mathcal{T}_\lambda(Wu_k). \quad (29)$$

(iii) Let u^* to the output of Step (ii). If $\eta(i) = 0$ for all $i \in \Lambda$ in (25), we set \mathbf{u}^* to be the solution (to the interpolation problem); otherwise, since \mathcal{T}_λ can remove the noise, we set $u^\diamond = W^T \mathcal{T}_\lambda(Wu^*)$ to be the solution (to the inpainting-plus-denoising problem).

Algorithm 1 was first proposed in [27], whose complete analysis of its convergence was given in [10] by re-formulating Algorithm 1 as an iteration for minimizing a special functional. Indeed, it was shown in [10] that $\alpha_n \equiv \mathcal{T}_\lambda W u_n$ converges to $\alpha^* \equiv \mathcal{T}_\lambda W u^*$ which is a minimizer of

$$\min_{\alpha} \left\{ \frac{1}{2} \|\mathcal{P}_\Lambda(W^T \alpha) - \mathcal{P}_\Lambda b\|^2 + \frac{1}{2} \|(\mathcal{I} - WW^T)\alpha\|^2 + \|\text{diag}(\lambda)\alpha\|_1 \right\}. \quad (30)$$

The idea of proof is that the iteration deriving sequence α_n can be written as a proximal forward-backward splitting iteration. Recall that for any proper, convex, lower semi-continuous function φ which takes its values in $(-\infty, +\infty]$, its proximal operator is defined by

$$\text{prox}_{\varphi}(\mathbf{x}) \equiv \arg \min_{\mathbf{y}} \left\{ \frac{1}{2} \|\mathbf{x} - \mathbf{y}\|_2^2 + \varphi(\mathbf{y}) \right\}, \quad (31)$$

The proximal forward-backward splitting iteration for α_n can be derived by using Algorithm 1 as:

$$\alpha_{k+1} = \text{prox}_{F_1}(\alpha_k - \nabla F_2(\alpha_k)), \quad (32)$$

where

$$F_1(\alpha) = \|\text{diag}(\lambda)\alpha\|_1, \text{ and } F_2(\alpha) = \frac{1}{2} \|\mathcal{P}_\Lambda(W^T \alpha) - \mathcal{P}_\Lambda b\|^2 + \frac{1}{2} \|(\mathcal{I} - WW^T)\alpha\|^2. \quad (33)$$

It was shown in [34] that when F_1 with range $(-\infty, +\infty]$ is a proper, convex, lower semi-continuous function, and F_2 with range in \mathbb{R} is a proper, convex, differentiable function with a L -Lipschitz continuous gradient, i.e.

$$\|\nabla F_2(\alpha) - \nabla F_2(\beta)\| \leq L \|\alpha - \beta\|, \quad \forall \alpha, \beta \quad (34)$$

for some $L > 0$. Then for any initial guess u_0 , the proximal forward-backward splitting iteration

$$\alpha_{k+1} = \text{prox}_{F_1/L}(\alpha_k - \nabla F_2(\alpha_k)/L)$$

converges to a minimizer of:

$$\min_{\alpha} \{F_1(\alpha) + F_2(\alpha)\}, \quad (35)$$

It is not difficult to check that F_1 and F_2 defined in (33) satisfy the conditions needed here, and furthermore, F_2 has 1-Lipschitz continuous gradient, hence, α_n converges.

3.2. Role of the redundancy. Tight frames are different from orthonormal systems because tight frames are redundant. What does the redundancy bring us here? We start with a sort of philosophical point of views on the algorithm and then give some quantitative analysis on the error being reduced at each iteration. Assume that some blocks of pixels are missing in a given image and we like to solve the inpainting problem in the wavelet frame domain as mentioned before. Since the framelets used are compactly supported, the coefficients of those framelets whose supports fall in the missing blocks are missing and the coefficients of those framelets whose supports overlap with the missing blocks are inaccurate. The main step of Algorithm 1 perturbs the frame coefficients Wu_n by thresholding so that information contained in the available coefficients will permeate into the missing frame coefficients. Here, the redundancy is very important, since the available coefficients and its associated atoms in the the system contain information of the missing coefficients only if the system is redundant, as the atoms in an orthonormal basis are orthogonal to each other and do not contain information of other atoms in L_2 - sense.

While applying the thresholding operator on the frame coefficients is a very important step in Algorithm 1 in order to remove the noises and perturb the coefficients and sharpen the edges, it, however, also brings in new errors and artifacts. To explain how the numerical errors and artifacts introduced by the thresholding can be reduced by the redundancy of the system W , we take the computed solution u^* as an example. Similar analysis holds for the computation of each iteration. Our computed solution u^* that interpolates the given data satisfies

$$u^* = \mathcal{P}_\Lambda b + (\mathcal{I} - \mathcal{P}_\Lambda)W^T\mathcal{T}_\lambda Wu^*.$$

That is, on Λ , $W^T\mathcal{T}_\lambda Wu^*$ is replaced by b . But since $\mathcal{P}_\Lambda b = \mathcal{P}_\Lambda u^* = \mathcal{P}_\Lambda W^T Wu^*$, we are actually replacing $\mathcal{P}_\Lambda W^T\mathcal{T}_\lambda Wu^*$ by $\mathcal{P}_\Lambda W^T Wu^*$, which generates artifacts. Hence to reduce the artifacts, we require that the norm of

$$\mathcal{P}_\Lambda W^T Wu^* - \mathcal{P}_\Lambda W^T\mathcal{T}_\lambda Wu^* = \mathcal{P}_\Lambda W^T(Wu^* - \mathcal{T}_\lambda Wu^*)$$

to be small.

Clearly the smaller the norm of $W^T\mathbf{e} := W^T(Wu^* - \mathcal{T}_\lambda Wu^*)$ is, the smaller the artifact is. Note that the reconstruction operator W^T can eliminate the error components sitting in the kernel of W^T . In fact, since W^T projects all sequences down to the orthogonal complement of the kernel of W^T , which is the range of W , the component of \mathbf{e} in the kernel of W^T does not contribute. The redundant system reduces the errors as long as the component of \mathbf{e} in the kernel of W^T is not zero. Therefore, the larger is the kernel of W^T , the more redundant is the frame system. In other words, higher redundancy will lead to more error reduction in general. To increase the redundancy, we use undecimated tight wavelet frame system (i.e. no down sampling in the decomposition). In contrast, if W is not a redundant system but an orthonormal system, then the kernel of W^T is just $\{0\}$. In this case, $\|W^T\mathbf{e}\| = \|\mathbf{e}\|$.

3.3. Accelerated algorithm. In this section we introduce accelerated proximal gradient algorithms for solving (20), similar algorithm for (24) can be obtained easily. Let

$$F_1(\alpha) = \|\text{diag}(\lambda)\alpha\|_1, \text{ and } F_2(\alpha) = \frac{1}{2}\|AW^T\alpha - b\|_D^2 + \frac{\kappa}{2}\|(I - WW^T)\alpha\|_2^2. \quad (36)$$

Then, the proximal forward-backward splitting iteration

$$\alpha_{k+1} = \text{prox}_{F_1/L}(\alpha_k - \nabla F_2(\alpha_k)/L)$$

converges to a minimizer of:

$$\min_{\alpha} \{F_1(\alpha) + F_2(\alpha)\}, \quad (37)$$

Here, the gradient of F_2 is given by

$$\nabla F_2(\alpha) = WA^T D(AW^T\alpha - b) + \kappa(I - WW^T)\alpha. \quad (38)$$

It can be proven easily that F_1 and F_2 given here satisfy the conditions for the convergence of the the proximal forward-backward splitting iteration. This generalizes the inpainting algorithm given in Section 3.1 to algorithms for various image restoration problems. Although the original development of algorithms took a different path, this idea is used in the proof of the convergence of the balanced approach frame based algorithms given in [8, 9, 10, 11, 20, 23, 24, 25, 28]. The accelerated proximal gradient is obtained by adjusting the term $\alpha_n - \nabla F_2(\alpha_n)/L$ in the the proximal forward-backward splitting iteration. Next, we describe exactly the accelerated proximal gradient algorithm for solving (20).

Algorithm 2. For a given nonnegative vector λ , choose $\alpha_0 = \alpha_{-1} \in \mathbb{R}^m$, $t_0 = t_{-1} = 1$. For $k = 0, 1, 2, \dots$, generate α_{k+1} from α_k according to the following iteration:

- (i) Set $\beta_k = \alpha_k + \frac{t_{k-1}-1}{t_k}(\alpha_k - \alpha_{k-1})$.
- (ii) Set $g_k = \beta_k - \nabla F_2(\beta_k)/L$.
- (iii) Set $\alpha_{k+1} = \mathcal{T}_{\lambda/L}(g_k)$.
- (iv) Compute $t_{k+1} = \frac{1 + \sqrt{1 + 4(t_k)^2}}{2}$.

When $F_2(\alpha) = \frac{1}{2}\|AW^T\alpha - b\|_D^2$, Algorithm 2 leads to an efficient algorithm for the synthesis based approach for the frame based image restoration as a side produce.

When the accelerated proximal gradient algorithm with $t^k = 1$ for all k is applied to the problem (20), it is the proximal forward-backward splitting

algorithm developed in [8, 9, 10, 11, 20, 23, 24, 25, 28] for the balanced approach in frame based image restorations, and it is also the popular iterative shrinkage/thresholding algorithms [38, 41, 52, 53]. The iterative shrinkage/thresholding algorithms and the proximal forward-backward splitting algorithms have been developed and analyzed independently by many researchers. These algorithms only require gradient evaluations and soft-thresholding operations, so the computation at each iteration is very cheap. But, for any $\epsilon > 0$, these algorithms terminate in $O(L/\epsilon)$ iterations with an ϵ -optimal solution [1, 84]. Hence the sequence $\{\alpha_k\}$ converges slowly. On the other hand, the accelerated proximal gradient algorithm proposed here gets an ϵ -optimal solution in $O(\sqrt{L/\epsilon})$ iterations. Thus the algorithm accelerates the proximal forward-backward splitting algorithms used in [8, 9, 10, 11, 20, 23, 24, 25, 28] for the balanced approach in frame based image restorations. In fact, it was proven in [83] (see Theorem 2.1 of [83]) that for given ϵ ,

$$F_1(\alpha_k) + F_2(\alpha_k) - F_1(\alpha^*) - F_2(\alpha^*) \leq \epsilon \quad \text{whenever } k \geq C\sqrt{\frac{2L}{\epsilon}}. \quad (39)$$

where α^* is a minimizer of (20). Furthermore, the constant C is explicitly given in Theorem 2.1 [83]. Numerical simulations in [83] illustrate and verify that Algorithm 2 is very effective for frame based image inpainting, decomposition, denoising and deblurring.

3.4. Some simulation results. This section gives a few simulation results to show the effectiveness of the frame based image restorations. We omit the detailed discussions on the numerical simulations and the interested reader should consult the relevant references for the details.

Table 1: Numerical results for the accelerated proximal decent algorithm in solving (20) and (24) arising from image inpainting without noise (i.e., $\sigma = 0$ in (18)).

inpainting $\sigma = 0$	one system $\lambda = 0.03$			two systems $\lambda_1 = \lambda_2 = 0.01$		
	iter	psnr	time	iter	psnr	time
peppers256	22	33.69	3.38	29	33.66	6.39
goldhill256	24	32.21	3.79	32	32.09	7.10
boat256	23	30.99	3.63	29	30.87	6.51
camera256	23	30.13	3.58	29	30.44	6.47
bridge256	26	31.31	4.15	33	31.08	7.47
bowl256	23	34.38	3.53	35	36.02	7.66
barbara512	27	31.33	22.47	31	33.82	34.12
baboon512	26	29.12	22.23	32	29.10	35.79
fingerprint512	25	26.51	21.23	34	28.00	38.20
zebra512	25	28.47	20.93	33	29.32	36.43

Table 1 is from [83], that gives the numerical performance of the accelerated proximal gradient algorithm applied to the balanced approach (20) and (24) for the image inpainting problem. As indicated, the accelerated proximal gradient algorithm takes no more than 27 iterations and 25 seconds to solve the model (20) for all the images. For the model (24), the accelerated proximal gradient algorithm takes no more than 35 iterations and 40 seconds to solve all the problems. More simulation results on the balanced approach of frame based image restoration by using the accelerated proximal gradient algorithm can be found in [83]

The next two figures illustrate examples of the frame based blind deblurring via the analysis based approach by using split Bregman algorithm. The assumption is that the blurring is caused by a convolution and the convolution kernel is sparse at the space domain. The convolution kernel and the deblurring image are solved alternatively and iteratively. The interested reader should consult [19] for the details.

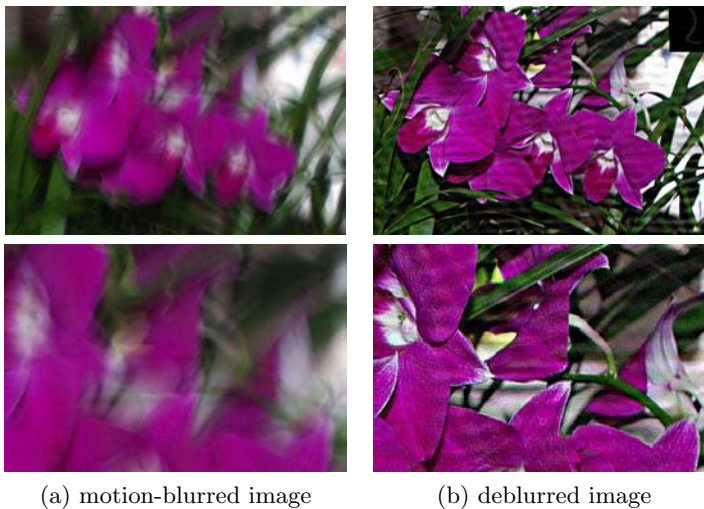


Figure 4. (a) Real motion-blurred image with one region after zooming in; (b) deblurred image with one region after zooming in by using blind motion deblur algorithm based on the analysis-based sparsity prior of images/kernels under wavelet tight frame system.

References

- [1] A. Beck and M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems, *SIAM J. Imaging Sci.* **2(1)** (2009), 183–202.
- [2] J. Benedetto and S. D. Li, The theory of multiresolution analysis frames and applications to filter banks, *Appl. Comp. Harmonic Anal.* **5** (1998), 389–427.

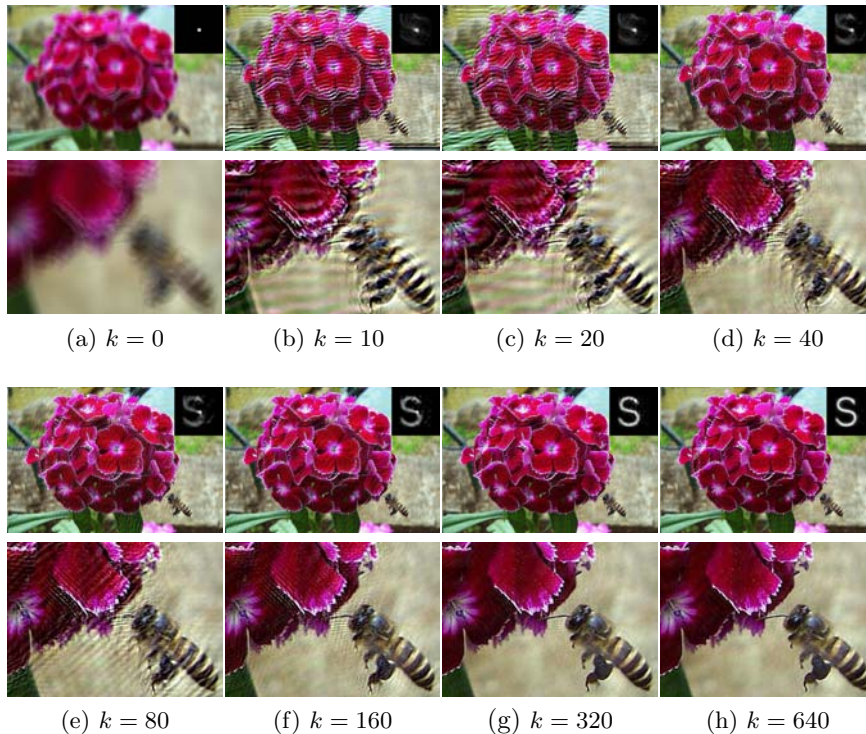


Figure 5. (a)–(h) are the intermediate results when de-blurring a synthesized motion-blurred image using the blind motion deblur algorithm based on the analysis-based sparsity prior of images/kernels under tight wavelet frame system, for the loop index $k = 0, 10, 20, 40, 80, 160, 320, 640$ respectively. De-blurred images are shown in the first row with the corresponding estimated blur kernels shown in their top right region respectively, and the zoomed regions are shown in the second row.

- [3] M. Bertalmío, G. Sapiro, V. Caselles and C. Ballester, Image inpainting, *in SIG-GRAPH 2000*, 417–424.
- [4] C. de Boor, R. DeVore and A. Ron, On the construction of multivariate (pre)wavelet, *Constr. Approx.* **9** (1993), 123–166.
- [5] L. Borup, R. Gribonval and M. Nielsen, Tight wavelet frames in Lebesgue and Sobolev spaces, *J. Funct. Spaces Appl.* **2** (2004), 227–252.
- [6] L. Borup, R. Gribonval and M. Nielsen, Bi-framelet systems with few vanishing moments characterize Besov spaces, *Appl. Comput. Harmon. Anal.* **17** (2004), 3–28.
- [7] J.-F. Cai, E. J. Candès and Z. Shen, A singular value thresholding algorithm for matrix completion, *SIAM J. Optimization* **20**(4) (2010), 1956–1982.
- [8] J.-F. Cai, R. H. Chan, L. Shen and Z. Shen, Restoration of chopped and noded images by framelets, *SIAM J. Sci. Comput.* **30**(3) (2008), 1205–1227.

- [9] J.-F. Cai, R. H. Chan, L. Shen and Z. Shen, Convergence analysis of tight framelet approach for missing data recovery, *Adv. Comput. Math.* **31** (2009), 87–113.
- [10] J.-F. Cai, R. H. Chan and Z. Shen, A framelet-based image inpainting algorithm, *Appl. Comput. Harmon. Anal.* **24** (2008), 131–149.
- [11] J.-F. Cai, R. H. Chan and Z. Shen, Simultaneous cartoon and texture inpainting, *Inverse Probl. Imaging*, to appear.
- [12] J.-F. Cai, S. Osher and Z. Shen, Convergence of the linearized Bregman iteration for ℓ_1 -norm Minimization, *Math. Comp.* **78** (2009), 2127–2136.
- [13] J.-F. Cai, S. Osher and Z. Shen, Linearized Bregman iterations for compressed sensing, *Math. Comp.* **78** (2009), 1515–1536.
- [14] J.-F. Cai, S. Osher and Z. Shen, Linearized Bregman iterations for frame-based image deblurring, *SIAM J. Imaging Sci.* **2**(1) (2009), 226–252.
- [15] J.-F. Cai, S. Osher and Z. Shen, Split Bregman methods and frame based image restoration, *Multiscale Model. Simul.* **8**(2) (2009), 337–369.
- [16] J.-F. Cai, H. Ji, C. Liu and Z. Shen, Blind motion deblurring using multiple images, *J. Comput. Phys.* **228** (14) (2009), 5057–5071.
- [17] J.-F. Cai, H. Ji, C. Liu and Z. Shen, High-quality curvelet-based motion deblurring using an image pair, *IEEE Conference on Computer Vision and Pattern Recognition CVPR* (2009).
- [18] J.-F. Cai, H. Ji, C. Liu and Z. Shen, Blind motion deblurring from a single image using sparse approximation, *IEEE Conference on Computer Vision and Pattern Recognition CVPR* (2009).
- [19] J.-F. Cai, H. Ji, C. Liu and Z. Shen, Framelet based blind motion deblurring from a single image, preprint, 2009.
- [20] J.-F. Cai and Z. Shen, Framelet based deconvolution, *Journal of Computational Mathematics* **28**(3) (2010), 289–308.
- [21] E. J. Candès and D. L. Donoho, New tight frames of curvelets and optimal representations of objects with piecewise- C^2 singularities, *Comm. Pure Appl. Math.* **57** (2002), 219–266.
- [22] E. J. Candès and D. L. Donoho, Continuous curvelet transform: II. Discretization of frames, *Appl. Comput. Harmon. Anal.* **19** (2005), 198–222.
- [23] A. Chai and Z. Shen, Deconvolution: A wavelet frame approach, *Numer. Math.* **106** (2007), 529–587.
- [24] R. H. Chan, T. F. Chan, L. Shen and Z. Shen, Wavelet algorithms for high-resolution image reconstruction, *SIAM J. Sci. Comput.* **24** (2003), 1408–1432.
- [25] R. H. Chan, S. D. Riemenschneider, L. Shen and Z. Shen, Tight frame: an efficient way for high-resolution image reconstruction, *Appl. Comput. Harmon. Anal.* **17** (2004), 91–115.
- [26] T. F. Chan and J. Shen, Variational image inpainting, *Comm. Pure Appl. Math.* **58** (2005), 579–619.

- [27] R. H. Chan, L. Shen and Z. Shen, A framelet-based approach for image inpainting, Research Report 2005-04(325), Department of Mathematics, The Chinese University of Hong Kong, 2005.
- [28] R. H. Chan, Z. Shen and T. Xia, A framelet algorithm for enhancing video stills, *Appl. Comput. Harmon. Anal.* **23** (2007), 153–170.
- [29] S. S. Chen, D. L. Donoho and M. A. Saunders, Atomic decomposition by basis pursuit, *SIAM Rev.* **43** (2001), pp. 129–159.
- [30] C. K. Chui and W. He, Compactly supported tight frames associated with refinable functions, *Appl. Comput. Harmon. Anal.* **8** (2000), 293–319.
- [31] C. K. Chui, W. He and J. Stöckler, Compactly supported tight and sibling frames with maximum vanishing moments, *Appl. Comput. Harmon. Anal.* **13** (2002), 224–262.
- [32] A. Cohen, I. Daubechies and J. C. Feauveau, Biorthogonal bases of compactly supported wavelets, *Comm. Pure Appl. Math.* **45** (1992), 485–560.
- [33] R. Coifman and D. Donoho. Translation-invariant de-noising, *Wavelet and Statistics, Springer Lecture Notes in Statistics* **103** (1994), 125–150, New York, Springer-Verlag.
- [34] P. L. Combettes and V. R. Wajs, Signal recovery by proximal forward-backward splitting, *Multiscale Model. Simul.* **4** (2005), 1168–1200.
- [35] I. Daubechies, Orthonormal bases of compactly supported wavelets, *Comm. Pure Appl. Math.* **41** (1988), 909–996.
- [36] I. Daubechies, *Ten Lectures on Wavelets*, SIAM, Philadelphia, PA, 1992.
- [37] I. Daubechies, A. Grossmann and Y. Meyer, Painless nonorthogonal expansions, *J. Math. Phys.* **27** (1986), 1271–1283.
- [38] I. Daubechies, M. De Friese and C. De Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Comm. Pure Appl. Math.* **57** (2004), 1413–1457.
- [39] I. Daubechies and B. Han, Pairs of dual wavelet frames from any two refinable functions, *Constr. Approx.* **20** (2004), 325–352.
- [40] I. Daubechies, B. Han, A. Ron and Z. Shen, Framelets: MRA-based constructions of wavelet frames, *Appl. Comput. Harmon. Anal.* **14** (2003), 1–46.
- [41] I. Daubechies, G. Teschke and L. Vese, Iteratively solving linear inverse problems under general convex constraints, *Inverse Probl. Imaging* **1** (2007), 29–46.
- [42] B. Dong and Z. Shen, Pseudo-splines, wavelets and framelets, *Appl. Comput. Harmon. Anal.* **22** (2007), 78–104.
- [43] S. Dubuc, Interpolation through an iterative scheme, *J. Math. Anal. Appl.* **114** (1986), 185–204.
- [44] D. L. Donoho, Denoising by soft-thresholding, *IEEE Trans. Inform. Theory* **41** (1995), 613–627.
- [45] D. L. Donoho, Interpolating wavelet transform, preprint, 1992.
- [46] D. L. Donoho, Ridge functions and orthonormal ridgelets, *J. Approx. Theory* **111(2)** (2001), 143–179.

- [47] R.J. Duffin and A.C. Schaeffer, A class of nonharmonic Fourier series, *Trans. Amer. Math. Soc.* **72** (1952), 147–158.
- [48] M. Elad, P. Milanfar and R. Rubinstein, Analysis versus synthesis in signal priors, *Inverse Problems* **23** (2007), 947–968.
- [49] M. Elad, J.-L. Starck, P. Querre and D. L. Donoho, Simultaneous cartoon and texture image inpainting using morphological component analysis, *Appl. Comput. Harmon. Anal.* **19** (2005), 340–358.
- [50] M. Fadili and J.-L. Starck, Sparse representations and bayesian image inpainting, *In Proc. SPARS'05*, Vol. I, Rennes, France, 2005.
- [51] M. Fadili, J.-L. Starck, and F. Murtagh, Inpainting and zooming using sparse representations, *The Computer Journal* **52** (2009), 64–79.
- [52] M. Figueiredo and R. Nowak, An EM algorithm for wavelet-based image restoration, *IEEE Trans. Image Proc.* **12** (2003), 906–916.
- [53] M. Figueiredo and R. Nowak, A bound optimization approach to wavelet-based image deconvolution, *IEEE Intern. Conf. on Image Processing-ICIP'05*, 2005.
- [54] M. Frazier, G. Garrigos, K. Wang and G. Weiss, A characterization of functions that generate wavelet and related expansion. *J. Fourier Anal. Appl.* **3** (1997), 883–906.
- [55] T. Goldstein and S. Osher, The split Bregman algorithm for L_1 regularized problems, *SIAM J. Imaging Sci.* **2(2)** (2009), 323–343.
- [56] R. Gribonval and M. Nielsen, On approximation with spline generated framelets. *Constr. Approx.* **20** (2004), 207–232.
- [57] G. Gripenberg, A necessary and sufficient condition for the existence of a father wavelet, *Studia Math.* **114** (1995), 207–226.
- [58] K. Gröchenig, *Foundations of Time-Frequency Analysis*, Birkhäuser, Boston, 2001.
- [59] B. Han, On dual tight wavelet frames, *Appl. Comput. Harmon. Anal.* **4** (1997), 380–413.
- [60] B. Han, G. Kutyniok and Z. Shen, Unitary extension principle for compactly supported shearlets, preprint, 2009.
- [61] B. Han and Z. Shen, Wavelets with short support, *SIMA J. Math. Anal.* **38** (2006), 530–556.
- [62] B. Han and Z. Shen, Compactly supported symmetric C^∞ wavelets with spectral approximation order, *SIAM J. Math. Anal.* **40** (2008), 905–938.
- [63] B. Han and Z. Shen, Dual wavelet frames and Riesz bases in Sobolev spaces, *Constr. Approx.* **29** (2009), 369–406.
- [64] E. Hernández and G. Weiss, *A First Course on Wavelets*, Studies in Advanced Mathematics, CRC Press (Boca Raton FL), 1996.
- [65] Y. Hur and A. Ron, CAPlets: wavelet representations without wavelets, preprint, 2005.
- [66] R. Jia and Z. Shen, Multiresolution and wavelets, *Proc. Edinburgh Math. Soc.* **37** (1994), 271–300.

- [67] G. Kutyniok and D. Labate, Resolution of the wavefront set using continuous shearlets, *Trans. Amer. Math. Soc.* **361** (2009), 2719–2754.
- [68] W. Lawton, Tight frames of compactly supported affine wavelets, *J. Math. Phys.* **31** (1990), 1898–1901.
- [69] S. G. Mallat, Multiresolution approximations and wavelet orthonormal bases of $L_2(\mathbb{R})$, *Trans. Amer. Math. Soc.* **315** (1989), 69–87.
- [70] S. G. Mallat, *A Wavelet Tour of Signal Processing*, 2nd Edition, Academic Press, 1999.
- [71] Y. Meyer, *Wavelets and Operators*, Cambridge University Press, 1992.
- [72] Y. Meyer, *Oscillating Patterns in Image Processing and Nonlinear Evolution Equations*, American Mathematical Society, Univeristy Lecture Series, **22**, 2001.
- [73] A. Nemirovski and D. Yudin, *Problem Complexity and Method Efficiency in Optimization*, Wiley, New York, 1983.
- [74] Y. Nesterov, A method of solving a convex programming problem with convergence rate $O(1/k^2)$, *Soviet Mathematics Doklady* **27** (1983), 372–376.
- [75] Y. Nesterov, On an approach to the construction of optimal methods of minimization of smooth convex functions, *Ėkonom. i. Mat. Metody* **24** (1988), 509–517.
- [76] Y. Nesterov, Smooth minimization of nonsmooth functions *Math. Prog.* **103** (2005), 127–152.
- [77] S. Osher, Y. Mao, B. Dong and W. Yin, Fast linearized Bregman iteration for compressed sensing and sparse denoising, *Communications in Mathematical Sciences*, to appear.
- [78] A. Ron and Z. Shen, Frames and stable bases for shift-invariant subspaces of $L_2(\mathbb{R}^d)$, *Canadian J. Math.* **47(5)** (1995), 1051–1094.
- [79] A. Ron and Z. Shen, Affine systems in $L_2(\mathbb{R}^d)$: the analysis of the analysis operator, *J. Funct. Anal.* **148** (1997), 408–447.
- [80] A. Ron and Z. Shen, Affine systems in $L_2(\mathbb{R}^d)$ II: dual systems, *J. Fourier Anal. Appl.* **3** (1997), 617–637.
- [81] A. Ron and Z. Shen, Generalized shift invariant systems, *Constr. Approx.* **22** (2005), 1–45.
- [82] A. Ron, and Z. Shen, Weyl-Heisenberg frames and Riesz bases in $L_2(\mathbb{R}^d)$, *Duke Math. J.* **89** (1997), 237–282.
- [83] Z. Shen, K.C. Toh and S. Yun, An accelerated proximal gradient algorithm for frame based image restorations via the balanced approach, preprint, 2009.
- [84] K.C. Toh and S. Yun, An accelerated proximal gradient algorithm for nuclear norm regularized least squares problems, *Pacific J. Optimization*, to appear.
- [85] J.-L. Starck, M. Elad and D. L. Donoho, Image decomposition via the combination of sparse representations and a variational approach, *IEEE Trans. Image Proc.* **14** (2005), 1570–1582.
- [86] P. Tseng, On accelerated proximal gradient methods for convex-concave optimization, preprint, 2008.

-
- [87] W. Yin, S. Osher, D. Goldfarb and J. Darbon, Bregman iterative algorithms for ℓ_1 -minimization with applications to compressed sensing, *SIAM J. Imaging Sci.* **1** (2008), 143–168.
- [88] X. Zhang, M. Burger, X. Bresson and S. Osher, Bregmanized nonlocal regularization for deconvolution and sparse reconstruction, CAM Report (09–03), UCLA, 2009.

Role of Computational Science in Protecting the Environment: Geological Storage of CO₂

Mary F. Wheeler*, Mojdeh Delshad[†], Xianhui Kong[‡],
Sunil Thomas[§], Tim Wildey[¶], and Guangri Xue^{||}

Abstract

Simulation of field-scale CO₂ sequestration (which is defined as the capture, separation and long-term storage of CO₂ for environmental purposes) has gained significant importance in recent times. Here we discuss mathematical and computational formulations for describing reservoir characterization and evaluation of long term CO₂ storage in saline aquifers as well as current computational capabilities and challenges.

Mathematics Subject Classification (2010). 65N12, 65N15, 65N30, 65N08, 65N22, 65Z06, 76T30, 76V05, 35J15, 35J70, 35K61, 35Q86, 35L02; 86-08.

Keywords. CO₂ sequestration, parallel computation, multiscale and multiphysics coupling, multiphase flow, reactive transport, mixed finite element, discontinuous Galerkin, and a-posteriori error estimation.

*Center for Subsurface Modeling (CSM), Institute for Computational Engineering and Sciences (ICES), The University of Texas at Austin, 1 University Station, C0200, Austin, TX 78712. E-mail: mfw@ices.utexas.edu

[†]CSM, ICES, Department of Petroleum and Geosystems Engineering, The University of Texas at Austin, Austin, TX 78712. E-mail: delshad@mail.utexas.edu.

[‡]Department of Petroleum and Geosystems Engineering, The University of Texas at Austin, Austin, TX 78712. E-mail: xhkong@mail.utexas.edu.

[§]CSM, ICES, The University of Texas at Austin, Austin, TX 78712. Present address: Chevron ETC, San Ramon, CA 94583. Chevron ETC, 6001 Bollinger Canyon Rd, San Ramon, CA 94583. E-mail: sunil.g.thomas@gmail.com.

[¶]CSM, ICES, The University of Texas at Austin, 1 University Station, C0200, Austin, TX 78712. E-mail: twildey@ices.utexas.edu

^{||}CSM, ICES, The University of Texas at Austin, 1 University Station, C0200, Austin TX 78712. E-mail: gxue@ices.utexas.edu.

1. Introduction

Currently the world obtains more than 80% of its energy (coal, oil, gas) for the global economy from the subsurface. The byproducts of consuming these fuels such as greenhouse gas accumulation in the atmosphere are serious and potentially devastating. Renewables such as solar energy and wind farms may take many decades to develop before becoming economically feasible alternatives capable of replacing or reducing fossil energy usage. A major hope for the near future is geologic sequestration, a proven means of permanent CO₂ greenhouse-gas storage. This method involves injecting CO₂, generally in supercritical form, directly into underground geological formations. Oil and gas fields, saline formations, unminable coal seams, and saline-filled basalt formations are considered as storage sites. Various physical processes such as highly impermeable caprock and geochemical trapping mechanisms would prevent the CO₂ from escaping to the surface. Unfortunately, it is difficult to design and manage such efforts. Predictive computational simulation may be the only means to account for the lack of complete characterization of the subsurface environment, the multiple scales of the various interacting processes, the large areal extent of storage sites such as saline aquifers, and the need for long time predictions.

In this paper we discuss mathematical and computational formulations for describing reservoir characterization and evaluation of long term CO₂ storage in saline aquifers as well as current computational capabilities and challenges. We note that numerical software must be based on high fidelity multiscale and multiphysics algorithms which are necessary for simulation of multiphase flow and transport coupled with geochemical reactions and related mineralogy and geomechanical deformation in porous media, to predict changes in rock properties during sequestration. Moreover, models need to be validated and verified against field-scale experimental tests; this is a necessary prerequisite for wide-spread acceptance of computational simulation.

The paper is organized as follows. Section 2 describes the flow model theory and approximation methods. It includes some details on the equation of state (EOS) and a very brief discussion of the two phase flash equations. Section 3 presents the thermal energy transfer model with details of the time-split method of solution as well as a flow chart showing the sequential coupling to the flow step. Section 4 presents numerical results obtained using The University of Texas at Austin Integrated Parallel Accurate Reservoir Simulator (IPARS). In Section 5 we discuss the needs of including specific physical processes and in Section 6 emerging developments in high fidelity discretizations, and solvers. While optimization and control, risk analysis and uncertainty quantification models are important in addressing sequestration problems, we are restricting our attention to forward modeling in this paper.

2. Compositional Flow Model

In the equations that follow, i and α represent component and phase indices respectively. Component mass balances for $i = 1, \dots, n_c$ are given by

$$\frac{\partial(\varphi N_i)}{\partial t} + \nabla \cdot \left(\sum_{\alpha} \mathbf{J}_i^{\alpha} \right) = q_i, \quad (1)$$

where φ is the medium porosity, N_i and q_i are the molar concentration and source/sink of component i . \mathbf{J}_i^{α} is the “net velocity” of component i in phase α given by

$$\mathbf{J}_i^{\alpha} = \rho_{\alpha} \xi_i^{\alpha} \mathbf{u}_{\alpha} - \varphi \rho_{\alpha} S_{\alpha} \mathbf{D}_i^{\alpha} \nabla \xi_i^{\alpha}. \quad (2)$$

An expression for the diffusion-dispersion tensor, \mathbf{D}_i^{α} can be found in [18]. In this work, only molecular diffusion is considered. This renders \mathbf{D}_i^{α} a diagonal tensor. In (2), ρ_{α} , S_{α} , and \mathbf{u}_{α} are the density, saturation and velocity, respectively of phase α and ξ_i^{α} is the mole fraction of component i in phase α . The velocity \mathbf{u}_{α} is given by Darcy’s constitutive law,

$$\mathbf{u}_{\alpha} = -\frac{\kappa_{r\alpha}}{\mu_{\alpha}} \mathbf{K} (\nabla p_{\alpha} - \gamma_{\alpha} \nabla D), \quad (3)$$

where $D(\mathbf{x})$ is the “depth” which is a function of space. The phase pressures p_{α} are related by a capillary pressure which is a known function of phase saturation. Thus all the phase pressures can be eliminated in terms of the reference phase pressure, by equations

$$p_{\alpha} = p + p_{c\alpha}(S_{\alpha}), \quad (4)$$

where α includes all phases except the reference phase. In (3), \mathbf{K} is the absolute permeability of the medium and $\kappa_{r\alpha}$, μ_{α} , and γ_{α} are the relative permeability, viscosity and specific gravity, respectively of phase α . It is assumed that $\kappa_{r\alpha}$ is a function of S_{α} . The reference phase pressure is used in the flash, well and geomechanical calculations. Porosity may be approximated as

$$\varphi = \varphi_0 (1 + c_r (p - p_0)), \quad (5)$$

where c_r is the (constant) rock compressibility factor and φ_0 is the prescribed porosity at the standard pressure, p_0 .

The water phase viscosity μ_w in (3), is computed using an Arrhenius-type exponential temperature correlation $\mu_w = \mu_{w,\text{ref}} \exp b_w \left(\frac{1}{T} - \frac{1}{T_{\text{ref}}} \right)$, where $\mu_{w,\text{ref}}$ is the reference viscosity of water prescribed at T_{ref} and b_w is a constant. Non-aqueous phase viscosities ($i = 2, \dots, n_c$) are calculated using a Lohrenz-Bray-Clark correlation [42], wherein, at low pressures, component viscosities are estimated using a Stiel and Thodos correlation [58] in the form, $\mu_i^{\text{low}} = \frac{\beta_i}{\lambda_i}$; β_i is a function of $T_{i,r}$ and λ_i is a function of $T_{i,\text{cr}}$ and $p_{i,\text{cr}}$. At high pressures, a different correlation is used to obtain μ_i^{high} . A linear mixing rule is then applied

to the component values to obtain the phase viscosities. The definitions of the terms $T_{i,cr}$, $p_{i,cr}$, and $T_{i,r}$ are given in Section 2.1.

After the component properties are expressed in terms of state variables p and T , their mole-fractions, ξ_i^α determined using the fugacity equations for phase equilibrium (which in turn yield phase properties) and the phase saturations expressed in terms of the state variables vector $[p, \mathbf{N}, \{\xi_i^\alpha\}, T]^T$ (detailed in section 2.1), it only remains to solve for the reference phase pressure p , component molar concentrations N_i and the reservoir temperature T .

Toward this end, the non-aqueous phase molar specific volumes, i.e., $\nu_\alpha \equiv 1/\rho_\alpha$, are given by the gas law

$$\nu_\alpha = \frac{RTZ_\alpha}{p_\alpha}, \quad (6)$$

where R is the Universal gas constant and Z_α is the “Z-factor” (also called “compressibility factor”) of phase α , obtained by solving the equation of state. The aqueous phase, molar specific volume is given by

$$\nu_w = \frac{\nu_w^0 B_w^0}{1 + c_w(p_w - p_w^0)}, \quad (7)$$

where c_w is the water phase compressibility and p_w^0 is the standard water phase pressure (usually, one atmosphere) at which the values ν_w^0, B_w^0 of the molar specific volume and the “formation volume factor”, respectively, are prescribed. The phase saturations are then expressed in terms of the state variables, viz., $[p, \mathbf{N}, \{\xi_i^\alpha\}, T]^T$ by the equations,

$$S_w = \nu_w N_w, \quad S_l = (1 - v)\nu_l \sum_{i=2}^{n_c} N_i, \quad S_g = v\nu_v \sum_{i=2}^{n_c} N_i, \quad (8)$$

where v is the vapor fraction and the subscripts w, l and g stand for the aqueous, oleic and gaseous phases, respectively. The saturations calculated in (8) will not, in general, sum to unity and therefore the iterative implicit pressure and explicit concentration (IMPEC) method needs an additional constraint,

$$\sum_{\alpha} S_\alpha = 1. \quad (9)$$

Equation (9) is the familiar volume balance criterion which forms the convergence condition for the iterative IMPEC method.

2.1. The Equation of State and Flash Implementation. The simulator uses the Peng-Robinson EOS [50] to determine the non-aqueous molar specific volumes of (6) in terms of the “Z-factors” which are functions of p and T . It can be shown [17] that this reduces to solving a cubic equation for \bar{Z}_α , which includes a volumetric shift parameter C_α so that, $Z_\alpha = \bar{Z}_\alpha - C_\alpha$. The cubic EOS for phase α then takes the form

$$\bar{Z}_\alpha^3 + h_1(B_\alpha)\bar{Z}_\alpha^2 + h_2(A_\alpha, B_\alpha)\bar{Z}_\alpha + h_3(A_\alpha, B_\alpha) = 0. \quad (10)$$

The arguments A_α, B_α in the coefficients h_j of the cubic EOS (10), and C_α , in the definition of \bar{Z}_α are functions of the component “reduced” pressures $\{p_{i,r}\}$ and temperatures $\{T_{i,r}\}$, where $p_{i,r} = p_i/p_{i,\text{cr}}, T_{i,r} = T/T_{i,\text{cr}}$. They are also functions of the composition of that phase (via mixing rules applied to the component mole fractions $\{\xi_i^\alpha\}$). It is noted that p_i is Dalton’s partial pressure of the component i in phase α . The quantities $p_{i,\text{cr}}, T_{i,\text{cr}}$ are component critical pressures and temperatures respectively and these are thermodynamic properties unique to every component. For details, the reader is referred to [17]. Equation (10) is solved using the Newton-Raphson method.

Vapor fractions are determined using the Rachford-Rice equation. It is given by

$$\sum_i \frac{(K_i - 1)z_i}{1 + (K_i - 1)v} = 0. \quad (11)$$

In (11), z_i is the overall nonaqueous mole fraction of component i . The fluid is a single-phase liquid when $v = 0$ and is a single-phase vapor when $v = 1$. Once v is determined using (11), the component mole fractions in the liquid and vapor phases can be calculated from

$$\xi_i^l = \frac{z_i}{1 + (K_i - 1)v}, \quad \xi_i^g = K_i \xi_i^l. \quad (12)$$

In (12), z_i is the overall nonaqueous mole fraction of component i . It remains then to determine the equilibrium phase composition $\{\xi_i^\alpha\}$ using a “flash” algorithm. From thermodynamics, phase equilibrium at constant pressure and temperature requires that the component fugacities in each of the non-aqueous phases be equal, i.e., for $i = 2, \dots, n_c$,

$$f_i^g = f_i^l. \quad (13)$$

Rather than solving (13) for component fugacities, the simulator solves for $\ln(K_i)$, working with fugacity coefficients, $\Phi_i^\alpha \equiv \frac{f_i^\alpha}{\xi_i^\alpha p_\alpha}$ instead of fugacities. This modifies (13) to

$$R_i \equiv \ln(\Phi_i^l) - \ln(\Phi_i^g) - \ln(K_i) = 0. \quad (14)$$

The $\ln(\Phi_i^\alpha)$ term in (14) is a function of Z_α , $\{\xi_i^\alpha\}$ and p_α for the case of the Peng-Robinson EOS. The flash algorithm then applies the Newton-Raphson method to find the root of (14). For further details, the reader can refer to [17].

2.2. Iterative IMPEC Implementation. The phase saturations are functions of the state variables $[p, \mathbf{N}, T, \{\xi_i^\alpha\}]^T$ as given by the (8). The volume balance at the $(k + 1)^{\text{th}}$ iteration of the n^{th} time step is written as

$$S_T^{k+1} \equiv \sum_\alpha S_\alpha^{k+1} = 1. \quad (15)$$

Expanding this, yields

$$\frac{\partial S_T}{\partial p} \delta p + \sum_{i=1}^{n_c} \frac{\partial S_T}{\partial N_i} \delta N_i + \sum_{i=2}^{n_c} \frac{\partial S_T}{\partial (\ln K_i)} \delta (\ln K_i) = 1 - S_T^k. \quad (16)$$

The fugacity equations (13) can be expanded in terms of δp , δN_i and $\delta (\ln K_i)$ and this expansion can be rearranged to express $\delta (\ln K_i)$ in terms of δp and δN_i . This expression can be substituted in (16) resulting in an equation in terms of δp and δN_i . The mass balance (1) can then be expanded to express δN_i in terms of δp which results in a single system for cell pressure changes of the form

$$\mathbf{A} \delta p = \mathbf{b}, \quad (17)$$

which is the linear pressure system and can be solved using any standard linear solver.

Once the pressure is updated, the change in porosity is calculated using (5) and the component accumulation term (denoted \mathcal{A}_i) in the mass balance (1) is calculated from

$$\mathcal{A}_i^{k+1} = \frac{\varphi N_i^k + N_i^k \delta \varphi + \varphi^k \delta N_i - \varphi^n N_i^n}{\Delta t}. \quad (18)$$

In (18) δN_i has been expressed in terms of δp using the mass balance (1). Then the $(k + 1)^{\text{th}}$ iteration concentrations are obtained using the equation

$$N_i^{k+1} = \frac{\Delta t \mathcal{A}_i^{k+1} + \varphi^n N_i^n}{\varphi^{k+1}}. \quad (19)$$

This technique of updating N_i avoids component material balance errors that arise due to the product of N_i^{k+1} and φ^{k+1} . Because of the explicit-in-time nature of the concentration (19), the IMPEC calculations can become unstable if time steps become too large. The simulator currently uses a saturation-type control to limit time step sizes for the iterative IMPEC implementation. The saturation change for a component is defined as

$$(\Delta S_T)_i = \frac{\partial S_T}{\partial N_i} \Delta N_i. \quad (20)$$

where ΔN_i is the change in concentration of the i^{th} component. The simulator then requires that $|(\Delta S_T)_i| \leq DS_{\max}$ for all components i during a timestep.

Once the $(k + 1)^{\text{th}}$ level iteration solution for p and N_i are available, the simulator returns to check if the volume balance (convergence) condition given in (15) is satisfied. If not, it repeats the process described between (16) and (19) until (15) holds upto a tolerance or until a maximum number of iterations is exceeded. At each iteration the solution from the most recently available iteration (or previous time step in the case $k = 1$) is used.

3. Thermal Energy Transfer Model

The IPARS simulator implements a “weak” or “sequential” coupling between the flow and thermal steps. This is justifiable in most of the practical subsurface applications, especially the CO₂ sequestration injection tests mentioned in the introduction, since the temperature changes encountered are typically relatively small. In this section, the equations governing thermal energy transfer and the solution scheme are presented.

3.1. Governing Equations. The thermal energy balance is described by the PDE

$$\frac{\partial U_T T}{\partial t} + \nabla \cdot \left(\sum_{\alpha} \rho_{\alpha} C_{p\alpha} T \mathbf{u}_{\alpha} - \lambda \nabla T \right) = q_H, \quad (21)$$

in the unknown reservoir temperature T , where U_T is the “effective isochoric specific heat capacity” (from thermodynamics, this is nothing but $\partial U/\partial T$). It is given by

$$U_T = (1 - \varphi) \rho_s C_{vs} + \varphi \sum_{\alpha} \rho_{\alpha} S_{\alpha} C_{v\alpha}. \quad (22)$$

In Eqs.(21)-(22), $C_{p\alpha}$ and $C_{v\alpha}$ are the isobaric and isochoric molar specific heat capacities of the phase α computed from their respective component counterparts using an appropriate mixing rule. Finally, λ is the “effective reservoir thermal conductivity” and q_H is the “heat source/sink per unit volume”, given by

$$q_H = \sum_{\alpha} C_{p\alpha} q_{\alpha} T_{src}, \quad (23)$$

where q_{α} is the injection or production flow rates of the phase α per unit volume, once again calculated from their component counterparts, q_i in (1) and the component mole fractions. It is noted that T_{src} is the temperature of the injected fluid T_{inj} at source points and equals the resident temperature T at the sink points. The subscript s in (22) represents the rock.

3.2. Time-Split Scheme. Let $t^{m+1} \in [t^n, t^{n+1}]$ be the time at which the thermal step is solved. In general, the simulator allows for multiple thermal steps nested within a flow step. The basic idea of the *time-split* scheme is to successively account (by accumulation) for the advection and diffusion (or thermal conduction, in this case) in time. Hence, it can be regarded as an operator-splitting method. Theoretical details of the method can be found in [21, 51] where it has been applied to the species transport problem. Accordingly, the (21) is split into an advection and a diffusion step. The advection step is given by the equation

$$\frac{\partial(U_T T)}{\partial t} + \nabla \cdot \left(\sum_{\alpha} \rho_{\alpha} C_{p\alpha} T \mathbf{u}_{\alpha} \right) = q_H. \quad (24)$$

A higher order Godunov method has been implemented using element slopes of the scalar variables in the advection term and carefully chosen flux limiters [22, 41, 20]. For brevity, the first order scheme is presented here. Let E be any element of the finite element mesh and $H_E^m \equiv \int_E U_T^m T^m dx$ be the local thermal energy content in E at time step m of the thermal algorithm. Integrating over E of (24) against the characteristic function on E , gives the weak form of the advection step,

$$\frac{\bar{H} - H_E^m}{\Delta t^m} + \int_E \sum_{\alpha} \bar{\mathbf{u}}_{\alpha} \cdot \mathbf{n} ds = \int_E q_H^{m+1/2} dx. \tag{25}$$

In (25) $\Delta t^m = t^{m+1} - t^m$, \mathbf{n} is the unit outward normal to element E and \bar{H} denotes the intermediate value of the local thermal energy content H_E^{m+1} (which is sought to obtain T_E^{m+1}) from the contribution due to advection only and becomes the initial condition for the conduction step. Further, the quantity $\bar{\mathbf{u}}_{\alpha} = (C_{p\alpha} T)^{m, \text{upw}} (\rho_{\alpha}^{\text{upw}} \mathbf{u}_{\alpha})^{m+1/2}$, where $(C_{p\alpha} T)^{m, \text{upw}}$ represents the up-winded value of $(C_{p\alpha} T)^m$ based on the sign of $\mathbf{u}_{\alpha}^{m+1/2} \cdot \mathbf{n}$. Similarly, $\rho_{\alpha}^{m+1/2, \text{upw}}$ denotes the upwinded value of $\rho_{\alpha}^{m+1/2}$ based on the sign of $\mathbf{u}_{\alpha}^{m+1/2} \cdot \mathbf{n}$. It is noted that the values of ρ_{α} and \mathbf{u}_{α} are known at flow time steps t^n and t^{n+1} . Thus, $(\rho_{\alpha} \mathbf{u}_{\alpha})^{m+1/2}$ is the linear interpolant computed at $t^{m+1/2}$.

Once \bar{H} is determined, the conduction step is solved, given by

$$\frac{\partial(U_T T)}{\partial t} + \nabla \cdot (\lambda \nabla T) = 0. \tag{26}$$

In the weak form, upon integration by the characteristic function on E , this becomes

$$\frac{H_E^{m+1} - \bar{H}}{\Delta t^m} + \int_E \nabla \cdot (\lambda \nabla T^{m+1}) = 0. \tag{27}$$

(27) is solved using the backward-Euler and mixed finite element methods for time and space discretizations respectively. The lowest order RT_0 approximation space in conjunction with the trapezoidal quadrature rule for the flux term $\lambda \nabla T^{m+1}$ reduces it to a cell-centered finite difference approximation [56]. The accumulation term H^{m+1} is linearized about the current temperature. The resulting linear system is solved using the bi-conjugate gradient method, with the temperature \bar{T} from the advection step as initial guess.

4. Numerical Results

In this section, the sequentially coupled “iterative-IMPEC, time-split thermal” scheme in IPARS is applied to some challenging problems. We first discuss results of a benchmark CO₂ injection problem proposed by Dahle et al. [19]. The benchmark definition is based on a homogeneous aquifer with a dip of 1% bounded by impermeable barriers on top and bottom boundaries. The aquifer

Table 1. Model data

Acquifer size, L, W, H	200 km, 100 km, 50 m
Dip	1%
Depth (at center)	2.5km
Surface temperature	10 °C
Geothermal gradient	25 C/km
Permeability	10^{-13} m ²
Porosity	15 %
Horizontal well length	1 km
Initial water saturation	1
Residual water saturation, s_{wr}	0.2
Residual CO ₂ saturation, s_{gr}	0.2
Injection rate	1 Mt / year
Injection period	20 years

properties are given in Table 1. The flow boundary conditions in the horizontal direction are constant head. A horizontal injection well is placed at the bottom vertical layer 50 km updip from the lowest point of the formation and in the center with respect to the horizontal direction. Drainage relative permeability and capillary pressure functions are used to describe the CO₂ injection in fully water saturated aquifer. The Corey power law functions are used as follows.

$$k_{rw} = \bar{s}_w^4 \quad \text{and} \quad k_{r,\text{CO}_2} = 0.4 (1 - \bar{s}_w^2) (1 - \bar{s}_w)^2, \quad (28)$$

where \bar{s}_w is the normalized water saturation defined as

$$\bar{s}_w = \frac{s_w - s_{wr}}{1 - s_{wr}}. \quad (29)$$

The primary drainage capillary pressure in bar is given by

$$p_c = 0.2(\bar{s}_w)^{-1/2}. \quad (30)$$

We have included the effect of trapping with a maximum residual CO₂ saturation of 0.2 corresponding to the primary imbibition curve. The effects of mineral reactions and rock mechanics were not considered in the example. Total distribution of carbon in free phase, residual phase, dissolved state at the end of injection and later times are given in Fig 1. Fig. 2 shows free CO₂ saturation profiles at 100, 1000, 5000, and 10000 years. The results indicate the upward movement of CO₂ to the impermeable aquifer boundary. It is interesting to note that there is significant free phase at top seal even after 10,000 yrs of dissolution and trapping. Using a conservatively fine grid near the well, the computation to 10,000 years took approximately 3 weeks on 24 processors with very conservative time steps.

The same case was repeated but including hysteresis in capillary pressure and relative permeability. The residual CO₂ saturation is not a constant as

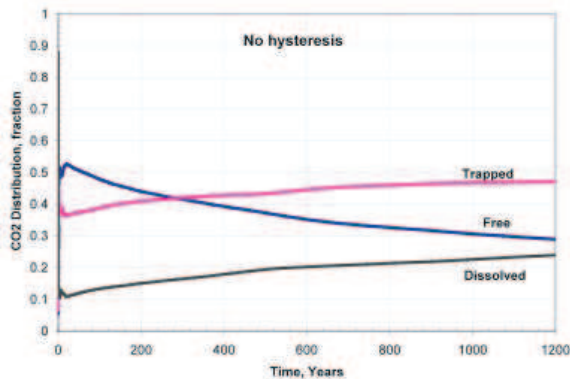


Figure 1. CO₂ distribution as a function of time

in the previous case and changes according to the history of CO₂ saturation [61]. Typically during the injection, CO₂ is not trapped but will trap during the subsequent redistribution as a function of CO₂ saturation [39]. The CO₂ distribution is given in Fig. 3 for the case with hysteresis.

5. Modeling of Specific Physical Processes

The four primary mechanisms for CO₂ trapping in brine formations are (1) residual trapping, in which CO₂ becomes a disconnected phase trapped in individual pores of the rock, (2) structural trapping due to low-permeability cap rocks, (3) dissolution, which works by dissolving CO₂ gas into brine, and (4) mineral trapping, in which dissolved CO₂ in brine reacts with rock to form minerals. We now briefly discuss processes that we feel must be modeled in a state-of-the-art simulator and that we are concurrently working on for inclusion in IPARS [60, 62]

5.1. Geomechanics, Faults and Fractures. The CO₂ structural and mineral trapping mechanisms alter the state of the porous medium. This can in turn alter the sequestration potential of the formation, so it is critical to model accurately the geomechanical response (compaction, subsidence) of the aquifer [44]. Few simulators have this capability, especially in the context of complex compositional flows. Multiphase flow and geomechanics are highly coupled phenomena that occur at multiple spatial and temporal scales. Given the size and complexity of CO₂ simulations and the multitude of scales involved, a major mathematical research issue involves finding efficient and provably accurate multiphysics coupling algorithms of mechanics and flow. Domain decomposition (DD) has been employed to thermohydromechanical problems in [64, 57] and using mortar spaces with different grids on different subdomains [2, 28, 25].

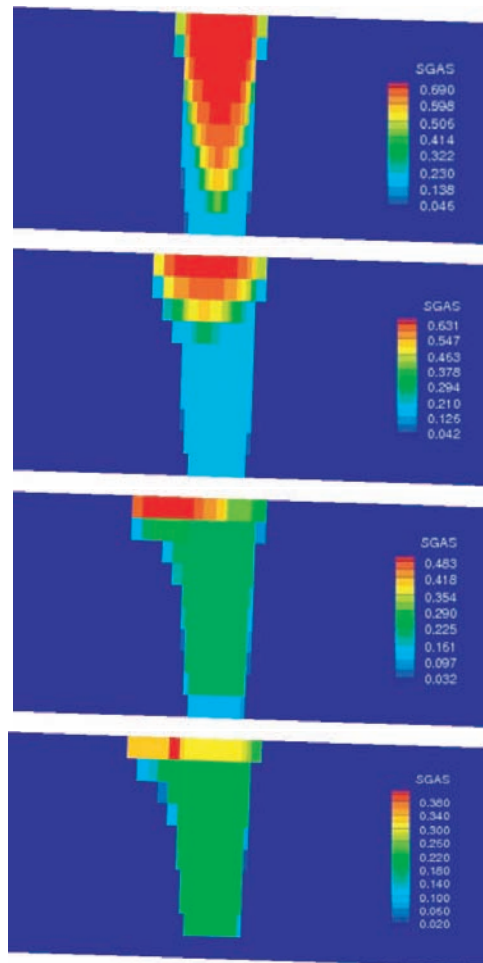


Figure 2. CO₂ free phase saturation at 100, 1000, 5000, and 10000 yrs

DD appears to be an accurate and practical approach for treating this computationally intensive problem; namely, multiscale or subgrid models and multiphysics models can be coupled across domains effectively. For example in [25] a time-dependent poroelastic model in a region is coupled with an elastic model in adjacent regions. Each model is discretized independently on non-matching grids and a domain decomposition is defined on the interface between the regions by introducing discontinuous Galerkin (DG) jumps and mortars. The unknowns are condensed on the interface, so that at each time step, the computation in each subdomain can be performed in parallel. In addition, by extrapolating the displacement, the computations of the pressure and displacement are decoupled. While this example address a linear problem, nonlinear multiphase examples have been treated in [52].

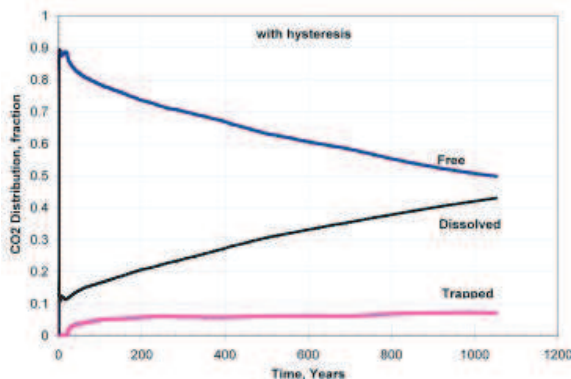


Figure 3. CO₂ distribution as a function of time including hysteresis

We remark that field studies are important to assess the need for including non-linear mechanics. The flow of fluids is significantly impacted by the geometry and distribution of natural fractures and faults. These features are small-scale in aperture, but large-scale in extent, so their effects are difficult to model accurately. Flow in highly fractured subsurface formations is often modeled by dual-porosity models. These have been shown through experiments and mathematical homogenization to be sound [4]. Recent work has extended applicability to compositional flows [46].

One possible approach for modeling discrete fractures and faults is to develop new discretization methods, such as DG that model the fault directly on a refined mesh [49]. A second approach is to use mortar techniques. Here a mortar is a 2-D function space that is used to “glue” together two 3-D regions. If we set the mortar to coincide with a fault, we have the opportunity to model complex flow and other processes within the fault. Further mortar space discretization work related to parallel processing and multiscale modeling is discussed later.

5.2. Phase Behavior and Fluid Properties. Trapping of CO₂ as a residual phase is the most significant means of sequestering CO₂ for long durations. The performance of CO₂ injection and sequestration is a strong function of relative permeability and capillary pressure vs. fluid saturation relationships. Injection of CO₂ into a brine formation follows a primary drainage path until the injection is terminated. CO₂ becomes less dense and tends to rise as the pressure drops near the injection well, allowing water to invade regions previously swept by CO₂. This is referred to as the imbibition process in which a substantial volume of CO₂ can become trapped [35].

The notion that relative permeabilities and capillary pressures are functions of both saturation and its history, known as hysteresis, is well established. We must model hysteresis to accurately predict the amount of CO₂ that is immobilized and permanently trapped. However, recent drainage and imbibition

tion experiments [29] indicate that the water-CO₂ system may require a new hysteretic model that takes into account not only saturation and saturation history, but also the pressure history. That is, the hysteretic effect of the CO₂ plume changing from drainage to imbibition is a strongly non-linear effect.

5.3. Thermal Effects. Non-isothermal effects are important during CO₂ injection into brine aquifers. The impact of temperature is generally restricted to the near injection wellbore regions. However, there are other situations where large variations in temperature can occur due to sudden expansion of CO₂ where pressure declines. Such instances are in fault zones, fractures, and abandoned wells. The temperature variations can have a large impact on the thermodynamic phase behavior and the dissolution of CO₂ in water. Moreover, geomechanical overburden and underburden effects can generate heat. Nonlinear convergence issues and simulator robustness are major issues that need to be addressed.

6. Numerical Algorithms, Discretization, Solvers, and Uncertainty Quantification

There have been many advances recently in the theoretical development of advanced numerical algorithms for the discretization of partial differential equations; however, incorporation in subsurface simulation software has not kept pace. Of specific importance is physics-preserving discretizations, which give numerical models that preserve basic physical principles. Two important physical properties, that should be reflected in the mathematics, are that (1) mass is conserved locally at every point of space and (2) diffusion or dispersion cannot produce local maxima or local minima in the solution. The spreading of a CO₂ plume cannot be correctly approximated if either principle is violated. For example, artificial creation of mass results in a plume that has too much CO₂ and spreads faster than is to be realized in practice. Violation of the maximum principle has deleterious effects on the computation of reaction dynamics. Most numerical approximation schemes fail to preserve these physical/mathematical principles, and so lead to erroneous simulation results.

The subsurface geometry also plays an important role. The computational mesh must follow the geologic strata for accurate results. However, these strata generally form irregular layers that often cease to exist in what is known as a “pinch-out.” Moreover, local refinement is generally desired to increase resolution, and therefore accuracy, in regions where the solution varies greatly.

6.1. Multiscale Temporal and Spatial Discretization. Two important physics-based discretizations are adaptive discontinuous Galerkin (DG) methods and variations of the mixed finite element (MFE) method, including multipoint flux approximations and mimetic methods. Both types

of methods exhibit local mass conservation. Important directions for research for these two types of discretization methods includes the improving the incorporation of the maximum principle and the application of mortar methods as described below.

The MFE methods approximate physical fluxes continuously and have a more economical use of degrees of freedom, so there are good solvers for these methods (although better solvers are needed). Although many researchers feel these methods are unable to handle general meshes, this is not the case in two dimensions; one can take simplicial meshes and apply a domain decomposition approach with local elimination of variables [63].

Standard formulation of MFE gives a saddle point system and requires solving pressure and velocity simultaneously. Another approach is hybrid MFE formulation [7, 13] that reduces to a symmetric positive definite system for the pressure Lagrange multipliers on the element faces. More efficient formulations that reduced to cell-centered pressure schemes have been developed. These are based on appropriate MFE spaces and numerical quadrature rule for the velocity mass term, see [55, 65] for diagonal permeability on rectangular grids based on the lowest order Raviart-Thomas MFE method [54], and expanded mixed finite element (EMFE) for smooth full tensor permeability on rectangular and cuboid grids [6]. For the case of discontinuous permeability, the EMFE loses the accuracy unless pressure Lagrange multipliers are introduced along the interfaces between discontinuities [3]. Multipoint flux mixed finite element (MFMFE) methods [71, 31, 67, 66] are designed to be a cell-centered pressure scheme that is accurate for both smooth and discontinuous full tensor permeability.

In addition, there has recently been the development of related MFEMs, mimetic approximations [30, 36, 8, 9, 11, 12, 40, 10] that can handle tensor permeability coefficients and general meshes.

DG methods apply to transport processes, and they relatively easily accommodate irregular, nonconforming meshes and hp-adaptivity, i.e., mesh spacing refinements (h-refinements) and also polynomial degree increases (p-refinements). Because they have many more degrees of freedom than competing methods, a challenge is to construct a good numerical solution procedure.

Above mentioned discretization techniques can be combined to solve multiphase flow problems. For example, in the two-phase oil-water simulation, we use iterative coupling method to solve the pressure equation implicitly and to solve the saturation equation explicitly, where the MFMFE method is used for the pressure equation and the DG method for the saturation equation. These two discretizations are both accurate on general hexahedral meshes. Fig. 4 shows the water saturation and pressure profile at 200 days.

6.2. A Posteriori Error Estimates. The modeling of CO₂ sequestration requires the simulation of complex nonlinear systems of equations over a long temporal scale, rendering a priori error estimates pessimistic and

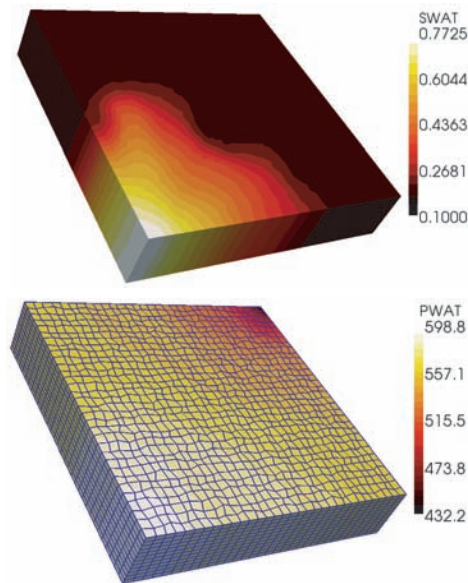


Figure 4. Water saturation (top) and water pressure (bottom) on hexahedral meshes at 200 days

impractical. The alternative is to utilize a-posteriori estimators based on local conservation principles for our discretization methods, in which the quality of the solution is quantified during and after its computation. If the level of error is below some acceptable tolerance, the solution is accepted. Otherwise, the estimator predicts where the error is greatest, and mesh refinement (or polynomial degree enhancement) is initiated and the problem is solved again. This process may be repeated. Recently we have developed a-posteriori estimators for MFEMs and DG discretizations, [68, 5, 48] in which the quality of the solution is quantified during and after its computation. While this latter work focuses on linear problems, preliminary results on some nonlinear equations such as slightly compressible single phase flow in porous media [32] and nonlinear Laplacian equation [1] indicate that these estimates can be useful in mesh selection and tolerances for inexact Newton methods.

In addition, it should also be noted that it is important to decide what information is being sought, since the estimator should be tailored to the needs of the engineering or scientific decision that is to be made. Such mesh refinement algorithms add considerable complexity to parallel computer codes: work on adaptive runtime management (e.g., for load balancing and dynamic repartitioning).

6.3. Multiphysics Couplings and Time-stepping. In the carbon sequestration, the treatment of multiphysics couplings is a major

computational difficulty, i.e. coupling of flow, thermal, geomechanics, and chemistry. For example, in some domains one needs to couple a plasticity model with compositional flow while in another domain poroelasticity is appropriate.

For interdomain coupling, i.e., coupling of physical processes through interfaces, the current state of the art is focused on developing new methods for handling interface boundary conditions and temporal/spatial scale variations. Such methods, for example, may use either a mortar space or discontinuous approximating spaces (e.g., DG) [5, 67, 27]. These methods have the advantage that boundary conditions are not imposed strongly on the solution but are enforced weakly, and they easily handle meshes that do not align at interfaces. Stability and error estimates for many such schemes have been derived for linear model problems. Extensions of the methods to some nonlinear applications, for example, in coupling different multiphase flow models, have also been explored [52]. Mortar methods in particular are of interest, as they provide a natural way of “decomposing” the problem using Schur-complement techniques and can provide the basis for domain-decomposition parallel solvers [2, 5]. As mentioned above, we are investigating issues of the proper definition of the mortar spaces for modeling geomechanics and faults, and a-posteriori estimation and multiscale mortar techniques [5], including curved mortars [23], tied to the temporal discretization.

Explicit time-stepping is inadequate for the CO₂ sequestration problem, since there are complex nonlinearities and multiple spatial and temporal scales. Moreover, fully implicit methods are too computationally expensive to be relied upon. We propose a flexible approach that involves iterative coupling with adaptive time-stepping. In the case of flow, the iterative coupling algorithm decouples a reference pressure equation from the concentration equations. We solve each in sequence, using the solution from the previous stage of the process. We then cycle back until convergence at a given time step is achieved, e.g., when the volume balance satisfies a specified tolerance. Similarly, in treating poroelasticity, multiphase flow and elasticity can be decoupled with convergence being based on the fluid fraction satisfying a given tolerance. The effectiveness of iterative coupling for both of these applications has been demonstrated in [43, 24] and shown to reduce computational time by roughly 40% to 50% over fully implicit methods. Moreover different discretizations can be applied to the decoupled problems.

Clearly, effective time-stepping schemes are essential for modeling reactive transport and multiphase flow. Practical questions arise concerning global and local time-step control and the use of different time-steps for different subdomains (of the parallel domain decomposition) and for different physical models. Generally one expects to use longer time steps for geomechanics, medium step sizes for flow and thermal effects, small time steps for transport processes, and micro-steps for reactive processes. Conservation of material mass and volume across space-time boundaries and maximum principles also affect the time-step. We are currently investigating the efficiency and accuracy of different types of couplings based on physical, numerical, and even stochastic considerations.

6.4. Linear and Nonlinear solvers. The development of accurate, efficient and robust parallel solvers for solving the large nonlinear dynamical systems that arise from finite element discretization represents a formidable challenge. General-purpose solver technology is inadequate for the complexity of the multiphysics and multiscale systems arising in CO₂ sequestration. Significantly faster run-times can be achieved by tailoring the solvers to the application, providing reasonable turn-around time for engineering and risk analysis [14, 15, 16]. Linear and nonlinear solvers frequently used in multiphase flow problems are described in [33, 34, 37, 38]. Emerging developments emphasize (1) multiscale and physics-driven deflation preconditioned algebraic multigrid (AMG) for modeling MFEM subdomain discretizations [33, 34, 59]; (2) balancing algorithms for preconditioning domain decomposition interface problems for solving the pressure equation [5] and for geomechanics [26]; (3) the robust hybrid banded linear solver [53, 45] for certain applications where a DG approximation may be used for subdomain problems for treating complex geometries [27, 47]; and (4) Krylov recycling methods [33]. Challenging aspects of the latter are determining updates for highly nonlinear problems, variable preconditioners, and exploiting Krylov basis information through different time steps.

We remark that the mortar formulation has also given rise to a multiscale mixed finite element method [5]. The solution is resolved locally on a fine scale while flux continuity across subdomain interfaces is imposed on the coarse scale via mortar finite elements. The method is comparable in cost to existing variational multiscale and subgrid upscaling methods, but it is more flexible since mortar degrees of freedom can be adapted locally if higher resolution is needed.

This method is made more efficient by precomputing a multiscale basis (or discrete Green's functions) by solving fine scale subdomain problems for each coarse scale mortar degree of freedom. With this approach the number of fine scale subdomain solves is independent of the number of interface iterations, which was not the case in the original implementation. While the multiscale flux basis is not directly applicable to nonlinear problems, the concept can be extended to a frozen Jacobian interface preconditioner for solving nonlinear interface problems [70]. A multiscale interface Jacobian is computed for a fixed state of the variables, which could be the state at the beginning of the nonlinear iteration or the state at the initial time in the case of time dependent problems. The preconditioner has been used for single phase slightly compressible and iteratively coupled or fully implicit two phase flow in porous media with up to a 95% reduction in the time to solve each nonlinear interface equation [70, 69].

Acknowledgments

A portion of this research was supported by the U.S. Department of Energy, Office of Science, Office of Basic Energy Sciences. The Center for Frontiers of Subsurface Energy Security (CFSES) is a DOE Energy Frontier Research

Center, under Contract No. DE-SC0001114. The authors gratefully acknowledge the financial support provided by the NSF-CDI under contract number DMS 0835745 and King Abdullah University of Science and Technology (KAUST)-AEA-UTA08-687. The last author is supported by Award no. KUS-F1-032-04, made by KAUST.

References

- [1] L. E. Alaoui, A. Ern, and M. Vohralik. Guaranteed and robust a posteriori error estimates and balancing discretization and linearization errors for monotone nonlinear problems. *Submitted to Comput. Methods Appl. Mech. Engrg.*, 2009.
- [2] T. Arbogast, L. C. Cowsar, M. F. Wheeler, and I. Yotov. Mixed finite element methods on nonmatching multiblock grids. *SIAM J. NUMER. ANAL.*, 37(4):1295–1315, 2000.
- [3] T. Arbogast, C. N. Dawson, P. T. Keenan, M. F. Wheeler, and I. Yotov. Enhanced cell-centered finite differences for elliptic equations on general geometry. *SIAM Journal on Scientific Computing*, 19(2):404–425, 1998.
- [4] T. Arbogast, J. Douglas, and U. Hornung. Derivation of the double porosity model of single phase flow via homogenization theory. *SIAM J. Math. Anal.*, 21(4):823–836, 1990.
- [5] T. Arbogast, G. Pencheva, M. F. Wheeler, and I. Yotov. A multiscale mortar mixed finite element method. *Multiscale Model. Simul.*, 6(1):319–346, 2007.
- [6] T. Arbogast, M. F. Wheeler, and I. Yotov. Mixed finite elements for elliptic problems with tensor coefficients as cell-centered finite differences. *SIAM Journal on Numerical Analysis*, 34(2):828–852, 1997.
- [7] D. N. Arnold and F. Brezzi. Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates. *Mathematical Modelling and Numerical Analysis*, 19(1):7–32, 1985.
- [8] M. Berndt, K. Lipnikov, M. Shashkov, M. F. Wheeler, and I. Yotov. A mortar mimetic finite difference method on non-matching grids. *Numer. Math.*, 102:203–230, 2005.
- [9] M. Berndt, K. Lipnikov, M. Shashkov, M. F. Wheeler, and I. Yotov. Superconvergence of the velocity in mimetic finite difference methods on quadrilaterals. *SIAM. J. Numer. Anal.*, 43(4):1728–1749, 2005.
- [10] F. Brezzi, K. Lipnikov, M. Shashkov, and V. Simoncini. A new discretization methodology for diffusion problem on generalized polyhedral meshes. *Comput. Methods Appl. Mech. Engrg.*, 196:3682–3692, 2007.
- [11] F. Brezzi, K. Lipnikov, and V. Simoncini. A family of mimetic finite difference methods on polygonal and polyhedral meshes. *Mathematical Models and Methods in Applied Sciences*, 15(10):1533–1551, 2005.
- [12] F. Brezzi, K. Lipnikov, and M. Shashkov. Convergence of the mimetic finite difference method for diffusion problems on polyhedral meshes. *SIAM J. NUMER. ANAL.*, 43(5):1872–1896, 2005.

-
- [13] Franco Brezzi and Michel Fortin. *Mixed and hybrid finite element methods*. Springer-Verlag, New York, 1991.
- [14] S. Chandra, X. Li, T. Saif, and M. Parashar. Enabling scalable parallel implementations of structured adaptive mesh refinement applications. *Journal of Supercomputing*, 39:177–203, 2007.
- [15] S. Chandra and M. Parashar. Towards autonomic application-sensitive partitioning for SAMR applications. *Journal of Parallel and Distributed Computing*, 65:519–531, 2005.
- [16] S. Chandra and M. Parashar. Addressing spatiotemporal and computational heterogeneity for structured adaptive mesh refinement. *Computing and Visualization in Science, Springer Verlag*, 9:145–163, 2006.
- [17] Y. Chang. *Development and application of an equation of state compositional simulator*. PhD Thesis, University of Texas at Austin, Austin, TX, August 1990.
- [18] Z. Chen, G. Huan, and Y. Ma. *Computational methods for multiphase flows in porous media*. SIAM, Philadelphia, PA, 2006.
- [19] H. K. Dahle, G. T. Eigestad, J. M. Nordbotten, and K. Pruess. A model-oriented benchmark problem for CO₂ storage. *Workshop on Modeling and Risk of Assessment of Geological Storage of CO₂*, 2009.
- [20] C. N. Dawson. Godunov mixed methods for advection-diffusion equations in multi-dimensions. *SIAM J. Numer. Anal.*, 30(5):1315–1332, Oct 1993.
- [21] C. N. Dawson and M. F. Wheeler. An operator-splitting method for advection-diffusion-reaction problems. In *The mathematics of finite elements and applications*, pages 463–482. Academic Press, London, UK, 1987.
- [22] R. Eymard, T. Gallouët, and R. Herbin. Finite volume methods. In P. G. Ciarlet and J. L. Lions, editors, *Handbook of numerical analysis*, pages 713–1020. Elsevier, North Holland, Amsterdam, Sep 2000.
- [23] B. Flemisch, J. M. Melenk, and B. I. Wohlmuth. Mortar methods with curved interfaces. *Applied Numerical Mathematics*, 54:339–361, 2005.
- [24] X. Gai. *A coupled geomechanics and reservoir flow model on parallel computers*. PhD thesis, The University of Texas at Austin, 2004.
- [25] V. Girault, G. Pencheva, M. Wheeler, and T. Wildey. Domain decomposition for poroelasticity and elasticity with DG jumps and mortars. *To appear*, 2010.
- [26] V. Girault, G. V. Pencheva, M. F. Wheeler, and T. M. Wildey. Domain decomposition for linear elasticity with DG jumps and mortars. *Comput. Methods Appl. Mech. Engrg.*, 198:1751–1765, 2009.
- [27] V. Girault, S. Sun, M. F. Wheeler, and I. Yotov. Coupling discontinuous Galerkin and mixed finite element discretizations using mortar finite elements. *SIAM. J. Numer. Anal.*, 46:949–979, 2008.
- [28] P. Hauret and M. Ortiz. BV estimates for mortar methods in linear elasticity. *Comput. Methods Appl. Mech. Engrg.*, 195:4783–4793, 2006.
- [29] R. N. Horne, R. Stacey, and K. Li. Physical modeling of CO₂ sequestration. *California Energy Commission, PIER Energy-Related Environmental Research Program, CES-500-2007-113*, 2008.

- [30] J. Hyman, J. Morel, M. Shashkov, and S. Steinberg. Mimetic finite difference methods for diffusion equations. *Computational Geosciences*, 6:333–352, 2002.
- [31] R. Ingram, M. F. Wheeler, and I. Yotov. A multipoint flux mixed finite element method on hexahedra. *Technical Report TR-MATH 09-22, Dept. Math., University of Pittsburgh, Submitted to SIAM J. Numer. Anal.*, 2009.
- [32] M. Kim, E. Park, S. G. Thomas, and M. F. Wheeler. A multiscale mortar mixed finite element method for slightly compressible flows in porous media. *J. Korean Math. Soc.*, (5):1–15, 2007.
- [33] H. Klie and M. F. Wheeler. Krylov-secant methods for accelerating the solution of fully implicit formulations. *In SPE Reservoir Simulation Symposium, Proceedings, Houston, TX, SPE 92863:57–65*, 2005.
- [34] H. Klie, M. F. Wheeler, T. Kees, and K. Stuben. Deflation AMG solvers for highly ill- conditioned reservoir simulation problems. *In SPE Reservoir Simulation Symposium, Houston, Texas, SPE 105820*, 2007.
- [35] A. Kumar, M. Noh, G. A. Pope, K. Sepehrnoori, S. Bryant, and L. W. Lake. Reservoir simulation of CO₂ storage in deep saline aquifers. *In SPE/DOE Fourteenth Symposium on Improved Oil Recovery, Tulsa, Oklahoma, SPE-89343*, 2004.
- [36] Y. Kuznetsov, K. Lipnikov, and M. Shashkov. The mimetic finite difference method on polygonal meshes for diffusion-type problems. *Computational Geosciences*, 8:301–324, 2004.
- [37] S. Lacroix, Y. Vassilevski, M. F. Wheeler, and J. Wheeler. Iterative solvers of the implicit parallel accurate reservoir simulator (IPARS). *Numer. Lin. Algebra Appl.*, 4:537–549, 2001.
- [38] S. Lacroix, Y. Vassilevski, M. F. Wheeler, and J. Wheeler. Iterative solution methods for modeling multiphase flow in porous media fully implicitly. *SIAM J. Sci. Comput.*, 25:905–926, 2003.
- [39] C. S. Land. Calculation of imbibition relative permeability for two- and three-phase flow from rock properties. *Soc. Pet. Eng. J.*, 8:149–156, 1968.
- [40] K. Lipnikov, M. Shashkov, and I. Yotov. Local flux mimetic finite difference methods. *Numerische Mathematik*, 112:115–152, 2009.
- [41] J. Liu, M. Delshad, G. A. Pope, and K. Sepehrnoori. Application of higher order flux-limited methods in compositional simulations. *Transport in Porous Media*, 16(1):1–29, Jul 1994.
- [42] J. Lohrenz, B. G. Bray, and C. R. Clark. Calculating viscosities reservoir fluids from their compositions. *SPE 915, Journal of Petroleum Technology*, 1171–1176, 1964.
- [43] B. Lu. *Iteratively coupled reservoir simulation for multiphase flow in porous media*. PhD thesis, The University of Texas at Austin, 2008.
- [44] A. Menin, V. A. Salomoni, R. Santagiuliana, L. Simoni, A. Gens, and B. A. Schrefler. A mechanism contributing to subsidence above reservoirs and its application to a case study. *International Journal for Computational Methods in Engineering and Mechanics*, 9:270–287, 2008.

- [45] I. Moulitsas and G. Karypis. Partitioning algorithms for simultaneously balancing iterative and direct methods. *Technical Report 04-014, Department of Computer Science and Engineering, University of Minnesota, Minneapolis, MN*, 2004.
- [46] R. Naimi-Tajdar, C. Han, K. Sepehrnoori, T. Arbogast, and M. A. Miller. A fully implicit, compositional, parallel simulator for IOR processes in fractured reservoirs. *SPE Journal*, 12, 2007.
- [47] L. Oliker and R. Biswas. PLUM: Parallel load balancing for adaptive unstructured meshes. *Journal of Parallel and Distributed Computing*, 52:150–177, 1998.
- [48] G. Pencheva, M. Vohralik, M. F. Wheeler, and T. Wildey. A posteriori error control and adaptivity for multiscale, multinumerics, and mortar coupling. *To appear*, 2010.
- [49] G. Pencheva, M. F. Wheeler, and S. G. Thomas. Mortar coupling of discontinuous Galerkin and mixed finite element methods. *11th European Conference on the Mathematics of Oil Recovery (ECMOR), Bergen, Norway*, 2008.
- [50] D. Peng and D. B. Robinson. A new two-constant equation of state. *Ind. Eng. Chem., Fundam.*, 15(1):59–64, 1976.
- [51] M. Peszynska and S. Sun. Multiphase reactive transport module (TRCHEM) in ipars. Technical Report 01-32, CSM, ICES, The University of Texas at Austin, Austin, TX, Oct 2001.
- [52] M. Peszynska, M. F. Wheeler, and I. Yotov. Mortar upscaling for multiphase flow in porous media. *Comput. Geosci.*, 6:73–100, 2002.
- [53] E. Polizzi and A. Sameh. A parallel hybrid banded system solver: the SPIKE algorithm. *Parallel Computing*, 32:177–194, 2006.
- [54] P. A. Raviart and J. Thomas. A mixed finite element method for 2-nd order elliptic problems. In I. Galligani and E. Magenes, editors, *Mathematical aspects of the Finite Elements Method*, Lectures Notes in Math. 606, pages 292–315. Springer, Berlin, 1977.
- [55] T. Russell and M. Wheeler. Finite element and finite difference method for continuous flows in porous media. *Frontiers in Applied Mathematics*, 1:35, 1983.
- [56] T. F. Russell and M. F. Wheeler. Finite element and finite difference methods for continuous flows in porous media. In R. Ewing, editor, *Frontiers in Applied Mathematics*, volume 1, pages 35–106. SIAM, Philadelphia, PA, 1984.
- [57] B. A. Schrefler, R. Matteazzi, D. Gawin, and X. Wang. Two parallel computing methods for coupled thermohydromechanical problems. *Computer-Aided Civil and Infrastructure Engineering*, 15:176–188, 2000.
- [58] L. I. Stiel and G. Thodos. The viscosity of polar substances in the dense gaseous and liquid regions. *AiChE J.*, 10:275–277, 1964.
- [59] K. Stuben, T. Clees, H. Klie, B. Lu, and M. F. Wheeler. Algebraic multigrid methods for the efficient solution of fully implicit formulations in reservoir simulation. In *SPE Reservoir Simulation Symposium, Houston, TX*, 2007.
- [60] S. G. Thomas. On some problems in the simulation of flow and transport through porous media. *The University of Texas at Austin*, 2009.

-
- [61] S. G. Thomas, M. F. Wheeler, and M. Delshad. Modeling CO₂ sequestration using a sequentially coupled iterative-impec-time-split-thermal compositional simulator. *11th European Conference on the Mathematics of Oil Recovery, Bergen, Norway, September 8–11, 2008*.
- [62] S. G. Thomas, M. F. Wheeler, and M. Delshad. Parallel numerical reservoir simulations of non-isothermal compositional flow and chemistry. *SPE paper 118847, the 2009 SPE Reservoir Simulation Symposium, Houston, TX, Feb 2–4, 2009*.
- [63] M. Vohralik and B. Wohlmuth. Mixed finite element methods: all order schemes with one unknown per element, local flux expressions, positivity, polygonal meshes, and relations to other methods. *Preprint*, 2010.
- [64] W. Wang, G. Kosakowski, and O. Kolditz. A parallel finite element scheme for thermo-hydro-mechanical (THM) coupled problems in porous media. *Computers and Geosciences*, 35(1631–1641), 2009.
- [65] A. Weiser and M. F. Wheeler. On convergence of block-centered finite differences for elliptic problems. *SIAM Journal on Numerical Analysis*, 25:351–375, 1988.
- [66] M. F. Wheeler, G. Xue, and I. Yotov. Multipoint flux mixed finite element method on distorted quadrilaterals and hexahedra. *To appear*, 2010.
- [67] M. F. Wheeler, G. Xue, and I. Yotov. A multiscale mortar multipoint flux mixed finite element method. *To appear*, 2010.
- [68] M. F. Wheeler and I. Yotov. A posteriori error estimates for the mortar mixed finite element method. *SIAM J. NUMER. ANAL.*, 43(3):1021–1042, 2005.
- [69] M. F. Wheeler, T. Wildey, and G. Xue. Recent advances in multiscale mortar methods. *Preprint*, 2010.
- [70] M. F. Wheeler, T. Wildey, and I. Yotov. Frozen Jacobian multiscale mortar preconditioner for a frozen Jacobian multiscale mortar preconditioner for nonlinear interface operators. *Preprint*, 2010.
- [71] M. F. Wheeler and I. Yotov. A multipoint flux mixed finite element method. *SIAM. J. Numer. Anal.*, 44(5):2082–2106, 2006.

Fast Poisson-based Solvers for Linear and Nonlinear PDEs

Jinchao Xu*

Abstract

Over the last few decades, developing efficient iterative methods for solving discretized partial differential equations (PDEs) has been a topic of intensive research. Though these efforts have yielded many mathematically optimal solvers, such as the multigrid method, the unfortunate reality is that multigrid methods have not been used much in practical applications. This marked gap between theory and practice is mainly due to the fragility of traditional multigrid methodology and the complexity of its implementation. This paper aims to develop theories and techniques that will narrow this gap. Specifically, its aim is to develop mathematically optimal solvers that are robust and easy to use for a variety of problems in practice. One central mathematical technique for reaching this goal is a general framework called the *Fast Auxiliary Space Preconditioning* (FASP) method. FASP methodology represents a class of methods that (1) transform a complicated system into a sequence of simpler systems by using auxiliary spaces and (2) produces an efficient and robust preconditioner (to be used with Krylov space methods such as CG and GMRes) in terms of efficient solvers for these simpler systems. By carefully making use of the special features of each problem, the FASP method can be efficiently applied to a large class of commonly used partial differential equations including equations of Poisson, diffusion-convection-reaction, linear elasticity, Stokes, Brinkman, Navier–Stokes, complex fluids models, and magnetohydrodynamics. This paper will give a summary of results that have been obtained mostly by the author and his collaborators on this topic in recent years.

Mathematics Subject Classification (2010). Primary 65N55 and 65N22; Secondary 65N30.

Keywords. Finite element, FASP, auxiliary space preconditioning, method of subspace correction, adaptivity, multigrid, domain decomposition, nearly singular systems, near-null space recovery condition, $H(\text{grad})$, $H(\text{curl})$, $H(\text{div})$, saddle-point, non-Newtonian models, MHD.

*Department of Mathematics, Pennsylvania State University, University Park, PA 16802.
E-mail: xu@math.psu.edu.

1. Introduction

Most scientific and engineering problems can be modeled by using certain partial differential equations (PDEs). These PDEs usually have to be solved numerically. With appropriate discretizations, such as the finite element or finite difference methods, the numerical solution of each of these PDEs is often reduced to the solution of one linear algebraic system of equations or a sequence of such equations:

$$Au = f. \quad (1.1)$$

In fact, the solution of the underlying equations like (1.1) often accounts for a major portion of the work required for the numerical solution of a PDE system. How to efficiently solve (1.1) is thus of fundamental importance in scientific and engineering computing. From a purely mathematical point of view, solving equation (1.1) is trivial as long as A is non-singular: $u = A^{-1}f$. But the point of concern here is the amount of work (the computational complexity) required to find a solution. The most commonly used method in practice is still the classic Gaussian elimination that requires $\mathcal{O}(N^3)$ (in general about $\frac{1}{3}N^3$) floating point operations if (1.1) has N unknowns. Therefore, a naive application of Gaussian elimination for a system of, say, one million (10^6) unknowns (which is not a lot in today's applications) would be a formidable cost even on today's most powerful computers.

The aim of this paper is to discuss more advanced numerical algorithms for solving equations like (1.1) that arise from the discretization of PDEs. These types of equations often possess some special properties (related to the underlying PDEs and their discretizations) that can be exploited so that tailored algorithms that are much more efficient than the Gaussian elimination can be designed. Among the various possible algorithms, the multigrid (MG) method is generally considered to be one of the most powerful techniques for this task. Indeed, for a large class of equations, the efficiency of MG can be optimal or nearly optimal theoretically (namely it requires only $\mathcal{O}(N)$ or $\mathcal{O}(N \log N)$ operations for a linear system with N unknowns); see [15, 28, 13, 49, 55] for details. Yet, the multigrid method has not been used as much in practical applications. The method, especially the traditional (geometric) multigrid (GMG) method, has some limitations from a practical point of view. The most critical limitation is that a GMG method requires a hierarchical sequence of geometric grids, which is not available in most existing codes and is difficult to construct for domains with complicated geometry.

In recent years, considerable progress has been made in developing multigrid methods that are more easily applicable in practical applications. One remarkable example is the so-called algebraic multigrid method (AMG). A typical AMG method only require the user to input the coefficient matrix A and the

right-hand side data b in (1.1) and hence is quite user-friendly!¹ The efficiency of AMG, however, varies from one application to another. There is, therefore, still a long way from rendering the classic AMG technology robustly applicable in practice.

The methods we will discuss in this paper are built on the success of the AMG method for some special types of PDEs, such as the Poisson-like equation, specifically the Poisson equation and its variants. The method we are proposing is a combination of AMG (for Poisson-like equations) with various analytic and geometric properties pertaining to a given linear algebraic system arising from the discretization of certain classes of PDEs. Among the many important factors to be considered for a given algorithm studied, the two most important ones considered in this paper are (1) *efficiency*: the algorithm is as close as possible to being optimal, and (2) *practicality*: the algorithm is user-friendly; that is, it would not take an extraordinary amount of programming effort for an ordinary user. These are two competing factors, but with careful mathematical study, we will demonstrate that it is possible to strike a good balance between the two for a large class of problems. To this end, the following *four*-stage strategy will be adopted:

1. develop user-friendly optimal solvers and relevant theories for the discrete Poisson-like equations;
2. extend the list of solver-friendly PDE systems (for which optimal and user-friendly solvers can be applied), such as discrete Stokes and Maxwell equations, by reducing them to the solution of a handful of Poisson-like equations;
3. develop solver-friendly discretization techniques for more complicated PDE systems such that the discretized systems will join the list of solver-friendly systems (such as the Eulerian–Lagrangian method for the Navier–Stokes equations, the Johnson–Segalman equations, and the magnetohydrodynamics equations);
4. solve the discretized system from a general discretization by using a solver-friendly discretization as a component to construct a FASP method for the original system (in case the solver-friendly discretization is not suitable to obtain the numerical solution by itself.)

The purpose of this paper is to summarize the many results that the author and collaborators have obtained in recent years in regard to realizing the

¹We note that the term “*user-friendly*” is used in a rather loose and somewhat subjective way in this paper to describe those algorithms requiring relatively little extra programming efforts. An algebraic multigrid method that may or may not require the basic grid information (readily available by a standard finite element or finite difference code) is considered to be user-friendly, while a geometric multigrid method requiring user to define elaborate hierarchical grid structure may not be considered to be so.

aforestated four-stage strategy. Examples of equations to which our methods apply include Poisson, linear elasticity, biharmonic, convection-reaction-diffusion, Stokes, Navier–Stokes, non-Newtonian models, Maxwell, and MHD. Our studies are based on numerous earlier related works in the literature, but not all these works can be mentioned here due to the page limit; a more comprehensive presentation will be given in a future paper.

The rest of the paper is organized as follows. In §2, we discuss the main ideas of the Fast Auxiliary Space Preconditioning (FASP) method and algebraic multigrid methods for the Poisson equation and its variants. In §3, we present a list of solver-friendly systems that can be solved by FASP. In §4, we give an example of a solver-friendly discretization, namely the Eulerian–Lagrangian method (ELM). In §5 and §6, we demonstrate how the ELM can be used to discretize a popular non-Newtonian fluid model and a model MHD equation, respectively, so that the resulting discrete systems are solver-friendly. We conclude the paper by offering a brief commentary in §7 along with a table of PDE systems that can be solved using FASP methods and by outlining some plan of future works.

2. The FASP and AMG Methods

A linear iterative method for solving a linear algebraic system of equations $Au = f$ can, in general, be written in the following form:

$$u^n = u^{n-1} + B(f - Au^{n-1}) \quad (n = 1, 2, 3, \dots),$$

where B can be viewed as an approximate inverse of A . As simple examples, if $A = (a_{ij}) \in \mathbb{R}^{N \times N}$ and $A = D - L - U$, we can take $B = D^{-1}$ to obtain the Jacobi method and $B = (D - L)^{-1}$ to obtain the Gauss–Seidel method.

The approximate inverse B , when it is symmetric and positive-definite (SPD), can be used as a preconditioner for the Conjugate Gradient (CG) method. The resulting method, known as the preconditioned conjugate gradient (PCG) method, admits the following error estimate:

$$\frac{\|u - u^n\|_A}{\|u - u^0\|_A} \leq 2 \left(\frac{\sqrt{\kappa(BA)} - 1}{\sqrt{\kappa(BA)} + 1} \right)^k \quad (k \geq 1), \quad \text{with } \kappa(BA) = \frac{\lambda_{\max}(BA)}{\lambda_{\min}(BA)}.$$

For non-SPD systems, MinRes and GMRes are often used.

There are various approaches to the construction of B , an approximate inverse or a preconditioner of A . One major example is the method of subspace corrections [14, 55]. This type of method aims to precondition a system of equations in a vector space by solving some appropriately chosen subspace problems of the original space. When there is a lack of adequate subspaces, the auxiliary space method (Xu [57], Hiptmair and Xu [32]) can be used for designing preconditioners using some auxiliary spaces that are not necessarily subspaces of the original space.

2.1. FASP: Fast Auxiliary Space Preconditioning. A general mathematical framework, the Fast Auxiliary Space Preconditioning (FASP) method, was first proposed in [57]; and it will be used to derive and analyze most of the algorithms presented in this paper. FASP gives a preconditioner for a symmetric positive-definite system $Au = f$ on a vector space V , equipped with an inner product $a(\cdot, \cdot) = (A\cdot, \cdot)$, by using solvers on the following product of auxiliary spaces:

$$\bar{V} = V \times W_1 \times \cdots \times W_J, \tag{2.2}$$

where W_1, \dots, W_J and $J \in \mathbb{N}$ are auxiliary (Hilbert) spaces endowed with inner products $\bar{a}_j(\cdot, \cdot) = (\bar{A}_j\cdot, \cdot)$, $j = 1, \dots, J$. With appropriate transformation operators $\Pi_j : W_j \mapsto V$ for each j , we have the following preconditioner:

$$B = S + \sum_{j=1}^J \Pi_j \bar{A}_j^{-1} \Pi_j^*. \tag{2.3}$$

A distinctive feature of the auxiliary space method is the presence of V in (2.2) as a component of \bar{V} and the presence of the operator $S : V \mapsto V$, which is usually called the smoother. It can be proved that the preconditioner (2.3) admits the estimate

$$\kappa(BA) \leq c_0^2(c_s^2 + c_1^2 + \cdots + c_J^2), \tag{2.4}$$

where $a(\Pi_j w_j, \Pi_j w_j) \leq c_j^2 \bar{a}(w_j, w_j)$, $w_j \in W_j$, $a(SAv, v) \leq c_s^2 a(v, v)$ for any $v \in V$; and, for each $v \in V$, there are $v_0 \in V$ and $w_j \in W_j$, such that $v = v_0 + \sum_{j=1}^J \Pi_j w_j$ and $a(SAv_0, v_0) + \sum_{j=1}^J \bar{a}_j(w_j, w_j) \leq c_0^2 a(v, v)$.

The estimate (2.4) can be improved in many ways, but the version presented here is sufficient for the applications discussed in this paper. An important special case of the FASP method is the Method of Subspace Correction (MSC) [55, 60] in which $W_j \subset V$ for all j . Similar to MSC, FASP can also have many variants. B in (2.3) can be called the parallel or additive FASP, and we can naturally have a successive or multiplicative FASP in which all the auxiliary spaces are used one after other. Instead of discussing details of the successive FASP method, let us now discuss a partially parallel and partially successive method. In (2.3), we set $\bar{B} = \sum_{j=1}^J \Pi_j \bar{A}_j^{-1} \Pi_j^*$ and we consider the following partially successive correction method ($u^{n-1} \rightarrow u^n$):

$$\begin{cases} u^{n-\frac{2}{3}} &= u^{n-1} + S(f - Au^{n-1}), \\ u^{n-\frac{1}{3}} &= u^{n-\frac{2}{3}} + \bar{B}(f - Au^{n-\frac{2}{3}}), \\ u^n &= u^{n-\frac{1}{3}} + S(f - Au^{n-\frac{1}{3}}). \end{cases} \tag{2.5}$$

It is easy to see that $u - u^n = (I - \tilde{B}A)(u - u^{n-1})$ where

$$I - \tilde{B}A = (I - SA)(I - \bar{B}A)(I - SA). \tag{2.6}$$

The problem with this type of successive correction method is that it may not be convergent, if \bar{B} is not properly scaled so that $\rho(I - \bar{B}A) < 1$. But the following simple result is quite useful in practice.

Theorem 2.1. *Assume that $S, \bar{B} : V \mapsto V$ is such that $\rho(I - SA) < 1$ and \bar{B} is SPD. Then the operator \tilde{B} defined in (2.6) is also SPD.*

Proof. By (2.6), we have, for any $v \in V$, that

$$(\tilde{B}Av, v)_A = (v, v)_A - ((I - SA)v, (I - SA)v)_A + (\bar{B}A(I - SA)v, (I - SA)v)_A.$$

For $v \in V \setminus \{0\}$, the first two terms combined on the right-hand side are positive by the assumption that $\rho(I - SA) < 1$ and the third term is nonnegative since \bar{B} is assumed to be SPD. \square

The above theorem also provides a general approach for enhancing an existing preconditioner \bar{B} by combining it with another convergent iterative method (such as smoother S) to obtain a new preconditioner. Our experiences have shown that such a simple process can sometimes lead to significant improvement in the performance of either S or \bar{B} .

We finally point out that the FASP method can also be generalized to nonsymmetric and/or indefinite problems.

2.2. AMG for discrete Poisson equations and variants. The Poisson equation $-\Delta u = f$ and its variants $(-\nabla \cdot (\mu(x)\nabla u) + c(x)u = f)$ arise in many applications. When these equations are discretized (by either the finite difference or the finite element method) on uniform grids on a tensor-product domain (such as a square or cube), solvers based on the Fast Fourier Transform (FFT) can be used. The FFT cannot, however, be used to solve discrete Poisson equations on irregular domains. For discrete problems where a hierarchy of grids can easily be obtained (such as a grid obtained by uniform refinements from a coarse grid on a polygon or polyhedron), geometric multigrid methods can be used. But geometric multigrid methods are often not user-friendly.

For more user-friendly multigrid methods, we turn to algebraic multigrid (AMG) methods. What makes AMG attractive in practical terms is that it generates coarse-level equations without using much geometric information or re-discretization on the coarse levels. Despite the lack of rigorous theoretical justification, in most cases AMG methods are very successful in practice for various Poisson-like equations and, in recent years, many AMG techniques and relevant subroutines have been developed for solving these and even more general equations, cf. [47].

AMG is still a subject of extensive research. Among the many different AMG approaches, are the Ruge-Stuben [47], smoothed aggregation [51], multi-graph [6] and energy-minimization [52, 61, 16] methods. Most of the existing AMG methods emphasize their purely algebraic nature; that is, they only use the underlying algebraic properties of the coefficient matrix and the right-hand side data. In our approach, though, we advocate using additional information to make the method more robust and more efficient. The idea is that we should use as much information as the user is able to/is willing to provide. As demonstrated in Shu, Sun, and Xu [46], a little bit of extra information such as the

type of finite element discretization could lead to a significant improvement in the efficiency of AMG. We emphasize that information pertaining to the underlying PDEs, finite element spaces and grids should be used as much as possible. While the use of the geometric grid makes the corresponding AMG slightly less convenient to use, the method is still user-friendly, as it only requires the input of grid arrays that are usually readily available.

2.3. A FASP AMG method based on the auxiliary grid.

As an example of how geometric information can be used to design AMG, let us describe briefly an AMG method based on the FASP framework for solving the Poisson equation discretized on an unstructured grid in both 2D and 3D dimensions. For more details, we refer to [57, 27].

Consider a homogeneous Dirichlet boundary condition problem for $-\Delta u = f$ on a domain $\Omega \subset R^n$ ($1 \leq n \leq 3$). Assuming that Ω is triangulated by a shape-regular grid as shown on the left figure in Fig 1, let $V \subset H_0^1(\Omega)$ be a linear finite element space on this unstructured grid. We are interested in constructing a FASP method for the finite element equation on V . To do this, we construct an auxiliary structured grid (which has the same local grid density as the original grid), by successively refining a uniform grid on those elements that intersect Ω , as shown on the right in Fig 1. A finite element space $W \subset H_0^1(\Omega)$ is associated with this auxiliary structured grid. This is the auxiliary space for V that consists of all finite element functions that vanish on all the elements that are not *completely* inside $\bar{\Omega}$. To facilitate the FASP method, we use a standard

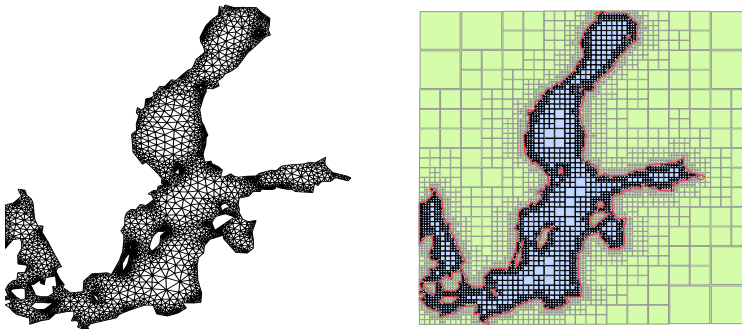


Figure 1. An unstructured grid and its auxiliary grid

nodal value interpolation $\Pi : W \mapsto V$ and a weighted inner product on V : $s(u, v) = (h^{-1}u, h^{-1}v)$. We claim that the resulting FASP preconditioner B leads to a condition number $\kappa(BA)$ that is uniformly bounded with respect to mesh parameters.

In fact, by localizing the analysis in Xu [57], we have, for the nodal-value interpolation $\Pi : W \mapsto V$ and $\Pi_0 : V \mapsto W$, that for $w \in W$ and $v \in V$,

$$\|h^{-1}(w - \Pi w)\| + |\Pi_1 w|_{1, \Omega} \lesssim |w|_{1, \Omega}, \text{ and } \|h^{-1}(v - \Pi_0 v)\| + |\Pi_0 v|_1 \lesssim |v|_1.$$

Using these estimates with the inverse inequality, $|v|_{1,\Omega} \lesssim \|h^{-1}v\|$, we have

$$\begin{aligned} s(v_0, v_0)^{1/2} &= \|h^{-1}(v - \mathbb{I}\mathbb{I}\mathbb{I}_0 v)\| \leq \|h^{-1}(v - \Pi_0 v)\| + \|h^{-1}(\Pi_0 v - \mathbb{I}\mathbb{I}\mathbb{I}_0 v)\| \\ &\lesssim |v|_1 + |\Pi_0 v|_1 \lesssim |v|_1. \end{aligned}$$

Optimal condition number estimates then follow from the estimate (2.4) by $v = v_0 + \Pi w$ with $w = \mathbb{I}\mathbb{I}\mathbb{I}_0 v$ and $v_0 = v - \Pi w$.

The auxiliary grid method outlined above was first developed in Xu [57] for quasi-uniform grids and then extended to general shape-regular unstructured grids in a recent work [27]. In [27], we developed an $\mathcal{O}(N \log N)$ algorithm to construct a hierarchical structured grid that has the same local density as the original grid with N vertices. This resulting FASP method can be viewed as a nonnested two-grid method because it makes use of a “structured” hierarchical grid as a “coarse” space for the original finite element space; the coarser grid equation is further solved by a standard nested geometric multigrid method. The convergence analysis for such a nested geometric multigrid can be established by using the techniques in Chen, Nochetto, and Xu [19] (see also Wu and Chen [53]). By combining all these results, we conclude that *a discrete Poisson equation on a general shape-regular unstructured grid can be solved with $\mathcal{O}(N \log N)$ operations by using the conjugate gradient method preconditioned by a FASP method.*

2.4. Building blocks: Fast solvers for Poisson-like systems.

Based on the theoretical results for optimal AMG given above (see [27] for details), and other vast numbers of existing computational experiences, we make the following basic assumption:

Assumption P. *The discretized system of the Poisson-like equation*

$$-\nabla \cdot (\mu(x)\nabla u) + c(x)u = f \quad (2.7)$$

is solver-friendly; that is, it can be solved efficiently (sometimes with optimal or nearly optimal computational complexity) by using geometric or algebraic multigrid methods (or a combination of the two) in a user-friendly way.

One central strategy of this paper is to adopt user-friendly solvers (such as AMG, which are either in existence or need to be further developed) for (2.7) to develop user-friendly solvers for more complicated PDE systems.

More specifically, in our development of efficient solvers for PDEs, we will propose core algebraic solvers (such as the multigrid method) mainly for the Poisson-like equation (2.7). We will then use mathematical techniques (such as FASP) and special discretization schemes (such as the Eulerian–Lagrangian method in § 4) to reduce the solution of other more complicated PDEs into the solution of a handful equations like (2.7).

While considerable work is still required to develop new technologies and also improve existing ones for (2.7), but, thanks to the contributions of many researchers during the last few decades, most of equations in the form of (2.7) are indeed solver-friendly. For example, the following boundary value problem with a highly oscillatory coefficient satisfying $0 < \mu_0 < \mu(x) < \mu_1$:

$$-\nabla \cdot (\mu(x/\epsilon)\nabla u) = f, \quad (2.8)$$

is also solver-friendly when it is discretized by a direct application of the finite element method (without using any numerical homogenization techniques). In fact, we can easily precondition it with the simple Poisson equation $-\Delta u = f$ due to this simple relation: $\mu_0(\nabla v, \nabla v) \leq (\mu\nabla v, \nabla v) \leq \mu_1(\nabla v, \nabla v)$. As a result, a direct finite element discretization of (2.8) is not much more difficult to solve than a Poisson equation.

3. Solver-friendly Systems

Using fast Poisson solvers as building blocks, we can develop user-friendly solvers for various discretized PDEs. In this section, we will identify a list of solver-friendly systems for which user-friendly solvers can be designed in terms of one or more basic solver-friendly Poisson-like systems.

First, we will study the H(grad), H(curl), and H(div) systems:

$$\text{Find } u \in V : \quad D^*(\mu(x)Du) + c(x)u = f. \quad (3.9)$$

Here $V = H(D) := \{v \in L^2(\Omega) : Dv \in L^2(\Omega)\}$ with the following special cases: $D = \text{grad}$ (Poisson equations), $D = \text{curl}$ (Maxwell equations), and $D = \text{div}$ (mixed finite elements for Darcy's law).

We will then study the following mixed systems:

$$\begin{cases} cu - \nabla \cdot (\mu\nabla u) + \nabla p = f & \text{in } \Omega, \\ \nabla \cdot u + \gamma p = g & \text{in } \Omega. \end{cases} \quad (3.10)$$

The following special cases are of particular interest: (1) the Stokes equation ($\mu > 0, \gamma = c = 0$), (2) linear elasticity ($\mu > 0, \gamma > 0, c = 0$), (3) mixed formulation for Darcy's Law ($\mu = 0, \gamma = 0, c > 0$), and (4) the Brinkman model to couple the Stokes equation and Darcy's Law ($\mu > 0, \gamma = 0$ and c is piecewise constant). With the proper choice of finite element spaces, the above systems can be discretized in a stable fashion. In fact, it is even possible to develop a discretization scheme that is uniformly stable with respect to all the aforesaid parameters (cf. [54]).

3.1. $H(\text{grad})$, $H(\text{curl})$, and $H(\text{div})$ systems. For the system (3.9), the main properties of various relevant spaces and operators are summarized in the following exact sequences and commutative diagrams:

$$\begin{array}{ccccccccccc}
 \mathbb{R} & \longrightarrow & C^\infty & \xrightarrow{\text{grad}} & C^\infty & \xrightarrow{\text{curl}} & C^\infty & \xrightarrow{\text{div}} & C^\infty & \longrightarrow & 0 \\
 & & \downarrow \Pi_h^{\text{grad}} & & \downarrow \Pi_h^{\text{curl}} & & \downarrow \Pi_h^{\text{div}} & & \downarrow \Pi_h^0 & & \\
 \mathbb{R} & \longrightarrow & H_h(\text{grad}) & \xrightarrow{\text{grad}} & H_h(\text{curl}) & \xrightarrow{\text{curl}} & H_h(\text{div}) & \xrightarrow{\text{div}} & L_h^2 & \longrightarrow & 0.
 \end{array}$$

Geometric multigrid methods have been studied in the literature for all these systems, e.g. see [28, 13, 55, 62] for the $H(\text{grad})$ systems and [23, 29, 30, 3, 4, 5] for the $H(\text{curl})$ and $H(\text{div})$ system. While AMG methods are well-developed for $H(\text{grad})$ systems, very few robust AMG methods have been developed for $H(\text{curl})$ and $H(\text{div})$ systems. One main difficulty is that both curl and divergence operators have large (near-) null spaces, which are not easily recoverable algebraically; on the other hand for $H(\text{grad})$, the null space of the gradient operator is at most one-dimensional and can easily be recovered algebraically. For related works, we refer to [7, 41, 10].

As the $H(\text{grad})$ systems are just the Poisson-like system discussed in (2.2), we will now study the $H(\text{curl})$ and $H(\text{div})$ systems.

$H(\text{curl})$ systems By means of the auxiliary space method framework by Xu [57], a family of preconditioners is obtained in [32]. For $H(\text{curl})$ systems, we have the optimal and user-friendly preconditioner:

$$B_h^{\text{curl}} := S_h^{\text{curl}} + \Pi_h^{\text{curl}} \mathbf{B}_1^{\text{grad}} (\Pi_h^{\text{curl}})^T + \text{grad } B_2^{\text{grad}} (\text{grad})^T, \tag{3.11}$$

where $\mathbf{B}_1^{\text{grad}}$ is a user-friendly preconditioner (see 2.2) for the *vectorial* $H(\text{grad})$ system or Poisson-like system $(\Pi_h^{\text{curl}})^T \mathbf{A}_h^{\text{curl}} \Pi_h^{\text{curl}}$ (see [36]) and B_2^{grad} is a user-friendly preconditioner for the operator $-\text{div}(\mu \text{grad})$ or $(\text{grad})^T \mathbf{A}_h^{\text{curl}} \text{grad}$. This preconditioner and its variants have been included and tested in LLNL’s *hypr* package [24] based on the parallel algebraic multigrid solver for Poisson equations; see the scalability test in Figure 2. Extensive numerical experiments demonstrate that this preconditioner is also efficient and robust for problems in which μ and c may be discontinuous, degenerating, and/or largely variant (see Hiptmair and Xu [32], and Kolev and Vassilevski [36]). The above FASP for the $H(\text{curl})$ system (which is called the Auxiliary-space Maxwell Solver (AMS)) along with its software package by Kolev and Vassilevski (see [36]) at the Lawrence Livermore National Laboratory (LLNL) has been featured in [22] as one of the *ten breakthroughs* in computational science in recent years.

$H(\text{div})$ systems Similarly, for the $H(\text{div})$ systems, we have

$$B_h^{\text{div}} := S_h^{\text{div}} + \Pi_h^{\text{div}} \mathbf{B}_3^{\text{grad}} (\Pi_h^{\text{div}})^T + \text{curl } B_h^{\text{curl}} (\text{curl})^T, \tag{3.12}$$

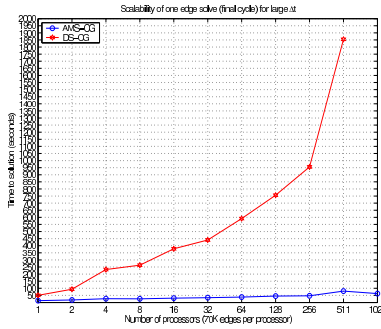


Figure 2. Parallel scalability test for the AMS preconditioner in *hypr* [22].

Quote from [22]: “AMS is the first solver for Maxwell’s equations that demonstrates theoretically supported weak parallel scalability. It has now been incorporated into several LLNL physics codes previously limited by their Maxwell solvers, most noticeably in terms of resolution. AMS has been tested in a number of applications and demonstrated a significant (4 to 25 times) improvement in the solution time when run on a large number of processors.”

where $\mathbf{B}_3^{\text{grad}}$ is a user-friendly preconditioner for the vectorial $H(\text{grad})$ system as in (3.9) (with $D = \text{grad}$) or $(\Pi_h^{\text{div}})^T \mathbf{A}_h^{\text{div}} \Pi_h^{\text{div}}$ and B_h^{curl} is given in (3.11).

We note that the preconditioners (3.11) and (3.12) are two typical examples of algorithms regarded as *user-friendly*. In the implementation of these two preconditioners, an existing user-friendly AMG-type solver is used together with smoothers and transformation operators that depend only on the stiffness matrix and basic geometric information of the finite element grid.

3.2. Mixed finite element methods. When the mixed finite element method is used to discretize incompressible porous media flow with Darcy’s Law, a symmetric but indefinite system (the discrete version of (3.10) with $\mu = \gamma = 0$ and $c > 0$) arise [17]. There are at least three different ways to develop user-friendly solvers for this indefinite system.

The first approach is to use the preconditioned MinRes method with a diagonal preconditioner $\text{diag}(I, (\text{div}^* \text{div} + I)^{-1})$ (see [43, 3]). This procedure can be made user-friendly if the preconditioner (3.12) is used for the $H(\text{div})$ system.

The second approach is to use an augmented Lagrangian method [26] based on the following equivalent formulation, with any $\epsilon > 0$):

$$\begin{bmatrix} A & B^* \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ p \end{bmatrix} = \begin{bmatrix} f \\ g \end{bmatrix} \iff \begin{bmatrix} A + \epsilon^{-1} B^* B & B^* \\ B & 0 \end{bmatrix} \begin{bmatrix} \mathbf{u} \\ p \end{bmatrix} = \begin{bmatrix} f + \epsilon^{-1} B^* g \\ g \end{bmatrix}.$$

We apply a simple Uzawa method to the above augmented system:

$$(A + \epsilon^{-1} B^* B) \mathbf{u}^{(k+1)} = f + \epsilon^{-1} B^* (g - p^{(k)}), p^{(k+1)} = p^{(k)} - \epsilon^{-1} (g - B \mathbf{u}^{(k+1)}).$$

Based on the error estimates [37] that $\|p - p^{(k)}\|_{0,\Omega} = \mathcal{O}(\epsilon^k)$ and $\|\mathbf{u} - \mathbf{u}^{(k)}\|_A = \mathcal{O}(\epsilon^{k+1/2})$, the Uzawa iteration converges within one or two iterations if $\epsilon \ll 1$. When $\epsilon \ll 1$, the nearly singular SPD matrix $A + \epsilon^{-1} B^* B = I + \epsilon^{-1} \text{div}^* \text{div}$ can be solved efficiently by the preconditioner (3.12) (see also [31] for the geometric multigrid method). As a result, the solution of the indefinite mixed

system is reduced to several Poisson equations in a user-friendly way. Preliminary results in Xu and Zhu [50] demonstrated that this is potentially a very efficient approach.

The third approach is to use a Lagrangian multiplier [1] to convert the indefinite system to a symmetric positive-definite system. This system is closely related to and sometimes equivalent to the system discretized by certain non-conforming finite element methods for the original Poisson-like PDE. Theoretical as well as numerical studies have shown that both geometric and algebraic multigrid methods can be applied efficiently to this system, see [18, 33].

3.3. Stokes equations. We consider a generalized Stokes system:

$$\begin{pmatrix} A & \nabla_h \\ -\nabla \cdot & 0 \end{pmatrix} \begin{pmatrix} u_h \\ p_h \end{pmatrix} = \begin{pmatrix} f_h \\ 0 \end{pmatrix}, \quad A = cI - \mu\Delta_h.$$

Many iterative methods have been developed for this system. Here we are mainly interested in user-friendly solvers. We apply the MinRes method with the diagonal block preconditioner [43, 8, 12]; namely, $\mathcal{P} = \text{diag}(\mathcal{P}_A, \mathcal{P}_S)$, where \mathcal{P}_A is a multigrid preconditioner for the matrix A and \mathcal{P}_S is a preconditioner corresponding to the Schur complement. The matrix A has a block-diagonal form with each diagonal block corresponding to a scalar Poisson-like equation that is solver-friendly. And, the Schur complement preconditioner can be chosen to be $\mathcal{P}_S = \mu D_M^{-1} + c(-\Delta_N)^{-1}$, where D_M is the diagonal of the mass matrix for the pressure space and $-\Delta_N$ is the auxiliary Laplace operator with the Neumann boundary condition.

The preconditioned minimum residual (MinRes) with this preconditioner is shown to be uniform with respect to c, μ and the size of the problem. Furthermore, this method is user-friendly and easily parallelizable because $(-\Delta)^{-1}$ can be replaced by a fast Poisson solver as described in §2.

3.4. Darcy–Stokes–Brinkman model. This model has been used for modeling the coupling of a porous media flow occupied in Ω_1 and a Stokes flow in Ω_2 :

$$-\nabla \cdot (\mu(x)\nabla u) + c(x)u + \nabla p = f, \quad \nabla \cdot u = g \quad (3.13)$$

where $\mu = \mu_i \in (0, 1]$ and $c = c_i \in [0, 1]$ for $x \in \Omega_i$ are constant on each Ω_i ($i = 1, 2$). According to Xie, Xu, and Xue [54], any stable Stokes element is uniformly stable for the following slightly modified and equivalent system:

$$-\nabla \cdot (\mu(x)\nabla u) - \nu\nabla\nabla \cdot u + c(x)u + \nabla p = f - \nu\nabla g, \quad \nabla \cdot u = g \quad (3.14)$$

where $\nu = \max(\mu_1, \mu_2, c_1, c_2, 1)$. For (3.14), the standard Stokes element is uniformly stable and the Schur complement is uniformly well-conditioned. Hence we can use the MinRes together with a block diagonal preconditioner $\text{diag}(A^{-1}, D_M^{-1})$ to solve this system. Here $Au = -\nabla \cdot (\mu(x)\nabla u) - \nu\nabla\nabla \cdot u + c(x)u$. The operator A (similar to the linear elasticity operator) is potentially solver-friendly. It is related to the problem considered in [44] where proper geometric

multigrid methods are proven to be robust with respect to large variations in the coefficients μ and c , as well as in regard to the size of the problem. More user-friendly robust solvers for this problem, however, require further research.

3.5. Plate models.

Kirchhoff plate As an example of high-order partial differential equations, we will demonstrate how Poisson equations can be used for the numerical solution of the biharmonic equation $\Delta^2 u = f$ on a polygonal domain $\Omega \subset \mathbb{R}^2$: Find $u \in H_0^2(\Omega)$, such that

$$a(u, v) := (\nabla^2 u, \nabla^2 v) = (f, v), \quad \forall v \in M. \tag{3.15}$$

As a common practice for 4th-order problems, we introduce an intermediate variable $v = -\Delta u$ for discretization and obtain a corresponding mixed finite element discretization. This type of mixed finite element is, however, not an appropriate discretization for (3.15) for various reasons (such as its lack of optimality) and in fact, for simply supported plate problems, such a mixed formulation could lead to the wrong approximation of (3.15) when the domain Ω is concave [63],

A more reliable finite element discretization for the original variational problem (3.15) can be obtained by choosing either conforming (such as Argyris and Bell) or nonconforming finite element methods (such as Morley, Zienkiewicz, Nilssen–Tai–Winther, Morley–Zienkiewicz). Let M_h be any such finite element space, and find $u_h \in M_h$, such that

$$a_h(u_h, v_h) := \sum_{K \in T_h} (\nabla^2 u_h, \nabla^2 v_h)_{L^2(K)} = (f, v_h), \quad v_h \in M_h. \tag{3.16}$$

We now propose to use two linear finite element spaces, $V_h \subset H^1(\Omega)$ and $V_{h,0} \subset H_0^1(\Omega)$, for the Poisson equation as auxiliary spaces to construct a FASP method for (3.16) based on the following stable decomposition [59]:

$$|w_h|_{2,h}^2 \approx \sum_T h_T^{-4} \|w_h - I_V^C w_h\|_{L^2,T}^2 + \|\tilde{\Delta}_h I_V^C w_h\|^2, \quad \forall w_h \in M_{h,0}, \tag{3.17}$$

where $I_V^C : M_{h,0} \mapsto V_{h,0}$ and $I_C^V : V_{h,0} \mapsto M_{h,0}$ are interpolants based on simple averages. Now we define $A_h : M_h \mapsto M_h$ by $(A_h u_h, v_h) = a_h(u_h, v_h)$ for $u_h, v_h \in M_h$ and $\Delta_h : V_{h,0} \mapsto V_{h,0}$ and $\tilde{\Delta}_h : V_{h,0} \mapsto V_h$ by $(-\Delta_h u_h, v_h) = (\nabla u_h, \nabla v_h)$ and $(-\tilde{\Delta}_h u_h, w_h) = (\nabla u_h, \nabla w_h)$ for $u_h, v_h \in V_{h,0}, w_h \in V_h$. We obtain a basic FASP method for (3.16) as follows:

$$\tilde{B}_h = S_h + I_C^V (\tilde{\Delta}_h^T \tilde{\Delta}_h)^{-1} (I_V^C)^T. \tag{3.18}$$

Here $S_h : M_{h,0} \mapsto M_{h,0}$ represents any smoother such as the symmetric Gauss–Seidel method. It can be proved that $\kappa(\tilde{B}_h A_h)$ is uniformly bounded. We can

simply replace $\tilde{\Delta}_h^T \tilde{\Delta}_h$ by Δ_h^2 to obtain a simplified preconditioner:

$$B_h = S_h + I_C^V \Delta_h^{-2} (I_C^V)^T. \tag{3.19}$$

It can be proved that $\kappa(B_h A_h) = \mathcal{O}(h^{-1})$. With a more sophisticated approach, we can precondition $\tilde{\Delta}_h^T \tilde{\Delta}_h$ optimally by combining Δ_h^2 with some boundary operations (which involve more Poisson solvers); see [59] for details.

A preconditioner similar to (3.18) was first obtained in [11] for Morley elements discretized on uniform grids by a different approach. A more general derivation and analysis was given in [59] using the FASP framework.

Reissner-Mindlin model Consider the Reissner-Mindlin model for a plate of thickness $t \in (0, 1)$:

$$\begin{cases} -\operatorname{div} \mathcal{C}D(\phi) - \zeta = 0, \\ -\operatorname{div} \zeta = g, \\ -\phi + \nabla \omega - \lambda^{-1} t^2 \zeta = 0, \end{cases} \tag{3.20}$$

on Ω with suitable boundary conditions. Here $D(\phi) = \frac{1}{2}(\nabla \phi + (\nabla \phi)^T)$, the scalar constant λ and tensor \mathcal{C} depend on the material properties of plate. With a class of appropriate discretizations for this model, Arnold, Falk and Winther [2] proposed the following preconditioner:

$$B = \operatorname{diag}((-\Delta_h)^{-1}, (-\Delta_h)^{-1}, I + (1 - t^2) \operatorname{curl}_h S_{t,h}^{-1} \operatorname{rot}_h) \tag{3.21}$$

and they proved this preconditioner leads to a uniformly (with respect to both h and t) convergent MinRes method for the discretized system of (3.20). Here

$$S_{t,h} = I + t^2 \operatorname{rot}_h \operatorname{curl}_h, \tag{3.22}$$

rot_h and $\operatorname{curl}_h (= \operatorname{rot}_h^*)$ are discrete versions of the operators:

$$\operatorname{rot} = (\partial/\partial y, -\partial/\partial x) \text{ and } \operatorname{curl} = \begin{pmatrix} -\partial/\partial y \\ \partial/\partial x \end{pmatrix}.$$

We notice that $S_{t,h}$ defined by (3.22) is a Laplacian-like operator, which can be preconditioned by a fast Poisson-like solver. As a result, the preconditioner (3.21) is user-friendly and its action amounts to, roughly speaking, four Poisson solvers.

4. Solver-friendly Eulerian–Lagrangian Method

We consider a discretization solver-friendly for a PDE if it yields some discrete solver-friendly systems (see §3). And, there are many discretization methods in the literature that can be categorized as solver-friendly. For example, the

two-grid method developed in [56, 58] can be viewed as a solver-friendly discretization, as it transforms a certain class of non-selfadjoint or nonlinear PDE systems into solver-friendly systems by using an extra coarse space.

One popular solver-friendly discretization known as the projection method has been much used for solving incompressible Navier–Stokes equations (NSEs) for Newtonian fluid flow:

$$(u_t + (u \cdot \nabla)u) - \text{Re}^{-1} \Delta u + \nabla p = 0 \text{ and } \nabla \cdot u = 0, \quad (4.23)$$

where Re is the Reynolds number, u is the velocity field, and p is the pressure.

The main idea of the projection method [20, 48] for Navier–Stokes equations is to transfer the following semi-implicit discretization at each time step:

$$\frac{1}{k}(u^{n+1} - u^n) - \text{Re}^{-1} \Delta u^{n+1} + \nabla p^{n+1} = -(u^n \cdot \nabla)u^n \text{ and } \nabla \cdot u^{n+1} = 0 \quad (4.24)$$

to a handful of Poisson equations for both velocity and pressure variables. As the Poisson equation discretized on a uniform grid could be solved by the Fast Poisson Solvers based on FFT, the projection method has been efficient for solving Navier–Stokes equations discretized on regular domains such as squares and cubes. Because of the availability of Fast Poisson Solvers for unstructured grids based on multigrid methods as noted in §2.2, the projection method should also be an efficient method for solving Navier–Stokes equations discretized on unstructured grids for domains with complicated geometries.

Given the availability of the fast Stokes solver (see §3.3), we argue that a natural way for solving semi-implicit discretization schemes (4.24) is solving the whole underlying Stokes equation directly without using any extra manipulations, e.g., on the artificial boundary conditions for pressure. While this method was not favored before as an efficient Stokes solver was not available, the situation is different now because Stokes equations can now be solved by an optimal solver on general unstructured grids.

The Eulerian–Lagrangian method Given a velocity vector field u , we define the flow map $\phi_{s,t}$ as follows:

$$\frac{d}{ds} \phi_{s,t}(x) = u(\phi_{s,t}(x), s), \quad \phi_{t,t}(x) = x. \quad (4.25)$$

The material derivative of v (a scalar, vector or tensor) is defined by

$$\frac{Dv}{Dt} = \frac{d}{ds} \phi_{s,t}^* v \Big|_{s=t} = \frac{d}{ds} v(\phi_{s,t}(x), s) \Big|_{s=t} = (v_t + (u \cdot \nabla)v)(x, t). \quad (4.26)$$

Here $\phi_{s,t}^*$ is called the pull-back operator for ϕ : $\phi_{s,t}^* u = u(\phi_{s,t}(x), s)$.

By approximating the particle trajectories, the Eulerian–Lagrangian [21, 40] (finite element) method seeks the positions of the particles at the previous time

(t^{n-1}) that have reached quadrature points at the current time (t^n) . As a result, we obtain a symmetric semi-discrete problem:

$$\frac{1}{k}(u^n - u_*^{n-1}) - \frac{1}{\text{Re}}\Delta u^n + \nabla p^n = 0 \quad \text{and} \quad \nabla \cdot u^n = 0, \quad (4.27)$$

where $k = t^n - t^{n-1}$ and $u_*^{n-1} = u(x_*^{n-1}, t^{n-1})$ is the velocity field at t^{n-1} evaluated at the position $x_*^{n-1} := \phi_{t_{n-1}, t_n}(x)$. We note that the ELM naturally works for reaction-convection-diffusion (R-C-D) equations also.

Since the invention of the ELM, many researchers have developed a number of variants and used them for different applications. ELM has many attractive features: it is solver-friendly, stable, and easily parallelizable. Despite these desirable features, ELM has not been without controversy: (1) some of its variants introduce excessive numerical diffusion that may degrade the accuracy of the method; (2) its accuracy is limited by the accuracy of the numerical integration; and (3) computational overhead (such as the back-tracking computation) is heavy—particularly on unstructured grids. We will address these pros and cons (and remedies) in some detail below. In particular, we demonstrate that the ELM method can be made more efficient and more important by an integrated application of modern numerical techniques:

1. user-friendly and optimal algebraic solvers for the discrete systems,
2. both temporal and spatial grid adaptation for improved stability and accuracy,
3. parallel implementation for reducing computational overheads, and
4. advanced techniques for accurate numerical integrations.

Numerical integration, stability, and artificial diffusions It has been observed that, if not treated carefully, some variants of ELM can cause excessive artificial diffusion, especially for finite difference discretization [42]. Actually, this issue is not as significant in finite element discretization. In fact, as shown in [34], ELM in the finite element setting is not more diffusive than other methods and, in fact, it has no or very little numerical diffusion when the solution is smooth if integration is evaluated exactly (or accurately enough). The numerical diffusion becomes more significant if the numerical integration is less accurate.

When using the finite element ELM to discretize material derivatives, a numerical solution at the departure feet is needed for numerical quadrature. Due to the nonalignment of the departure feet with the underlying mesh grid points, the function to be integrated—piecewise polynomial on each triangle or tetrahedron—is of low regularity. Hence, achieving an accurate numerical integration is a challenge in finite element ELM.

In [34], we have the following observations on ELM: (1) ELM with exact integration is unconditionally stable as a fully implicit scheme and it remains

stable for a relatively large time stepsize (see also [40, 38]); (2) ELM with a nodal interpolation is unconditionally stable, but it introduces excessive numerical diffusion and the convergence rate $O(k + h)$ is suboptimal for the linear finite element approximation; (3) ELM with Gaussian quadrature is conditionally stable and introduces less numerical diffusion. In some cases, the ELM may converge with the optimal rate $O(k + h^2)$, when k is chosen appropriately; (4) Spatial adaptivity can reduce numerical diffusion substantially and make the solving procedure more stable.

With these observations in mind, we have the following guidelines for using the Eulerian–Lagrangian method: (1) Numerical integration should be carried out as accurately as possible; (2) The nodal interpolation approach is preferred when diffusion is relatively small, and the Gauss quadrature is preferred when diffusion is relatively large; (3) Spatial grid adaptivity always helps to achieve stability and accuracy. For example, instability observed for the Gauss quadrature on uniform grids can be improved on properly adapted grids.

A general solver-friendly discretization What we find most attractive for the ELM is that it is a solver-friendly discretization: at each time step, the major work amounts to the solution of solver-friendly Stokes equations (see §3.3) and the nonlinearity in the original PDE is reduced to a set of independent nonlinear ordinary differential equations.

As a final note, one important feature that has not been much explored for the ELM method is that the method is highly parallelizable. Two major sources of overhead of the method are the calculation of the characteristic feet and numerical integrations. However these two computations can be carried out completely independently from element to element and can hence be easily realized with a parallel implementation.

5. Non-Newtonian Flows

The following system of equations is the commonly used Johnson–Segalman model for non-Newtonian fluids:

$$\operatorname{Re} \frac{Du}{Dt} = \nabla \cdot [\tau + \mu_s \mathcal{D}(u)] - \nabla p, \quad \nabla \cdot u = 0 \quad (5.28)$$

$$\tau + \operatorname{Wi} \left[\frac{D\tau}{Dt} - R(u)\tau - \tau R(u)^T \right] = 2\mu_p \mathcal{D}(u), \quad (5.29)$$

where $\frac{Du}{Dt}$ and $\frac{D\tau}{Dt}$ are material derivatives as defined in (4.26), and

$$\mathcal{D}(u) = \frac{1}{2}(\nabla u + \nabla u^T) \quad \text{and} \quad R(u) = \frac{a+1}{2}\nabla u + \frac{a-1}{2}\nabla u^T. \quad (5.30)$$

We notice that the above systems are reduced to the Oldroyd-B model, if $a = 1$, and to the Navier–Stokes equations when the Weissenberg number is $\operatorname{Wi} = 0$.

5.1. Reformulation of the constitutive equation. In addition to the difficulties that already exist in Navier–Stokes equations, the constitutive equation (5.29) presents a major challenge to properly discretizing the Johnson–Segalman model. Following Lee and Xu [38], this equation can be reformulated into a Riccati equation in terms of the following derivative along the particle trajectory defined for a symmetric tensor ξ :

$$\mathcal{L}_{u,R}\xi(t) = E(t,s) \frac{D}{Ds} \left(E(s,t)^T \xi(t,s) E(s,t) \right) E(t,s)^T \Big|_{s=t}, \quad (5.31)$$

where $E(t,s)$ satisfies

$$\frac{DE(t,s)}{Ds} = R(u)E(t,s), \quad E(t,t) = I. \quad (5.32)$$

Defining the conformation tensor $\sigma := \tau + \frac{\mu_p}{aWi} I$, we have (noticing that $\nabla \cdot u = 0$)

$$\mathcal{L}_{u,R}\sigma(t) = \partial_t \sigma + (u \cdot \nabla) \sigma - R\sigma - \sigma R^T.$$

It is easy to see that the constitutive equation (5.29) can be rewritten in terms of the conformation tensor as follows:

$$\mathcal{L}_{u,R}\sigma + \frac{1}{Wi} \sigma = \frac{\mu_p}{aWi^2} I. \quad (5.33)$$

We note that $\sigma(t)$ is symmetric positive-definite for any $t \geq 0$ physically. There are many indications [39] that preserving such a positivity on the discrete level is important. To solve (5.33), we can extend the positivity-preserving scheme for Riccati equations (ODE along the particle trajectory). We can further use piecewise constant or linear polynomials to discretize the spatial variable for σ to preserve such as positivity.

5.2. A solver-friendly fully discrete scheme. The material derivative Du/Dt and the derivative $\mathcal{L}_{u,R}\sigma$ are both derivatives along the particle trajectory. The ELM can discretize the material derivative in a straightforward way (see §4). Now we will discuss how to discretize $\mathcal{L}_{u,R}\sigma$. By definition (5.31), we can employ a first-order difference approximation for the time derivative to obtain

$$\mathcal{L}_{u,R}\sigma(s,t) \approx \frac{\sigma(t,t) - E(t-k,t)\sigma(t,t-k)E(t-k,t)^T}{k}. \quad (5.34)$$

Now we still need to approximate E . Let \tilde{E} be an approximate solution to (5.32) by the implicit Euler method:

$$\frac{1}{k} (\tilde{E}(s,t) - I) = R(t)\tilde{E}(s,t). \quad (5.35)$$

In the ELM, special integration schemes need to be carefully designed (see Feng and Shang [25]) to assure volume preservation of characteristic feet. For $d = 2$, the midpoint rule can be applied:

$$\frac{1}{k}(x - x_*^{n-1}) = u_h^n \left(\frac{x + x_*^{n-1}}{2} \right) \tag{5.36}$$

For stability concerns, we use special finite element discretizations for both velocity and pressure variables, such as the Scott and Vogelius element [45]. We can use the piecewise constant matrix space to approximate the conformation tensor σ .

A simple example of full discretization schemes reads like

$$\begin{aligned} & \frac{\text{Re}}{k} (u_h^n - \Pi_h^V u_{h*}^{n-1}) + \nabla_h p_h^n - \mu_s \Delta_h u_h^n = \nabla_h \cdot \sigma_h^n, \quad \nabla \cdot u_h^n = 0 \\ & \frac{1}{k} (\sigma_h^n - E_h^{n-1} \Pi_h^S(\sigma_{h*}^{n-1})(E_h^{n-1})^T) = -\alpha \sigma_h^n + \beta I, \quad \frac{1}{k} (E_h^{n-1} - I) = R_h(t^n) E_h^{n-1}. \end{aligned}$$

Here, Π_h^V is the L^2 -projection to the finite element space for approximating the velocity field; and Π_h^S is an entry-wise averaging operator, namely, $\Pi_h^S(\sigma_{ij})(x) = \frac{1}{|K|} \int_K \sigma_{ij}$, if $x \in K$ for any $K \in \mathcal{T}_h$.

As shown in [38], the above method satisfies the positivity-preserving property; that is, if $\Pi_h^S \geq 0$ and $\sigma_h^0 \geq 0$, then $\sigma_h^n \geq 0$, for $n \geq 0$. Secondly, this method satisfies the following discrete energy estimates,

$$\mathcal{E}_n \leq c_0 e^{-c_1 t^n} \mathcal{E}_0 + c_2 \frac{\mu p}{(aWi)^2} \quad \text{with} \quad \mathcal{E}_n := \text{Re} \|u_h^n\|_0^2 + \frac{1}{2a} \|\sigma_h^n\|_{L^1}, \tag{5.37}$$

where c_0, c_1 and c_2 are generic constants. We note that this type of scheme works for a whole range of models including the Oldroyd-B ($a = 1$), the FENE-PM, and the Phan-Thien and Tanner (PTT) models. For details, we refer to Lee and Xu [38] and Lee, Xu, and Zhang [39].

The above discretization scheme can be solved by, for example, a fixed-point iteration as follows. For a given n , we first compute the departure feet $x_*^{n-1} = x - k u_h^{n-1} \left(\frac{x + x_*^{n-1}}{2} \right)$ for all integration points x (which is easily solvable for an appropriately small k). We can then set $u_h^n := u_h^{(L+1)}$, $p_h^n := p_h^{(L+1)}$, and $\sigma_h^n := \sigma_h^{(L+1)}$ for some $L \geq 1$ from the following fixed-point iteration (for $\ell = 0, 1, 2, \dots$), with $u_h^{(0)} := u_h^{n-1}$, $p_h^{(0)} := p_h^{n-1}$, and $\sigma_h^{(0)} := \sigma_h^{n-1}$:

- (1) solve the Stokes-type system

$$\begin{cases} \text{Re} u_h^{(\ell+1)} - k \Delta_h u_h^{(\ell+1)} + k \nabla_h p_h^{(\ell+1)} = \text{Re} \Pi_h^V (u_h^{n-1}(x_*^{n-1})) + k \nabla_h \cdot \sigma_h^{(\ell)}, \\ \nabla \cdot u_h^{(\ell+1)} = 0. \end{cases}$$

(2) update the conformation tensor

$$\begin{aligned} E_h^{(\ell+1)} &= (I - k\Pi_h^S R(u_h^{(\ell+1)}))^{-1}, \\ (1 + k\alpha)\sigma_h^{(\ell+1)} &= E_h^{(\ell+1)}\Pi_h^S(\sigma_h^{n-1}(x_*^{n-1}))(E_h^{(\ell+1)})^T + k\beta I. \end{aligned}$$

Thanks to the energy estimate (5.37) and the optimal algorithm for solving the generalized Stokes equation in §3.3, we obtain the following result.

Theorem 5.1 ([39]). *If the time step size k is small enough, the fixed-point iteration above converges uniformly with respect to Re , Wi , and h . The computational complexity of the algorithm is of $\mathcal{O}(N \log N)$ for each iteration, where N is the total number of spatial degrees of freedom.*

6. Magnetohydrodynamics

MHD equations have been much studied in the literature. We demonstrate here how a solver-friendly scheme can discretize this kind of equations. We consider the following simple model magnetohydrodynamic (MHD) equation (for incompressible media) that consists of the Navier–Stokes equations, Maxwell’s equations and Ohm’s law:

$$u_t + (u \cdot \nabla)u + \nabla p = \mu\Delta u + \mu_0 J \times B, \quad \nabla \cdot u = 0, \tag{6.38}$$

$$B_t + \nabla \times E = 0, \quad \nabla \times B = \mu_0 J, \quad \nabla \cdot B = 0, \quad E + u \times B = \frac{1}{\sigma_0} J. \tag{6.39}$$

Here u is the fluid velocity, B is the magnetic field, $\eta = 1/(\sigma_0\mu_0)$ is the magnetic diffusivity, μ is the viscosity, μ_0 is the magnetic permeability constant, and σ_0 is the constant electrical conductivity of the fluid.

We view the magnetic field B as a 2-form and consider its Lie derivative as follows (noticing that $\nabla \cdot u = 0$):

$$\mathcal{L}_u B(x, t) = \left. \frac{d}{ds} \phi_{s,t}^* B \right|_{s=t} = \partial_t B + \nabla \times (B \times u) + (\nabla \cdot B)u = \partial_t B - \nabla \times (u \times B).$$

Here $\phi_{s,t}$ is the pull-back operator in terms of the flow-map $\phi_{s,t}$ (4.25) given by

$$\phi_{s,t}^* B = (\det D\phi_{s,t}(x))(D\phi_{s,t}(x))^{-1} B(\phi_{s,t}(x), s).$$

In terms of the material derivative (4.26) and the above Lie derivative, the MHD system can be rewritten as

$$\frac{Du}{Dt} + \nabla p = \mu\Delta u + (\nabla \times B) \times B, \quad \nabla \cdot u = 0 \tag{6.40}$$

$$\mathcal{L}_u B + \eta \nabla \times (\nabla \times B) = 0, \quad \nabla \cdot B = 0. \tag{6.41}$$

We notice that the condition $\nabla \cdot B = 0$ is actually contained in the first equation of (6.41) as long as $\nabla \cdot B = 0$ at time $t = 0$ because of the commutation of the exterior derivative and Lie derivative.

Let $F(s, t) = \nabla \phi_{s,t}$. Since $\nabla \cdot u = 0$, we have $\det(\nabla \phi_{s,t}(x)) = 1$. We now discretize both derivatives for u and B along the particle trajectory via the Eulerian–Lagrangian framework. The Lie advection for B can, for example, be discretized by a simple Euler method as follows:

$$\mathcal{L}_u B \approx \frac{B(x, t) - [F(t - k, t)]^{-1} B(\phi_{t-k,t}(x), t - k)}{k}, \tag{6.42}$$

where $\phi_{t-k,t}(x)$ can be computed by solving the ODE (4.25) and $F(s, t)$ can be computed by solving

$$\frac{DF^{-1}(s, t)}{Ds} = -F^{-1}(s, t) \nabla u(\phi_{s,t}(x), s) \quad (s < t), \quad F(t, t) = I. \tag{6.43}$$

We use the simple discretization of (6.42) to illustrate our main idea.

Using these ELM discretizations, we obtain the following implicit semi-discrete system:

$$\begin{aligned} \frac{u^n - u_*^{n-1}}{k} - \mu \Delta u^n + \nabla p^n - (\nabla \times B^n) \times B^n &= 0, \quad \nabla \cdot u^n = 0, \\ \frac{B^n - [F(t^{n-1}, t^n)]^{-1} B_*^{n-1}}{k} + \eta \nabla \times \nabla \times B^n &= 0, \end{aligned}$$

where k is the time-step size. By moving the known quantities related to time t^n to the right-hand sides, we obtain this system of equations:

$$\begin{cases} (k^{-1}I - \mu \Delta)u^n + \nabla p^n - (\nabla \times B^n) \times B^n &= f^{n-1}, \\ \nabla \cdot u^n &= 0, \\ (\eta \nabla \times \nabla \times + k^{-1}I)B^n &= g^{n-1}. \end{cases} \tag{6.44}$$

We can use finite element methods for the above system by: (1) discretizing (u, p) variables by standard finite elements for the Stokes equations; and (2) discretizing the B variable by standard edge elements for $H(\text{curl})$ systems.

The most noticeable feature of this discretization procedure is that the resulting discrete systems are solver-friendly. Roughly speaking, the third equation in (6.44) can be solved by applying the HX-preconditioner (3.11), and the first two equations are just the Stokes equations that can be solved by the method described in §3.3.

Another important advantage of our discretization scheme is that it is robust when the resistivity constant η becomes very small. This is analogous to the convection-dominated situation in the convection-diffusion equations, and our ELM scheme is related to the traditional upwinding scheme.

7. Concluding Remarks

In this paper, we have presented a systematic approach to designing mathematically optimal and practically user-friendly numerical methods for a large class of linear and nonlinear partial differential equations. Thus far, we have demonstrated that the partial differential equations listed in Table 1 can be solved by the techniques presented in this paper.

Table 1. Extended list of solver-friendly partial differential equations. Here #P is the number of Poisson solvers needed and §# is the number of the section where the relevant solvers are discussed.

Problems	Partial Differential Equations	#P	§#
Poisson	$-\Delta u = f$	1	§2.2
Reaction-Diffusion	$-\nabla \cdot (\mu(x)\nabla u) + c(x)u = f$	1	§2.2
H(curl)	$\nabla \times \nabla \times u + u = f$	4	§3.1
H(div)	$-\nabla \nabla \cdot u + u = f$	3–6	§3.1
Maxwell	$B_t + \nabla \times E = 0, \nabla \cdot D = \rho,$ $D_t - \nabla \times H = -J, \nabla \cdot B = 0$	4	§3.1
Darcy	$-\nabla \cdot u = f, u = \mu \nabla p$	3–6	§3.2
Stokes	$-\Delta u + \nabla p = f, \nabla \cdot u = 0$	3	§3.3
Brinkman	$-\nabla \cdot (\mu(x)\nabla u) + c(x)u + \nabla p = f, \nabla \cdot u = g$	3	§3.4
Linear Elasticity	$-\Delta u - \frac{\nu}{1-2\nu} \nabla \nabla \cdot u = f$	3	§3.4
Biharmonic (plate)	$\Delta^2 u = f$	2	§3.5
Reissner-Mindlin	$-\operatorname{div}(\mathcal{C}D(\phi)) = \lambda t^{-2}(\phi - \nabla \omega) = 0,$ $\lambda t^{-2}(-\Delta \omega + \operatorname{div} \phi) = g$	4	§3.5
R–C–D	$u_t + v(x) \cdot \nabla u - \nabla \cdot (\mu(x)\nabla u) + c(x)u = f$	1	§4
Navier–Stokes	$u_t + (u \cdot \nabla)u - \mu \Delta u + \nabla p = f, \nabla \cdot u = 0$	4	§4
Johnson–Seglman	$u_t + (u \cdot \nabla)u - \mu_s \Delta u + \nabla p = \nabla \cdot \tau, \nabla \cdot u = 0,$ $\tau + \operatorname{Wi}[u_t + (u \cdot \nabla)u - R\tau - \tau R^T] = \frac{1}{2} \mu_p \mathcal{D}u$ $R = \frac{a+1}{2} \nabla u + \frac{a-1}{2} \nabla u^T$	3	§5
MHD	$u_t + (u \cdot \nabla)u + \nabla p = \mu \Delta u + \mu_0 J \times B, \nabla \cdot u = 0,$ $B_t + \nabla \times E = 0, \nabla \times B = \mu_0 J,$ $\nabla \cdot B = 0, E + u \times B = \sigma_0^{-1} J$	7	§6

Let us give a brief summary on relevant algorithms for the equations listed in Table 1. For the Poisson, reaction-diffusion equations, and sometimes linear elasticity equations, AMG (sometimes enhanced with analytic and geometric information) are the methods of choice. H(curl) and H(div) systems can be solved by Hiptmair–Xu preconditioners. The Reissner–Mindlin plate model can be preconditioned by the Arnold–Falk–Winther preconditioner. The mixed finite element systems for Darcy–Stokes–Brinkman models and sometimes linear elasticity equations can be solved in most cases by preconditioned Min-Res using a block diagonal preconditioner consisting of Poisson-like solvers

and sometimes by augmented Lagrangian methods or hybridization techniques. The time-dependent reaction-convection-diffusion, Navier–Stokes, Johnson–Segalman and MHD equations, if discretized by a Eulerian–Lagrangian method, can be reduced to a handful of aforementioned equations. And, if discretized by other methods, they can be preconditioned by a FASP method by combining a Eulerian–Lagrangian discretization with an appropriate smoother.

Ongoing and future works While it is our view that solver-friendly discretizations should be used whenever possible, there could be situations where solver-friendly discretizations may not be desirable or available. In this event, we advocate the use of a possible solver-friendly discretization as an auxiliary discretization in order to design an efficient solver for the original discrete systems. We are now developing a general framework, to be known as the *auxiliary discretization method*. For example, as demonstrated in [35], a monotone scheme can be used to construct an efficient iterative method for a standard or streamline diffusion finite element discretization for convection-diffusion equations. The auxiliary space method presented in Xu [57] and Hiptmair and Xu [32] and the two-grid method [58] are also examples of auxiliary discretization methods.

We plan to expand these special auxiliary discretization methods into a general algorithmic design and theoretical analysis framework. In particular, we will explore the use of ELM as an auxiliary discretization. Similar to the auxiliary space method and the two-grid method, ELM can be used as a major component of an iterative method or as a preconditioner for other given discretization methods. Furthermore, for a given discretization of nonlinear steady-state problems or evolution problems (at each time step), we will also explore ELM’s potential for obtaining an initial guess for a linearization scheme such as the Newton’s method.

The Poisson-based Solver project It is the intention of the author to continue this line of work and to extend as much as possible the list of PDEs in the Table 1. We will call this “*The Poisson-based Solver Project*”. More details on this project (including its relevant numerical packages and references) and discussion pertaining to its future development can be found on the website www.multigrid.org.

References

- [1] D. N. Arnold and F. Brezzi. Mixed and nonconforming finite element methods: Implementation, postprocessing and error estimates. *RAIRO Model Math. Anal. Numer.*, 19:7–32, 1985.
- [2] D. N. Arnold, R. S. Falk, and R. Winther. Preconditioning discrete approximations of the Reissner–Mindlin plate model. *RAIRO Modél. Math. Anal. Numér.*, 31(4):517–557, 1997.

-
- [3] D. N. Arnold, R. S. Falk, and R. Winther. Preconditioning in $H(\text{div})$ and applications. *Mathematics of Computation*, 66:957–984, 1997.
- [4] D. N. Arnold, R. S. Falk, and R. Winther. Multigrid in $H(\text{div})$ and $H(\text{curl})$. *Numerische Mathematik*, 85:197–218, 2000.
- [5] T. M. Austin, T. A. Manteuffel, and S. McCormick. A robust multilevel approach for minimizing $\mathbf{H}(\text{div})$ -dominated functionals in an \mathbf{H}^1 -conforming finite element space. *Numer. Linear Algebra Appl.*, 11(2–3):115–140, 2004.
- [6] R. E. Bank and R. K. Smith. An algebraic multilevel multigraph algorithm. *SIAM Journal on Scientific Computing*, 23(5):1572–1592, 2002.
- [7] R. Beck. Algebraic multigrid by component splitting for edge elements on simplicial triangulations. Technical report, Tech. Report SC 99–40, ZIB, Berlin, Germany, 1999.
- [8] M. Bercovier and O. Pironneau. Error estimates for finite element method solution of the stokes problem in the primitive variables. *Numerische Mathematik*, 33:211–224, 1979.
- [9] P. B. Bochev and M. D. Gunzburger. *Least-squares finite element methods*, volume 166 of *Applied Mathematical Sciences*. Springer, New York, 2009.
- [10] P. B. Bochev, J. J. Hu, C. M. Siefert, and R. S. Tuminaro. An algebraic multigrid approach based on a compatible gauge reformulation of Maxwell’s equations. *SIAM J. Sci. Comput.*, 31(1):557–583, 2008.
- [11] D. Braess and P. Peisker. A conjugate gradient method and a multigrid algorithm for Morley’s finite element approximation of the biharmonic equations. *Numerische Mathematik*, 50:567–586, 1987.
- [12] J. Bramble and J. Pasciak. Iterative techniques for time dependent Stokes problems. *Computer Methods in Applied Mechanics and Engineering*, 1–2:13–30, 1997.
- [13] J. H. Bramble. *Multigrid Methods*, volume 294 of *Pitman Research Notes in Mathematical Sciences*. Longman Scientific & Technical, Essex, England, 1993.
- [14] J. H. Bramble, J. E. Pasciak, J. Wang, and J. Xu. Convergence estimates for multigrid algorithms without regularity assumptions. *Mathematics of Computation*, 57:23–45, 1991.
- [15] A. Brandt. Multi-Level adaptive solutions to Boundary-Value problems. *Mathematics of Computation*, 31:333–390, 1977.
- [16] J. Brannick and L. Zikatanov. Algebraic multigrid methods based on compatible relaxation and energy minimization. In *Domain decomposition methods in science and engineering XVI*, volume 55, pages 15–26. Springer, Berlin, 2007.
- [17] F. Brezzi and M. Fortin. *Mixed and hybrid finite element methods*. Springer-Verlag, 1991.
- [18] L. Chen, M. Holst, X. Wu, J. Xu, and Y. Zhu. Auxiliary space preconditioners for hybridized mixed finite element methods. in preparation.
- [19] L. Chen, R. H. Nochetto, and J. Xu. Local multilevel methods on graded bisection grids: $H(\text{curl})$ and $H(\text{div})$ systems. *Preprint*, 2008.
- [20] A. J. Chorin. On the convergence of discrete approximations to the Navier-Stokes equations. *Math. Comp.*, 23:341–353, 1969.

- [21] J. Douglas and T. F. Russell. Numerical methods for convection-dominated diffusion problems based on combining the method of characteristics with finite element or finite difference procedures. *SIAM J. Numer. Anal.*, 19(5):871–885, 1982.
- [22] A. M. et al. Panel report on recent significant advances in computational science, http://sc.doe.gov/ascr/ProgramDocuments/Docs/Breakthroughs_2008.pdf. Technical report, Department of Energy (DOE), 2008.
- [23] R. E. Ewing and J. Wang. Analysis of multilevel decomposition iterative methods for mixed finite element methods. *M2AN*, 28(4):377–398, 1994.
- [24] R. D. Falgout, J. E. Jones, and U. M. Yang. The design and implementation of hypre, a library of parallel high performance preconditioners. In *Numerical solution of partial differential equations on parallel computers*, volume 51 of *Lect. Notes Comput. Sci. Eng.*, pages 267–294. Springer, Berlin, 2006.
- [25] K. Feng and Z.-J. Shang. Volume-preserving algorithms for source-free dynamical systems. *Numer. Math.*, 71(4):451–463, 1995.
- [26] R. Glowinski and P. Le Tallec. *Augmented Lagrangians and operator splitting methods in nonlinear mechanics*. SIAM, Philadelphia 1989.
- [27] L. Grasedyck and J. Xu. Algebraic multigrid methods based on auxiliary grids. preprint.
- [28] W. Hackbusch. *Multigrid Methods and Applications*, volume 4 of *Computational Mathematics*. Springer-Verlag, Berlin, 1985.
- [29] R. Hiptmair. Multigrid method for $H(\text{div})$ in three dimensions. *Electronic Transactions on Numerical Analysis*, 6:133–152, 1997.
- [30] R. Hiptmair. Multigrid method for Maxwell’s equations. *SIAM Journal on Numerical Analysis*, 36(1):204–225, 1999.
- [31] R. Hiptmair, T. Schiekofer, and B. Wohlmuth. Multilevel preconditioned augmented Lagrangian techniques for 2nd order mixed problems. *Computing*, 57(1):25–48, 1996.
- [32] R. Hiptmair and J. Xu. Nodal auxiliary space preconditioning in $H(\text{curl})$ and $H(\text{div})$ spaces. *SIAM J. Numer. Anal.*, 45(6):2483–2509 (electronic), 2007.
- [33] X. Hu, X. Wu, J. Xu, C.-S. Zhang, and S. Zhang. Algebraic multigrid methods for hybridized mixed finite element systems arising from reservoir simulations. In preparation.
- [34] J. Jia, X. H. Hu, J. Xu, C. Zhang, and T. F. Russel. Numerical study on the Eulerian-Lagrangian method: Effects of integrations and spatial adaptivity. *J. Comp. Math.*, 2010.
- [35] H. Kim, J. Xu, and L. Zikatanov. Uniformly convergent multigrid methods for convection diffusion problems without any constraint on coarse grids. *Advances in Comp. Math.*, 20(4):385–399, 2004.
- [36] T. V. Kolev and P. S. Vassilevski. Parallel auxiliary space AMG for $H(\text{curl})$ problems. *J. Comput. Math.*, 27(5):604–623, 2009.
- [37] Y. Lee, J. Wu, J. Xu, and L. Zikatanov. Robust subspace correction methods for nearly singular systems. *Mathematical Models and Methods in Applied Sciences*, 17(11):1937–1963, 2007.

- [38] Y.-J. Lee and J. Xu. New formulations, positivity preserving discretizations and stability analysis for non-Newtonian flow models. *Comput. Methods Appl. Mech. Engrg.*, 195(9–12):1180–1206, 2006.
- [39] Y.-J. Lee, J. Xu, and C.-S. Zhang. Stable finite element discretizations for non-newtonian flow models: Applications to viscoelasticity. *Handbook of Numerical Analysis (HNA) Vol. 16*, Numerical Methods for Non-Newtonian Fluids Vol. 16 ISBN: 978-0-444-53047-9, 2009.
- [40] O. Pironneau. On the transport-diffusion algorithm and its applications to the Navier-Stokes equations. *Numer. Math.*, 38(3):309–332, 1981/82.
- [41] S. Reitzinger and J. Schöberl. An algebraic multigrid method for finite element discretizations with edge elements. *Numer. Linear Algebra Appl.*, 9(3):223–238, 2002.
- [42] T. Russell. Numerical dispersion in Eulerian-Lagrangian methods. *Computational Methods in Water Resources*, 2, 2002.
- [43] T. Rusten and R. Winther. A preconditioned iterative method for saddlepoint problems. *SIAM Journal on Matrix Anal. Appl.*, 1992.
- [44] J. Schöberl. Multigrid methods for a parameter dependent problem in primal variables. *Numer. Math.*, 84(1):97–119, 1999.
- [45] L. R. Scott and M. Vogelius. Conforming finite element methods for incompressible and nearly incompressible continua. In *Large-scale computations in fluid mechanics, Part 2 (La Jolla, Calif., 1983)*, volume 22 of *Lectures in Appl. Math.*, pages 221–244. Amer. Math. Soc., Providence, RI, 1985.
- [46] S. Shu, D. Sun, and J. Xu. An algebraic multigrid method for higher order finite element discretizations. *Computing*, 77(4):347–377, 2006.
- [47] K. Stüben. A review of algebraic multigrid. *J. Comput. Appl. Math.*, 128(1–2):281–309, 2001. Numerical analysis 2000, Vol. VII, Partial differential equations.
- [48] R. Temam. Sur l’approximation de la solution des équations de Navier-Stokes par la méthode des pas fractionnaires (I). *Archive for Rational Mechanics and Analysis*, 32(2):135–153, 1969.
- [49] U. Trottenberg, C. Oosterlee, and A. Schüller. *Multigrid*. Academic Press Inc., San Diego, CA, 2001. With contributions by A. Brandt, P. Oswald and K. Stüben.
- [50] R. Tuminaro, J. Xu, and Y. Zhu. Auxiliary space preconditioners for mixed finite element methods. In *Domain Decomposition Methods in Science and Engineering*, volume XVIII, pages 99–109. Springer, 2009.
- [51] P. Vaněk, M. Brezina, and J. Mandel. Convergence of algebraic multigrid based on smoothed aggregation. *Numerische Mathematik*, 88(3):559–579, 2001.
- [52] W. L. Wan, T. F. Chan, and B. Smith. An energy-minimizing interpolation for robust multigrid methods. *SIAM Journal on Scientific Computing*, 21(4):1632–1649 (electronic), 2000.
- [53] H. Wu and Z. Chen. Uniform convergence of multigrid V-cycle on adaptively refined finite element meshes for second order elliptic problems. *Sci. China Ser. A*, 49(10):1405–1429, 2006.

- [54] X. Xie, J. Xu, and G. Xue. Uniformly-stable finite element methods for Darcy-Stokes-Brinkman models. *J. Comput. Math.*, 26(3):437–455, 2008.
- [55] J. Xu. Iterative methods by space decomposition and subspace correction. *SIAM Review*, 34(4):581–613, 1992.
- [56] J. Xu. Some two-grid finite element methods. In A. Quarteroni, J. Périaux, Y. A. Kuznetsov, and O. B. Widlund, editors, *Domain Decomposition Methods in Science and Engineering: The Sixth International Conference on Domain Decomposition*, volume 157 of *Contemporary Mathematics*, pages 79–87, Providence, Rhode Island, 1994. American Mathematical Society.
- [57] J. Xu. The auxiliary space method and optimal multigrid preconditioning techniques for unstructured grids. *Computing*, 56(3):215–235, 1996.
- [58] J. Xu. Two-grid discretization techniques for linear and nonlinear PDEs. *SIAM Journal on Numerical Analysis*, 33(5):1759–1777, 1996.
- [59] J. Xu and S. Zhang. Poisson-based solvers for biharmonic problems. In preparation.
- [60] J. Xu and L. Zikatanov. The method of alternating projections and the method of subspace corrections in Hilbert space. *Journal of the American Mathematical Society*, 15:573–597, 2002.
- [61] J. Xu and L. Zikatanov. On an energy minimizing basis for algebraic multigrid methods. *Computing and Visualization in Science*, 7(3–4):121–127, 2004.
- [62] H. Yserentant. Old and new convergence proofs for multigrid methods. *Acta Numerica*, pages 285–326, 1993.
- [63] S. Zhang and Z. Zhang. Invalidity of decoupling a biharmonic equation to two Poisson equations on non-convex polygons. *Int. J. Numer. Anal. Model.*, 5(1):73–76, 2008.

Section 17

**Control Theory and
Optimization**

This page is intentionally left blank

Optimal Control under State Constraints

Hélène Frankowska*

Abstract

Optimal control under state constraints has brought new mathematical challenges that have led to new techniques and new theories. We survey some recent results related to issues of regularity of optimal trajectories, optimal controls and the value function, and discuss optimal synthesis and necessary optimality conditions. We also show how abstract inverse mapping theorems of set-valued analysis can be applied to study state constrained control systems.

Mathematics Subject Classification (2010). 49K15, 34A60, 47J07, 49N35, 49N60.

Keywords. Optimal control, state constraints, value function, optimal synthesis, normal maximum principle, smoothness of optimal trajectories, regularity of the adjoint variable.

1. Introduction

We consider here finite dimensional control systems by which we mean ordinary differential equations of the form

$$\begin{cases} x'(t) &= f(t, x(t), u(t)), \quad u(t) \in U \quad \text{a.e. in } [0, 1], \\ x(0) &= x_0, \end{cases} \quad (1)$$

where U is a complete separable metric space, $f : [0, 1] \times \mathbf{R}^n \times U \rightarrow \mathbf{R}^n$, $x_0 \in \mathbf{R}^n$, $x(t)$ is called the state of the system, t denotes the time, $x'(\cdot)$ is the derivative of $x(\cdot)$ with respect to time, and the function $u(\cdot)$ has to be chosen so that the corresponding solution $x(\cdot)$ has some desirable properties; in other words, $u(\cdot)$ “controls” the solution $x(\cdot)$ of (1). As a set of controls we choose the set of all Lebesgue measurable functions $u(\cdot) : [0, 1] \rightarrow U$, while a solution

*Combinatoire & Optimisation, Université Pierre et Marie Curie, case 189, 4 place Jussieu, 75252 Paris Cedex 05, France. E-mail: frankowska@math.jussieu.fr.

$x(\cdot)$ (called here a trajectory of control system) is an absolutely continuous function satisfying $x(0) = x_0$ and $x'(t) = f(t, x(t), u(t))$ almost everywhere in $[0, 1]$ (in the sense of Lebesgue measure) for some control $u(\cdot)$. Once we choose a control $u(\cdot)$, system (1) becomes an ordinary differential equation for which conditions for existence, uniqueness and properties of trajectories are classically known. Measurable controls have proven to be well adapted for investigation of existence of solutions to optimal control problems in the general case.

The importance of control systems in mathematics and science is nowadays widely acknowledged. Indeed, not only do they respond to basic issues arising in engineering and social sciences, but they serve as subsumption of previous theories - e.g., the classical calculus of variations - and stimulus to progress in related mathematical fields as well. For example, some fundamental research directions such as weak solutions of nonlinear first (and second) order partial differential equations, set-valued, variational and nonsmooth analysis, have found their inspiration and motivation in control theory and differential games.

The analysis of properties of trajectories of (1) becomes much more challenging to study if the states $x(t)$ are required to belong to a certain region; then we say that the control system (1) is subject to *state constraints*. To be specific, let K be a given closed subset of \mathbf{R}^n and consider state constraint of the form

$$x(t) \in K \quad \text{for all } t \in [0, 1]. \quad (2)$$

A trajectory $x(\cdot)$ of (1) satisfying the state constraint (2) is called a *viable* (or *feasible*) trajectory of the control system. Properties of viable trajectories could be quite different from those of system (1) only.

From now on, we denote by \mathcal{C} the space of continuous functions from $[0, 1]$ to \mathbf{R}^n with the supremum norm $\|\cdot\|_{\mathcal{C}}$ and by $W^{1,1}([0, 1]; \mathbf{R}^n)$ the space of absolutely continuous functions from $[0, 1]$ to \mathbf{R}^n with the norm $\|w\|_{W^{1,1}} = \|w\|_{L^1} + \|w'\|_{L^1}$. Let $\mathcal{B}_{\mathcal{C}}$ and $\mathcal{B}_{W^{1,1}}$ denote the closed unit balls in the corresponding spaces.

Consider the set $\mathcal{S}(x_0)$ of all trajectories of (1) and let $\mathcal{S}_K(x_0)$ denote the set of all trajectories of (1), (2). If f is Lipschitz with respect to x with a constant independent of t and u , then the set-valued map $x_0 \rightsquigarrow \mathcal{S}(x_0)$ is Lipschitz continuous in the sense that for some $L \geq 0$, $\mathcal{S}(x_0) \subset \mathcal{S}(y_0) + L|x_0 - y_0|\mathcal{B}_{W^{1,1}}$ for all $x_0, y_0 \in \mathbf{R}^n$. This is no longer the case for the set-valued map $\mathcal{S}_K(\cdot)$ even for simple sets K and even with $\mathcal{B}_{W^{1,1}}$ replaced by $\mathcal{B}_{\mathcal{C}}$.

Example. $K = \{x = (x_1, x_2) \in \mathbf{R}^2 \mid x_2 \leq x_1^2\}$, $U = [-1, 1] \times \{0\}$. Consider the following control system under a state constraint

$$x'(t) = u(t), \quad u(t) \in U \quad \text{a.e. in } [0, 1], \quad x(t) \in K \quad \forall t \in [0, 1].$$

Then $x(t) := (1 - t, 0)$ is a viable trajectory with the initial state $x(0) = (1, 0)$. Pick any $x_2 \in (0, 1]$ and $y(\cdot) \in \mathcal{S}_K((1, x_2))$. Denoting by $|\cdot|$ the Euclidean norm in \mathbf{R}^2 , we get $|x(1) - y(1)| \geq \sqrt{x_2}$ and therefore $\mathcal{S}_K(\cdot) : K \rightsquigarrow \mathcal{C}$ is not Lipschitz on any neighborhood of $(1, 0)$. \square

• **Existence of viable trajectories.** In general, for some initial conditions, trajectories of (1) satisfying the state constraint (2) may not exist; for instance when for every $u \in U$, $f(0, x_0, u)$ points strictly outside of K and f is continuous. The tangent vectors to the set K turn out to be very instrumental for investigating existence of trajectories of a control system under state constraint.

Denote the distance from $y \in \mathbf{R}^n$ to K by $\text{dist}(y; K)$. The contingent cone and the Clarke tangent cone to K at $x \in K$ are defined respectively by

$$\begin{aligned} T_K(x) &:= \{v \in \mathbf{R}^n \mid \liminf_{h \rightarrow 0^+} \frac{1}{h} \text{dist}(x + hv; K) = 0\}; \\ C_K(x) &:= \{v \in \mathbf{R}^n \mid \lim_{h \rightarrow 0^+, K \ni y \rightarrow x} \frac{1}{h} \text{dist}(y + hv; K) = 0\}; \end{aligned}$$

and the normal cone to K at x by $N_K(x) := \{p \in \mathbf{R}^n \mid \langle p, v \rangle \leq 0 \ \forall v \in C_K(x)\}$.

Set $f(t, x, U) := \bigcup_{u \in U} \{f(t, x, u)\}$ and $W(t, x) := \{u \in U \mid f(t, x, u) \in T_K(x)\}$. Observe that $x(\cdot) \in \mathcal{S}_K(x_0)$ if and only if $x(0) = x_0$ and for almost every $t \in [0, 1]$

$$x'(t) \in f(t, x(t), W(t, x(t))) = f(t, x(t), U) \cap T_K(x(t)).$$

In other words $x(\cdot)$ is a trajectory of the following control system

$$x'(t) = f(t, x(t), u(t)), \quad u(t) \in W(t, x(t)) \quad \text{a.e. in } [0, 1], \quad x(0) = x_0. \quad (3)$$

The main difficulty in investigating the above system consists in high irregularity of the set-valued map $(t, x) \rightsquigarrow W(t, x)$. For this reason usually one constructs trajectories of (1) satisfying state constraint (2) instead of solving directly (3).

Existence of viable trajectories can be studied using viability theory, which was developed for systems described by differential inclusions. Control systems are a particular case of differential inclusions for set-valued maps $(t, x) \rightsquigarrow f(t, x, U)$.

Let $F : \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$ be a Marchaud map, i.e. an upper semicontinuous set-valued map with nonempty convex compact values and linear growth and let $x_0 \in K$. Consider the differential inclusion

$$\begin{cases} x'(t) \in F(x(t)) & \text{for a.e. } t \geq 0 \\ x(t) \in K & \text{for all } t \geq 0 \\ x(0) = x_0. \end{cases} \quad (4)$$

A locally absolutely continuous function $x(\cdot) : [0, \infty) \rightarrow \mathbf{R}^n$ satisfying the above relations is called a viable (in K) trajectory. A necessary and sufficient condition for the existence of a trajectory of (4) for every $x_0 \in K$ is the viability condition

$$F(x) \cap T_K(x) \neq \emptyset \quad \forall x \in K. \quad (5)$$

(see for instance [3, Theorems 3.3.2 and 3.3.5]). Similar conditions allowing to handle time dependent constraints $K(t)$ with F depending also measurably

on time as well as infinite dimensional control systems and stochastic systems can be found for instance in [3]. Viability theory has numerous applications to control, such as for investigation of uniqueness of solutions to Hamilton-Jacobi-Bellman equations, of invariance of stochastic control systems, of optimal synthesis, etc.

• **Inward pointing condition.** In the example above the mapping $f(x, u) = u$ is Lipschitz, $f(x, U) = [-1, 1] \times \{0\}$ is compact and convex and viability condition (5) is satisfied. Thus (5) is not sufficient for the local Lipschitz continuity of $\mathcal{S}_K(\cdot) : K \rightsquigarrow \mathcal{C}$ on K (and on the interior $\text{Int } K$ of K) even when f is Lipschitz. For guaranteeing such property a stronger tangential condition is needed

$$F(x) \cap \text{Int } C_K(x) \neq \emptyset \quad \forall x \in \partial K. \quad (6)$$

For K having a smooth boundary, the control system (1) with the time independent f and $F(x) := f(x, U)$, (6) is equivalent to the so called *inward pointing condition*

$$\forall x \in \partial K, \exists u_x \in U \text{ such that } \langle n_x, f(x, u_x) \rangle < 0 \quad (7)$$

for the outward unit normal n_x to K at x . Condition (7) was introduced in [79] to investigate continuity of the value function of an infinite horizon problem and then to study uniqueness of viscosity solutions to a Hamilton-Jacobi PDE.

• **Inward pointing condition and linearization of control systems.** When K is an intersection of closed sets with smooth boundaries, a generalization of (6) to the time dependent f implies that, under some mild assumptions on f and a transversality assumption on K , for any $n \times n$ matrices $A(t) = (a_{ij}(t))$, $i, j = 1, \dots, n$ with $a_{ij}(\cdot) \in L^1(0, 1)$ and every trajectory/control pair $(\bar{x}(\cdot), \bar{u}(\cdot))$ of (1), (2), there exists a solution $w(\cdot)$ to the following linear control system

$$\begin{cases} w'(t) = A(t)w(t) + v(t), & v(t) \in T_{co f(t, \bar{x}(t), U)}(f(t, \bar{x}(t), \bar{u}(t))) \text{ a.e.} \\ w(0) = 0 \\ w(t) \in \text{Int } T_K(\bar{x}(t)) \text{ for all } t \in (0, 1], \end{cases} \quad (8)$$

where co states for the convex hull (see [11], and [53] for $w(0) = w_0 \in \text{Int } T_K(\bar{x}(0))$). When $A(t) = \frac{\partial f}{\partial x}(t, \bar{x}(t), \bar{u}(t))$, the control system in (8) is a linearization of (1) along the trajectory/control pair $(\bar{x}(\cdot), \bar{u}(\cdot))$, while the relation $w(t) \in \text{Int } T_K(\bar{x}(t))$ can be considered as linearization of the state constraint along $\bar{x}(\cdot)$.

Existence of a solution to (8) is important in various applications. For instance it yields normality of necessary optimality conditions for some optimal control problems. Observe that it resembles a constraint qualification condition in mathematical programming, which guarantees existence of Lagrange multipliers in normal form. In Section 2 existence of a solution to (8) is used to investigate local Lipschitz continuity of $\mathcal{S}_K(\cdot) : \text{Int } K \rightsquigarrow \mathcal{C}$ and in Section 4 it is applied to derive normal first order necessary optimality conditions.

• **Value function of the Mayer optimal control problem.** Let $\varphi : \mathbf{R}^n \rightarrow \mathbf{R} \cup \{+\infty\}$ be an extended-real-valued function, bounded from below.

Consider the Mayer problem

$$\text{minimize } \{\varphi(x(1)) \mid x(\cdot) \in \mathcal{S}_K(x_0)\}. \quad (9)$$

A trajectory $\bar{x}(\cdot) \in \mathcal{S}_K(x_0)$ is called optimal if $\varphi(\bar{x}(1)) = \min_{x(\cdot) \in \mathcal{S}_K(x_0)} \varphi(x(1)) < +\infty$. Let $t_0 \in [0, 1]$, $y_0 \in K$ and consider the control system

$$x'(t) = f(t, x(t), u(t)), \quad u(t) \in U \text{ a.e. in } [t_0, 1], \quad x(t_0) = y_0. \quad (10)$$

The value function $V : [0, 1] \times \mathbf{R}^n \rightarrow \mathbf{R} \cup \{+\infty\}$ associated to (9) is defined by

$$V(t_0, y_0) = \inf\{\varphi(x(1)) \mid x(\cdot) \text{ is a trajectory of (10), } x([t_0, 1]) \subset K\}, \quad (11)$$

(we adopt the convention that the infimum *inf* over an empty set is equal to $+\infty$).

Value functions arising in various optimal control problems have been extensively used since their introduction by Bellman and Isaacs in the fifties. In general, even for smooth f , φ and in the absence of state constraints, the value function of Mayer's problem may be not differentiable. Its lack of differentiability is related to the multiplicity of optimal trajectories (see [19]). This may be also explained by the shocks of characteristics of the associated Hamilton-Jacobi equation, see [25, 26, 55]. Conversely, as it was shown in [16] - [18], the absence of shocks guarantees smoothness of the value function. The Hamiltonian $H : [0, 1] \times \mathbf{R}^n \times \mathbf{R}^n \rightarrow \mathbf{R} \cup \{+\infty\}$ of the Mayer problem is defined by

$$H(t, x, p) = \sup_{u \in U} \langle p, f(t, x, u) \rangle.$$

Under appropriate assumptions, V is a unique solution in a generalized sense to the Hamilton-Jacobi equation

$$-\frac{\partial V}{\partial t} + H\left(t, x, -\frac{\partial V}{\partial x}\right) = 0, \quad V(1, x) = \varphi(x), \quad (t, x) \in [0, 1] \times K. \quad (12)$$

It is well known that (12) does not have smooth solutions and for this reason the notion of solution was extended to non differentiable functions. The most popular are continuous viscosity solutions using superdifferentials and subdifferentials instead of gradients for defining super and subsolutions to (12). See for instance [6, 32, 33, 46] and the references contained therein.

The Hamiltonian H defined above is convex with respect to the last variable. When $K = \mathbf{R}^n$ (no state constraint) this actually allowed to get a simpler definition of lower semicontinuous solution involving only subdifferentials and equalities (see [7] for an approach based on PDE arguments and [50, 51, 58] for the one based on viability theory). An extension to systems under state constraints is given in [57].

The value function is also an important tool for investigating optimality conditions. For instance it follows from [19] that in the absence of state constraints, sufficient conditions for optimality for the Mayer problem can be expressed using extremal points of the generalized gradients of the value function and that the optimal trajectories are unique at points of differentiability of the value function whenever H is smooth enough. Furthermore, the adjoint variable in the maximum principle (discussed below) satisfies some additional relations involving superdifferentials of the value function. These additional relations make the maximum principle not only a necessary but also a sufficient optimality condition (see [19] and [81]). When φ and f are smooth enough, the value function is semiconcave (see [24] for a nice collection of results on semiconcave functions and applications of semiconcavity to problems of optimal control or [19] for both a proof of semiconcavity of the value function of Mayer's problem and sufficient optimality conditions). In the presence of state constraint, in general, V is not semiconcave even for smooth f , φ . Observe that if $\varphi(\cdot)$ is locally Lipschitz, then local Lipschitz continuity of $\mathcal{S}_K(\cdot) : K \rightsquigarrow \mathcal{C}$ yields local Lipschitz continuity of the value function.

• **Maximum principle.** Assume f differentiable with respect to x and φ differentiable. Let $\bar{x}(\cdot)$ be optimal for problem (9) and let $\bar{u}(\cdot)$ be a corresponding control. Then, under some technical assumptions, the celebrated maximum principle under state constraint holds true (see [43] and also [60] for an earlier version): there exist $\lambda \in \{0, 1\}$, an absolutely continuous mapping $p(\cdot) : [0, 1] \rightarrow \mathbf{R}^n$ and a mapping $\psi(\cdot) : [0, 1] \rightarrow \mathbf{R}^n$ of bounded total variation satisfying

(i) the adjoint equation (where $*$ states for the transposition)

$$-p'(t) = \frac{\partial f}{\partial x}(t, \bar{x}(t), \bar{u}(t))^*(p(t) + \psi(t)) \quad \text{a.e. in } [0, 1], \quad (13)$$

(ii) the maximum principle

$$\langle p(t) + \psi(t), f(t, \bar{x}(t), \bar{u}(t)) \rangle = \max_{u \in U} (\langle p(t) + \psi(t), f(t, \bar{x}(t), u) \rangle) \quad \text{a.e. in } [0, 1] \quad (14)$$

and the transversality condition $-p(1) - \psi(1) = \lambda \nabla \varphi(\bar{x}(1))$. Furthermore $\psi(\cdot)$ is linked to the state constraint in the following way : there exist a positive (scalar) Radon measure μ on $[0, 1]$ and a Borel measurable $\nu(\cdot) : [0, 1] \rightarrow \mathbf{R}^n$ satisfying

$$\nu(s) \in N_K(\bar{x}(s)) \cap B \quad \mu - a.e., \quad (15)$$

$$\psi(t) = \int_{[0,t]} \nu(s) d\mu(s) \quad \forall t \in (0, 1]. \quad (16)$$

See the monograph [82] for different forms of maximum principle under state constraints and [72] for some historical comments on the maximum principle.

The above necessary optimality condition is said to be normal if $\lambda = 1$. It was shown in [11] and [53] that a generalized inward pointing condition yields normality of the maximum principle for a class of state constraints with non smooth boundaries. Normality is very useful for investigation of Lipschitz continuity of optimal trajectories (with respect to the time), to avoid the Lavrentieff phenomenon, see [21, 41, 54, 59, 62, 67, 78]. Let us underline that regularity of optimal trajectories is important in discrete approximations and hence for numerical solutions.

• **Adjoint state and gradient of the value function.** It is well known that if $K = \mathbf{R}^n$ and if the value function is differentiable, then $-p(t) = \nabla_x V(t, \bar{x}(t))$ for all t . In [30], for $K = \mathbf{R}^n$, this relation was extended to a locally Lipschitz value function using generalized gradient instead of gradient. It follows from [27] that in the presence of a state constraint if $V(0, \cdot)$ is locally Lipschitz at $x_0 \in \text{Int } K$, then $p(\cdot)$ in the maximum principle satisfies $-p(0) \in \lambda \partial_x V(0, x_0)$, where $\partial_x V(0, x_0)$ denotes the generalized gradient of $V(0, \cdot)$ at x_0 . For K with a smooth boundary, the relation $-p(t) - \psi(t) \in \partial_x V(t, \bar{x}(t))$ for a.e. $t \in [0, 1]$ was recently obtained in [14], using a slightly different notion of generalized gradient on the boundary of K .

• **Regularity of optimal trajectories.** The mapping $\psi(\cdot)$ in the maximum principle may be very irregular and have an infinite number of jumps (see [71] for a relevant example in \mathbf{R}^n with $n \geq 3$). For this reason optimal controls may also be highly discontinuous with respect to the time. However for some classes of nonlinear constrained optimal control problems of Bolza type (see (34) in Section 4 below) this is no longer the case. This was observed first in [62] for linear problems with convex cost and convex state constraints and extended in [67] to constrained control systems that are nonlinear with respect to the state. Further generalizations to systems affine with respect to control under nonlinear state constraints were obtained in [59]. In [53] it was shown that for the Bolza optimal control problem, whose Hamiltonian has a coercive gradient in the directions normal to constraint, $\psi(\cdot)$ is continuous on $(0, 1)$. This helps to investigate the continuity of optimal controls. Moreover, under a uniform coercivity assumption in the directions normal to constraint, $\psi(\cdot)$ becomes absolutely continuous on $(0, 1)$, implying in turn that optimal trajectories have absolutely continuous derivatives. For some classes of control systems this allows to get absolutely continuous and even Lipschitz continuous optimal controls.

• **Outline.** In the next section we discuss the local Lipschitz continuity of $\mathcal{S}_K(\cdot)$ and in Section 3 the local Lipschitz continuity of the value function and optimal synthesis. Section 4 relates the adjoint state $p(0)$ of the maximum principle to the generalized gradient of the value function. Finally, Section 5 is devoted to smoothness of $\psi(\cdot)$ in the maximum principle and regularity of optimal trajectories and controls for the Bolza optimal control problem.

2. Lipschitz Dependence of Viable Trajectories on Initial States and Inverse Mapping Theorems

By $B(x_0, \varepsilon)$ (or $B_X(x_0, \varepsilon)$) we denote the closed ball in a metric space X of center $x_0 \in X$ and radius $\varepsilon > 0$ and by B or B_Y the closed unit ball centered at zero in a Banach space Y . The Euclidean norm in \mathbf{R}^n is denoted by $|\cdot|$.

Let (X, d_X) be a metric space, Y be a Banach space and $G : X \rightsquigarrow Y$ be a set-valued map. G is said to be locally Lipschitz, if it has nonempty values and for every $x_0 \in X$ there exist $\varepsilon > 0$, $L \geq 0$ such that $G(x_1) \subset G(x_2) + L|x_1 - x_2|B_Y$ for all $x_1, x_2 \in B_X(x_0, \varepsilon)$. The graph of G is defined by $\text{Graph}(G) := \{(x, y) \mid y \in G(x)\}$.

Consider a set-valued map $F : [0, 1] \times \mathbf{R}^n \rightsquigarrow \mathbf{R}^n$, a closed set $K \subset \mathbf{R}^n$, $x_0 \in K$ and the differential inclusion

$$\begin{cases} x'(t) & \in F(t, x(t)) \quad \text{for a.e. } t \in [0, 1], \\ x(0) & = x_0. \end{cases} \quad (17)$$

It is worth to underline that if the mapping f from the introduction is measurable with respect to t and continuous with respect to x, u , then the set of trajectories of control system (1) coincides with the set of trajectories of differential inclusion (17) for $F(t, x) = f(t, x, U)$, see for instance [5, Theorem 8.2.10]. Define

$$\mathcal{S}(x_0) := \{x(\cdot) \in W^{1,1}([0, 1]; \mathbf{R}^n) \mid x(\cdot) \text{ satisfies (17)}\},$$

$$\mathcal{S}_K(x_0) := \{x(\cdot) \in \mathcal{S}(x_0) \mid x(t) \in K \text{ for all } t \in [0, 1]\}.$$

We say that $\mathcal{S}_K(\cdot)$ is locally \mathcal{C} -Lipschitz (respectively $W^{1,1}$ -Lipschitz) on a subset $\mathcal{D} \subset K$ if it is locally Lipschitz as a set-valued map from \mathcal{D} into the space \mathcal{C} (respectively into the space $W^{1,1}([0, 1]; \mathbf{R}^n)$).

Theorem 2.1. *Assume there exists $\gamma > 0$ such that $\sup_{v \in F(t, x)} |v| \leq \gamma(1 + |x|)$ and $F(t, x)$ is nonempty and closed for all $(t, x) \in [0, 1] \times \mathbf{R}^n$, that F is locally Lipschitz and that the “inward pointing condition”*

$$F(t, x) \cap \text{Int } C_K(x) \neq \emptyset \quad \forall x \in \partial K, \quad \forall t \in [0, 1] \quad (18)$$

holds true. Then the set-valued map $\mathcal{S}_K(\cdot)$ is locally \mathcal{C} -Lipschitz on K .

The above theorem is an extension of a result due to Filippov [45] to systems under state constraints. In the absence of a state constraint a stronger conclusion holds true : $\mathcal{S}(\cdot)$ is locally $W^{1,1}$ -Lipschitz under less restrictive assumptions, for instance F may be unbounded and measurably dependent on time. Furthermore [45] provides also estimates of the $W^{1,1}$ -distance from an arbitrary $x(\cdot) \in W^{1,1}([0, 1]; \mathbf{R}^n)$ to the set $\mathcal{S}(x_0) \subset W^{1,1}([0, 1]; \mathbf{R}^n)$.

There exist several approaches dealing with the question of Lipschitz continuity of $\mathcal{S}_K(\cdot)$. We briefly describe some of them.

- The **first** one was initiated in [79] for the time independent control systems when the boundary of K is C^2 and extended in [23] to Hilbert spaces. It is based on the local Lipschitz continuity of $\mathcal{S}(\cdot)$ and on a modification of controls in a suitable way to satisfy state constraints.

Actually the very same approach can be used to prove Theorem 2.1. More precisely fix $r > 0$ and $x_0 \in K$. Then there exists $L_r \geq 0$ such that for any $x_1, x_2 \in K \cap B(x_0, r)$ and every $y(\cdot) \in \mathcal{S}_K(x_1)$ we can find $\tilde{z}(\cdot) \in \mathcal{S}(x_2)$ satisfying $\|y - \tilde{z}\|_C \leq L_r|x_1 - x_2|$. Then to prove Theorem 2.1 it is sufficient to find $z(\cdot) \in \mathcal{S}_K(x_2)$ verifying $\|z - \tilde{z}\|_C \leq \bar{c}|x_1 - x_2|$ for a constant $\bar{c} \geq 0$ depending only on (the magnitude of) $r + |x_0|$.

To construct such $z(\cdot)$, assume that $\tilde{z}(\cdot) \notin \mathcal{S}_K(x_2)$ and define $t_0 = \inf\{t \mid \tilde{z}(t) \notin K\}$. By the inward pointing condition, it can be shown that for some $v_0 \in F(t_0, z(t_0)) \cap \text{Int } C_K(z(t_0)), \varepsilon > 0$ and $\tau > t_0$ there exists a trajectory $z(\cdot)$ of the differential inclusion $z'(s) \in F(s, z(s))$ a.e. in $[t_0, \tau]$ such that

$$z(s) \in z(t_0) + (s - t_0)v_0 + B(0, \varepsilon(s - t_0)) \subset \text{Int } K \quad \forall s \in (t_0, \tau].$$

Filippov’s theorem from [45] and the local Lipschitz continuity of $F(\cdot, \cdot)$ imply the existence of a trajectory $x(\cdot)$ to the differential inclusion $x'(s) \in F(s, x(s))$ a.e. in $[\tau, 1]$ such that $|x'(s) - \tilde{z}'(s - \tau + t_0)| \leq c(\tau - t_0)$ for all $s \in [\tau, 1]$ and $x(\tau) = z(\tau)$, where the constant $c \geq 0$ depends only on $|x_0| + r$. It follows that for some $t_1 > \tau$, $x([\tau, t_1]) \subset \text{Int } K$ and either $t_1 = 1$ or $x(t_1) \in \partial K$ and $t_1 < 1$. Denote by $z(\cdot)$ the restriction of $x(\cdot)$ to $[\tau, t_1]$. By assumptions of Theorem 2.1 it is possible to choose $\beta > 0$ and $\alpha > 0$ depending only on $|x_0| + r$ in such way that for some $\tau > t_0$ satisfying $\tau - t_0 \leq \beta|x_1 - x_2|$ we have $t_1 \geq \min\{\tau + \alpha, 1\}$ for sufficiently small $|x_1 - x_2|$.

Repeating the described process (a finite number of times) we construct $z(\cdot)$ on $[0, 1]$ as required. This approach uses a time shift in the definition of $z(\cdot)$ on $[t_0 + \tau, t_1]$ (which is not convenient when applied to some questions arising in differential games, where players have to adapt to each other strategies without knowing the future, i.e. using non anticipative controls).

- The **second** approach uses the so called neighbouring feasible trajectories theorems. These theorems provide a sufficient condition for the existence of $L_r \geq 0$ depending only on $|x_0| + r$ such that for any $y_0 \in K \cap B(x_0, r)$ and $x(\cdot) \in \mathcal{S}(y_0)$ we can find $\bar{x}(\cdot) \in \mathcal{S}_K(y_0)$ satisfying $\|x - \bar{x}\|_C \leq L_r \max_{s \in [0, 1]} \text{dist}(x(s); K)$. This approach was initiated in [47] for differential inclusions under a much stronger inward pointing condition. Constructions proposed in proofs of these theorems are still “anticipative”. Neighbouring feasible trajectories theorems imply the local C –Lipschitz continuity of $\mathcal{S}_K(\cdot)$. We refer to [8] for the most recent neighbouring feasible trajectories theorem in the space $W^{1,1}([0, 1]; \mathbf{R}^n)$ for F depending measurably on time and K having a smooth boundary.

When the boundary of K is non smooth and F is discontinuous with respect to the time, neighbouring feasible trajectories theorems are no longer valid neither in $W^{1,1}([0, 1]; \mathbf{R}^n)$ nor even in \mathcal{C} . Some counterexamples are proposed in [8] with a state independent F and K being a convex cone in \mathbf{R}^2 .

- The **third** approach was initiated in [9] for control systems of the form (1) with bounded f and when the boundary of K is smooth. It was further assumed that the sets $f(t, x, U)$ are convex and closed. Then an extension to unbounded set-valued maps was proposed in [10], where, instead of Lipschitz continuity of $\mathcal{S}_K(\cdot)$, its pseudo-Lipschitz continuity was investigated. The advantage of this construction is of formulating the problem into the viability theory framework and therefore proceeding in a non anticipative way. Let $y(\cdot) \in \mathcal{S}_K(x_1)$ and let $u(\cdot)$ be a control corresponding to $y(\cdot)$. Set $r(t, x) := \text{dist}(f(t, x, u(t)); f(t, x, U) \cap T_K(x))$,

$$G(t, x) := f(t, x, U) \cap B(f(t, x, u(t)), r(t, x))$$

and consider the differential inclusion

$$\begin{cases} z'(t) \in G(t, z(t)) \\ z(0) = x_2 \in K. \end{cases}$$

By the measurable viability theorem from [56] it has a viable trajectory $z(\cdot) \in \mathcal{S}_K(x_2)$. An analysis of $z(\cdot)$ yields estimates $\|y - z\|_{W^{1,1}} \leq L_r|x_1 - x_2|$ for a constant $L_r \geq 0$ depending only on $|x_0| + r$.

- To **summarize**, the first construction allows us to prove the local \mathcal{C} -Lipschitz continuity of $\mathcal{S}_K(\cdot)$. When the boundary of K is sufficiently smooth, the second and third approaches imply $W^{1,1}$ -Lipschitz continuity of $\mathcal{S}_K(\cdot)$ even when F is only measurable in time. Still counterexamples to neighbouring feasible trajectories theorems do exist when K is an intersection of sets with smooth boundaries and F is discontinuous in time. The third approach provides a non anticipative construction when in addition the sets $f(t, x, U)$ are convex and closed.

- We propose now an **inverse mapping theorem** approach to \mathcal{C} -Lipschitz continuity of $\mathcal{S}_K(\cdot)$ on $\text{Int } K$ for constraints with possibly nonsmooth boundary and f measurable with respect to the time.

Recall that for a subset $Q \subset \mathbf{R}^n$ with nonempty boundary ∂Q the oriented distance to ∂Q is the function $d_Q(\cdot) : \mathbf{R}^n \rightarrow \mathbf{R}$ defined by

$$d_Q(x) := \text{dist}(x; Q) - \text{dist}(x; \mathbf{R}^n \setminus Q).$$

See [37, 38] for interesting and detailed discussions about relations between smoothness of the oriented distance function and regularity of the boundary of Q . We write $d_Q \in \mathcal{C}_{loc}^{1,1}$ if for any $r > 0$ there exists $\varepsilon > 0$ such that the gradient of $d_Q(\cdot)$ is Lipschitz continuous on $(\partial Q \cap B(0, r)) + B(0, \varepsilon)$.

In Theorem 2.2 below we suppose that K is an intersection of closed sets having smooth boundaries

$$\left\{ \begin{array}{l} \emptyset \neq K = \cap_{j=1}^m K_j \\ \text{for some closed subsets } K_j \subset \mathbf{R}^n \text{ such that } d_j \in C_{loc}^{1,1} \quad \forall j = 1, \dots, m ; \\ 0 \notin co\{\nabla d_j(x) \mid j \in J(x)\} \quad \forall x \in \partial K, \end{array} \right. \tag{19}$$

where $d_j = d_{K_j}$ and $J(x) = \{j \mid x \in \partial K_j\}$. Then $T_{K_j}(x) = \{v \mid \langle \nabla d_j(x), v \rangle \leq 0\}$ for every $j \in J(x)$, $T_K(x) = \cap_{j=1}^m T_{K_j}(x)$ and $T_K(x) = C_K(x)$.

Can we expect the local \mathcal{C} -Lipschitz continuity of the set-valued map $\mathcal{S}_K(\cdot)$ on the interior of K when f is only measurable in time and the inward pointing condition holds true ? A positive answer is provided in [13] on the basis of an inverse mapping theorem of set-valued analysis. Lipschitz-like properties of inverse maps in an abstract setting were studied by many authors, see for instance [2, 5, 42, 64] and the references contained therein.

Consider the Banach space $\mathcal{C}_0 := \{x(\cdot) \in \mathcal{C} \mid x(0) = 0\}$ with the norm $\|\cdot\|_{\mathcal{C}}$ and define for every $y_0 \in K$ the set

$$\mathcal{K}(y_0) := \{x(\cdot) \in \mathcal{C} \mid x(0) = y_0, x(t) \in K \quad \forall t \in [0, 1]\}.$$

Then $\mathcal{K}(y_0)$ is a complete metric space with the metric induced by the \mathcal{C} -norm. We associate with every $y_0 \in K$ the set-valued map $G_{y_0} : \mathcal{K}(y_0) \rightsquigarrow \mathcal{C}_0$ defined by $G_{y_0}(y(\cdot)) = \mathcal{S}(y_0) - y(\cdot)$ and consider the problem: find $x(\cdot; y_0) \in \mathcal{K}(y_0)$ such that $0 \in \mathcal{S}(y_0) - x(\cdot; y_0)$. Observe that $x(\cdot; y_0) \in \mathcal{S}_K(y_0)$. Thus \mathcal{C} -Lipschitz continuity of $\mathcal{S}_K(\cdot)$ may be studied by investigation of Lipschitz behaviour of the set-valued map $G_{y_0}^{-1}(0)$ with respect to the parameter y_0 . This question is related to set-valued implicit function theorems (see [42] for a very clear exposition of this topic).

Let U be a complete separable metric space and $f : [0, 1] \times \mathbf{R}^n \times U \rightarrow \mathbf{R}^n$. In the theorem below we impose the following assumptions on f

$$\left\{ \begin{array}{l} f \text{ is Lebesgue measurable in } t \text{ and continuous in } u ; \\ \forall r > 0, \exists k_r \geq 0 \text{ such that } f(t, \cdot, u) \text{ is } k_r \text{-Lipschitz on } B(0, r) \quad \forall t, u ; \\ \exists \gamma > 0 \text{ such that } \sup_{u \in U} |f(t, x, u)| \leq \gamma(1 + |x|) \quad \forall t, x ; \\ f(t, x, U) \text{ is compact } \quad \forall t, x \end{array} \right. \tag{20}$$

and the inward pointing condition

$$\left\{ \begin{array}{l} \forall r > 0, \exists \rho_r > 0 \text{ such that } \forall x \in \partial K \cap B(0, r), \forall t \in [0, 1], \\ \exists v_{t,x} \in co f(t, x, U) \text{ satisfying } \langle \nabla d_j(x), v_{t,x} \rangle \leq -\rho_r \quad \forall j \in J(x). \end{array} \right. \tag{21}$$

Theorem 2.2 ([13]). *Assume (19) - (21) and that f is differentiable with respect to x . Then the set-valued map $\mathcal{S}_K(\cdot)$ is locally \mathcal{C} -Lipschitz on $\text{Int } K$.*

To simplify the discussion of methodology for proving Theorem 2.2 via set-valued inverse mapping theorems we consider only the case when in addition

$$f(t, x, U) \text{ is convex for all } t, x.$$

The relaxation theorem and some variational arguments allow to remove this assumption. Then $\text{Graph}(G_{y_0})$ is closed for every $y_0 \in K$. Assumptions (19)-(21), convexity of $f(t, x, U)$ and the measurable viability theorem from [56] imply that $\mathcal{S}_K(x_0) \neq \emptyset$ for all $x_0 \in K$. Furthermore, for every $x_0 \in K$ and $r > 0$ there exists $c(x_0, r) \geq 0$ such that for any $x_1, x_2 \in K \cap B(x_0, r)$ and all $\bar{z}(\cdot) \in \mathcal{S}(x_1)$, $\bar{y}(\cdot) \in \mathcal{K}(x_1)$, we can find $z(\cdot) \in \mathcal{S}(x_2)$ and $y(\cdot) \in \mathcal{K}(x_2)$ satisfying $\|z(\cdot) - \bar{z}(\cdot)\|_C + \|y(\cdot) - \bar{y}(\cdot)\|_C \leq c(x_0, r)|x_1 - x_2|$.

Theorem 2.2 is deduced from a result similar to the classical implicit function theorem. Indeed the underlying idea is to show that for any $y_0 \in \text{Int } K$ and any $\bar{x}(\cdot) \in \mathcal{S}_K(y_0) \subset \mathcal{K}(y_0)$ the “derivative” of G_{y_0} at \bar{x} is surjective. However $\mathcal{K}(y_0)$ being a metric space and G_{y_0} being a set-valued map, derivatives have to be replaced by *set-valued variations* and surjectivity by an *uniform covering property* of variations. Furthermore, Lipschitz continuity of the inverse is replaced by *pseudo-Lipschitz continuity* (also called Aubin continuity in [42]), because “surjectivity” at \bar{x} implies Lipschitz-like behaviour of the inverse only in a neighborhood of $(0, \bar{x})$. In such framework a general inverse mapping theorem from [49] can be applied to deduce the local C -Lipschitz continuity of $\mathcal{S}_K(\cdot)$ on $\text{Int } K$.

Definition 2.3 ([49]). Let $\Phi : X \rightsquigarrow Y$ be a set-valued map from a metric space X to a Banach space Y . The variation of Φ at $(\bar{x}, \bar{y}) \in \text{Graph}(\Phi)$ is the closed subset of Y defined by

$$\Phi^{(1)}(\bar{x}, \bar{y}) := \text{Limsup}_{h \rightarrow 0^+} \frac{\Phi(B(\bar{x}, h)) - \bar{y}}{h}.$$

In the above Limsup stands for the Painlevé-Kuratowski upper limit of sets (see for instance [65, 5]). In other words $v \in \Phi^{(1)}(\bar{x}, \bar{y})$ whenever there exist $h_i > 0$ converging to 0 and $v_i \in Y$ converging to v such that $\bar{y} + h_i v_i \in \Phi(B(\bar{x}, h_i))$.

Let (Y, d_Y) be a metric space. The Hausdorff semidistance between two subsets A and C of Y is defined by

$$e(A; C) := \sup_{a \in A} \text{dist}_Y(a; C) \in \mathbf{R}_+ \cup \{+\infty\},$$

where we set $e(A; C) = +\infty$ if one of the subsets A, C is empty.

Definition 2.4 ([2]). Let $\Psi : X \rightsquigarrow Y$ be a set-valued map from a metric space (X, d_X) to a metric space (Y, d_Y) . Ψ is called pseudo-Lipschitz at $(\bar{\zeta}, \bar{\xi}) \in \text{Graph}(\Psi)$ if there exist $L > 0$ and $\eta > 0$ such that

$$e(\Psi(\zeta) \cap B_Y(\bar{\xi}, \eta); \Psi(\zeta')) \leq L d_X(\zeta, \zeta'), \quad \forall \zeta, \zeta' \in B_X(\bar{\zeta}, \eta).$$

The main assumption of the inverse mapping theorem [49, Theorem 6.1] is a uniform covering property of variations. In terms of our setting, denoting by $\mathcal{B}_{\mathcal{C}_0}$ the closed unit ball in \mathcal{C}_0 , the uniform covering property at $x_0 \in \text{Int } K$ means that for some $\rho > 0$, $\varepsilon_0 > 0$ and $\varepsilon > 0$,

$$\rho \mathcal{B}_{\mathcal{C}_0} \subset G_{y_0}^{(1)}(y(\cdot), z(\cdot) - y(\cdot)) \tag{22}$$

for all $x_1, y_0 \in B(x_0, \varepsilon_0) \subset K$, $\bar{x}(\cdot) \in \mathcal{S}_K(x_1)$, $y(\cdot) \in \mathcal{K}(y_0)$ and $z(\cdot) \in \mathcal{S}(y_0)$ satisfying $\|\bar{x}(\cdot) - y(\cdot)\|_{\mathcal{C}} + \|\bar{x}(\cdot) - z(\cdot)\|_{\mathcal{C}} \leq \varepsilon$. If such uniform covering condition holds true, then the inverse set-valued map $G_{y_0}^{-1}$ defined by

$$G_{y_0}^{-1}(\zeta) = \{\xi \mid \zeta \in G_{y_0}(\xi)\}$$

is pseudo-Lipschitz on a neighbourhood of $(z(\cdot) - y(\cdot), y(\cdot))$ for all $z(\cdot), y(\cdot)$ as above. Furthermore L and η of Definition 2.4 do not depend on $\bar{x}(\cdot), y(\cdot)$ and $z(\cdot)$. That is for some $c \geq 0$, $\eta_0 > 0$, $\eta > 0$ and all $x_1, y_0 \in B(x_0, \eta_0) \subset K$, $\bar{x}(\cdot) \in \mathcal{S}_K(x_1)$, $y(\cdot) \in \mathcal{K}(y_0)$ and $z(\cdot) \in \mathcal{S}(y_0)$ satisfying $\|\bar{x}(\cdot) - y(\cdot)\|_{\mathcal{C}} + \|\bar{x}(\cdot) - z(\cdot)\|_{\mathcal{C}} \leq \eta$ the following inequality holds

$$\text{dist}_{\mathcal{C}}(y(\cdot); G_{y_0}^{-1}(0)) \leq c \|z(\cdot) - y(\cdot)\|_{\mathcal{C}}.$$

This implies Theorem 2.2, because it is enough to choose $z(\cdot) \in \mathcal{S}(x_2)$ and $y(\cdot) \in \mathcal{K}(x_2)$ satisfying $\|z(\cdot) - \bar{x}(\cdot)\|_{\mathcal{C}} + \|y(\cdot) - \bar{x}(\cdot)\|_{\mathcal{C}} \leq c(x_0, \eta_0) \|x_1 - x_2\|$.

The following lemma allows to check the covering property.

Lemma 2.5 ([13]). *Assume (19) - (21), that f is differentiable with respect to x and $x_0 \in \text{Int } K$. Then for some $\alpha > 0$, $\varepsilon > 0$ and $\sigma > 0$, for any $x_1 \in B(x_0, \varepsilon) \subset K$ and $\bar{x}(\cdot) \in \mathcal{S}_K(x_1)$ with $\max_{t \in [0,1]} d_K(\bar{x}(t)) > -\alpha$, the following holds true: if $y_0 \in B(x_0, \varepsilon)$, $y(\cdot) \in \mathcal{K}(y_0)$, $z(\cdot) \in \mathcal{S}(y_0)$ are such that $\|y(\cdot) - \bar{x}(\cdot)\|_{\mathcal{C}} + \|z(\cdot) - \bar{x}(\cdot)\|_{\mathcal{C}} \leq \varepsilon$, then for any control $u(\cdot)$ satisfying $z'(t) = f(t, z(t), u(t))$ a.e., there exist $\delta > 0$, a measurable selection $v(t) \in \text{co } f(t, z(t), U)$ a.e. in $[0, 1]$ and a solution $\bar{w}(\cdot)$ to the linear system*

$$\begin{cases} w'(t) = \frac{\partial f}{\partial x}(t, z(t), u(t))w(t) + r(t)(v(t) - z'(t)), & r(t) \geq 0 \text{ a.e. in } [0, 1] \\ w(0) = 0, \end{cases} \tag{23}$$

such that $\|\bar{w}(\cdot)\|_{\mathcal{C}} \leq \frac{1}{2}$,

$$\max_{t \in [0, \delta]} d_K(y(t)) < 0 \text{ and } \langle \nabla d_j(y(t)), \bar{w}(t) \rangle \leq -\sigma \quad \forall t \in (\delta, 1], j \in J(y(t)). \tag{24}$$

Observe that (23) is a linear control system with non negative scalar controls. To check that variations of G_{y_0} do have a uniform covering property, consider $\alpha, \varepsilon, \sigma$ as in Lemma 2.5. We may assume that $\sigma < 1$ and $\varepsilon < \frac{\alpha}{2}$. Let $x_1, y_0 \in B(x_0, \varepsilon)$, $\bar{x}(\cdot) \in \mathcal{S}_K(x_1)$, $y(\cdot) \in \mathcal{K}(y_0)$ and $z(\cdot) \in \mathcal{S}(y_0)$ be such that $\|y(\cdot) - \bar{x}(\cdot)\|_{\mathcal{C}} + \|z(\cdot) - \bar{x}(\cdot)\|_{\mathcal{C}} \leq \varepsilon$. If $\max_{t \in [0,1]} d_K(\bar{x}(t)) \leq -\alpha$, then

$y(\cdot) + h\mathcal{B}_{\mathcal{C}_0} \subset \mathcal{K}(y_0)$ for all sufficiently small $h > 0$. Hence $z(\cdot) - y(\cdot) + h\mathcal{B}_{\mathcal{C}_0} \subset G_{y_0}(B(y(\cdot), h))$ and therefore $\mathcal{B}_{\mathcal{C}_0} \subset G_{y_0}^{(1)}(y(\cdot), z(\cdot) - y(\cdot))$. Consider next the case $\max_{t \in [0,1]} d_K(\bar{x}(t)) > -\alpha$ and let $\bar{w}(\cdot)$ be as in Lemma 2.5.

By the variational equation of control theory (see for instance [48]), there exist $\bar{w}_h(\cdot)$ converging uniformly to $\bar{w}(\cdot)$ as $h \rightarrow 0+$ such that $z(\cdot) + h\bar{w}_h(\cdot) \in \mathcal{S}(y_0)$. Let $w(\cdot) \in \mathcal{C}_0$ be such that $\|w(\cdot)\|_{\mathcal{C}} \leq \frac{\sigma}{2}$. From (24) we deduce that for all small $h > 0$, $y(\cdot) + h(\bar{w}_h(\cdot) - w(\cdot)) \in \mathcal{K}(y_0) \cap \mathcal{B}_{\mathcal{C}}(y(\cdot), h)$. Therefore for all small $h > 0$,

$$z(\cdot) - y(\cdot) + hw(\cdot) = z(\cdot) + h\bar{w}_h(\cdot) - (y(\cdot) + h(\bar{w}_h(\cdot) - w(\cdot))) \in G_{y_0}(\mathcal{B}_{\mathcal{C}}(y(\cdot), h)),$$

implying that $w(\cdot) \in G_{y_0}^{(1)}(y(\cdot), z(\cdot) - y(\cdot))$. Thus (22) holds true with ρ replaced by $\frac{\sigma}{2}$. Therefore variations do have the announced uniform covering property.

3. Value Function and Optimal Synthesis

Let $\varphi : \mathbf{R}^n \rightarrow \mathbf{R} \cup \{+\infty\}$ be an extended-real-valued lower semicontinuous, bounded from below function and f, U, K be as in the introduction. Consider the Mayer optimal control problem

$$\text{minimize } \{\varphi(x(1)) \mid x(\cdot) \in \mathcal{S}_K(x_0)\} \tag{25}$$

and let $V : [0, 1] \times \mathbf{R}^n \rightarrow \mathbf{R} \cup \{+\infty\}$ be the value function associated to it by (11). Then $\bar{x}(\cdot) \in \mathcal{S}_K(x_0)$ is optimal for the Mayer problem if and only if $V(\cdot, \bar{x}(\cdot)) \equiv \text{const} \neq +\infty$. Therefore if $K = \mathbf{R}^n$ and $V \in C^1$, then, by (12), the set-valued map $\Lambda : [0, 1] \times \mathbf{R}^n \rightsquigarrow U$ given by

$$\Lambda(t, x) := \{u \in U \mid H(t, x, -V'_x(t, x)) = \langle -V'_x(t, x), f(t, x, u) \rangle\}$$

can be seen as an optimal synthesis for the optimal control problem (25). Indeed a trajectory/control pair $(\bar{x}(\cdot), \bar{u}(\cdot))$ is optimal for (25) if and only if $\bar{u}(t) \in \Lambda(t, \bar{x}(t))$ a.e. Thus the set of optimal trajectories coincides with the set of trajectories of

$$x'(t) \in f(t, x(t), \Lambda(t, x(t))) \quad \text{a.e. in } [0, 1], \quad x(0) = x_0. \tag{26}$$

If V is not differentiable, but f is sufficiently smooth with respect to x , then it is still possible to express the optimal synthesis using superdifferentials of the value function. Recall [33] that the superdifferential of a function $g : \mathbf{R}^n \rightarrow \mathbf{R}$ at x is a closed convex, possibly empty, subset of \mathbf{R}^n defined by

$$\partial_+ g(x) = \left\{ p \in \mathbf{R}^n \mid \limsup_{y \rightarrow x} \frac{g(y) - g(x) - \langle p, y - x \rangle}{|y - x|} \leq 0 \right\}.$$

Assume next that f is time independent, that $f(x, U)$ is closed and convex for every x and define for all $t \in (0, 1)$ and $x \in \mathbf{R}^n$

$$\Lambda(t, x) := \{u \in U \mid p_t + \langle p_x, f(x, u) \rangle = 0 \quad \forall (p_t, p_x) \in \partial_+ V(t, x)\}. \tag{27}$$

If $\partial_+ V(t, x) = \emptyset$, then we set $\Lambda(t, x) = \emptyset$. Observe that $f(x, \Lambda(t, x))$ is closed and convex (possibly empty) for all $t \in (0, 1)$ and $x \in \mathbf{R}^n$. From [52, Theorems 4.1 and 4.3], it follows that for smooth enough f and φ , the differential inclusion (26) with $\Lambda(t, x)$ defined by (27) characterizes all optimal trajectories. There is no analogue of this result when f is only Lipschitz continuous. Also, in general, the set-valued map $(t, x) \rightsquigarrow f(x, \Lambda(t, x))$ is not upper semicontinuous.

Directional derivatives of V seem to be better adapted to express “synthesis equations” for optimal trajectories to encompass problems with Lipschitz dynamics and state constraints. For all $t \in [0, 1]$, $x \in K$ such that $V(t, x) \neq +\infty$ and all $\bar{v} \in \mathbf{R}^n$ the contingent derivative of V at (t, x) in the direction $(1, \bar{v})$ is defined by

$$D_{\uparrow} V(t, x)(1, \bar{v}) := \liminf_{h \rightarrow 0+, v \rightarrow \bar{v}} \frac{V(t + h, x + hv) - V(t, x)}{h}.$$

We associate to it the set

$$G(t, x) = \{v \in f(t, x, U) \mid D_{\uparrow} V(t, x)(1, v) \leq 0\}.$$

The proof of the next result is immediate.

Proposition 3.1. *Assume that V is locally Lipschitz on $[0, 1] \times K$. Then $x(\cdot) \in S_K(x_0)$ is optimal for the Mayer problem if and only if*

$$x'(t) \in G(t, x(t)) \text{ a.e. in } [0, 1], \quad x(0) = x_0. \tag{28}$$

A refinement of the results of the previous section allows to deduce the following two theorems about the local Lipschitz continuity of the value function.

Theorem 3.2. *Assume that φ is locally Lipschitz, that for every $r > 0$ there exists $L_r > 0$ such that $f(\cdot, \cdot, u)$ is L_r -Lipschitz on $[0, 1] \times B(0, r)$ for all $u \in U$, that f is continuous with respect to u , that the sets $f(t, x, U)$ are closed and for some $\gamma > 0$, $\sup_{u \in U} |f(t, x, u)| \leq \gamma(1 + |x|)$ for all $(t, x) \in [0, 1] \times \mathbf{R}^n$. If for every $t \in [0, 1]$ and $x \in \partial K$, $f(t, x, U) \cap \text{Int } C_K(x) \neq \emptyset$, then V is locally Lipschitz on $[0, 1] \times K$.*

Theorem 3.3. *Assume (19) - (21), that f is differentiable with respect to x and that φ is locally Lipschitz. Then V is locally Lipschitz on $[0, 1] \times \text{Int } K$.*

Theorem 3.2 and Proposition 3.1 allow to characterize all optimal trajectories of the optimal control problem (25) as trajectories of the differential inclusion (28) when the inward pointing condition is satisfied. The differential inclusion (28) is not simple to handle because, in general, the set-valued map G neither has convex values nor it is upper semicontinuous. If $K = \mathbf{R}^n$ and V is semiconcave (see [19] for sufficient conditions for such regularity of V), then the set-valued map $(t, x) \rightsquigarrow G(t, x)$ is upper semicontinuous. However, in general, for control systems under state constraints V is not semiconcave.

In **conclusion**, the optimal synthesis problem presents the same difficulty than investigation of control systems under state constraints mentioned in the introduction - it leads to control systems (differential inclusions) having highly irregular right-hand sides.

An **alternative** way to characterize optimal trajectories is to consider an extended constrained control system under an extended state constraint that we now describe. Below we denote by $\text{epi}(V)$ the epigraph of V defined by

$$\text{epi}(V) = \{(t, x, r) \in [0, 1] \times K \times \mathbf{R} \mid r \geq V(t, x)\}.$$

Assume that $V(0, x_0) < +\infty$ and consider the following viability problem

$$\left\{ \begin{array}{l} s'(t) = 1, \quad s(0) = 0 \\ x'(t) = f(t, x(t), u(t)), \quad u(t) \in U \text{ a.e. in } [0, 1], \quad x(0) = x_0 \\ z'(t) = 0, \quad z(0) = V(0, x_0) \\ (s(t), x(t), z(t)) \in \text{epi}(V) \text{ for all } t \in [0, 1]. \end{array} \right. \quad (29)$$

Then a trajectory $\bar{x}(\cdot)$ of (1), (2) is optimal for the Mayer problem (25) if and only if for some real-valued absolutely continuous functions $s(\cdot)$ and $z(\cdot)$ defined on $[0, 1]$, the triple $(s(\cdot), \bar{x}(\cdot), z(\cdot))$ satisfies (29).

Observe that if (20) holds true and $f(t, x, U)$ is convex for every $(t, x) \in [0, 1] \times \mathbf{R}^n$, then V is lower semicontinuous and therefore $\text{epi}(V)$ is a closed set. The viability problem (29) is a new control system under a state constraint where two very simply evolving variables (s, z) were added. Such transformations, introduced in [3], now became standard in various applications of viability theory.

It is worth to underline that algorithms for solving (12) approximate numerically the value function and optimal controls, but not super/subdifferentials. Note that once a viable trajectory of (29) has been found, an optimal control can be associated to it by a measurable selection theorem.

Regular optimal synthesis should not be expected for general nonlinear control systems. However a locally Lipschitz continuous approximate optimal synthesis can be derived via non smooth analysis techniques, see [29].

4. Value Function and Maximum Principle

Let $\varphi : \mathbf{R}^n \rightarrow \mathbf{R}$ be locally Lipschitz and f, U, K be as in the introduction. Consider again the Mayer minimization problem (25). This section illustrates how Lemma 2.5, Theorem 3.3 and arguments of convex analysis can be applied to derive a normal first order necessary optimality condition involving the generalized gradient of $V(0, \cdot)$ at $x_0 \in \text{Int } K$. The result we state below is by no means the most general. Notations $\partial\varphi(x_0)$ and $\partial_x V(0, x_0)$ stand respectively for the Clarke generalized gradient of φ and of $V(0, \cdot)$ at x_0 .

Let $NBV([0, 1]; \mathbf{R}^n)$ (Normalized Bounded Variations) denote the space of functions $\psi : [0, 1] \rightarrow \mathbf{R}^n$ of bounded variation on $[0, 1]$, continuous from the right on $(0, 1)$ and such that $\psi(0) = 0$. The norm of $\psi(\cdot) \in NBV([0, 1]; \mathbf{R}^n)$, $\|\psi\|_{TV}$, is the total variation of $\psi(\cdot)$ on $[0, 1]$.

Theorem 4.1 ([27]). *Assume (19) - (21), that f is differentiable with respect to x and let $x_0 \in \text{Int } K$. If $\bar{x}(\cdot) \in \mathcal{S}_K(x_0)$ is optimal for problem (25) and $\bar{u}(\cdot)$ is a control corresponding to $\bar{x}(\cdot)$, then there exist $\psi(\cdot) \in NBV([0, 1]; \mathbf{R}^n)$ and $p(\cdot) \in W^{1,1}([0, 1]; \mathbf{R}^n)$ satisfying the adjoint equation (13), the maximum principle (14), the transversality condition $-p(1) - \psi(1) \in \partial\varphi(\bar{x}(1))$ and linked to the value function by the inclusion*

$$-p(0) \in \partial_x V(0, x_0). \tag{30}$$

Furthermore $\psi(\cdot)$ satisfies (16) for a positive (scalar) Radon measure μ on $[0, 1]$ and a Borel measurable $\nu(\cdot) : [0, 1] \rightarrow \mathbf{R}^n$ as in (15).

It follows from [14], that for a state constraint K with smooth boundary and under more general assumptions on f , $-p(t) - \psi(t) \in \partial_x V(t, \bar{x}(t))$ for a.e. $t \in [0, 1]$ satisfying $\bar{x}(t) \in \text{Int } K$ and also that a related inclusion holds true for a.e. $t \in [0, 1]$ such that $\bar{x}(t) \in \partial K$.

Sketch of proof of Theorem 4.1. By Theorem 3.3, V is locally Lipschitz on $[0, 1] \times \text{Int } K$. Consider the linearized control system

$$w'(t) = \frac{\partial f}{\partial x}(t, \bar{x}(t), \bar{u}(t))w(t) + v(t), \quad v(t) \in T_{\text{co}f(t, \bar{x}(t), U)}(f(t, \bar{x}(t), \bar{u}(t))) \text{ a.e.} \tag{31}$$

and define the convex sets

$$\begin{aligned} \mathcal{S}^L &= \{w(\cdot) \in W^{1,1}([0, 1]; \mathbf{R}^n) \mid w(\cdot) \text{ is a trajectory of (31)}\}, \\ \mathcal{K}^L &= \{w(\cdot) \in \mathcal{C} \mid w(t) \in C_K(\bar{x}(t)) \quad \forall t \in [0, 1]\}. \end{aligned}$$

Then $\text{Int } \mathcal{K}^L = \{w(\cdot) \in \mathcal{C} \mid w(t) \in \text{Int } C_K(\bar{x}(t)) \quad \forall t \in [0, 1]\}$. As in Lemma 2.5, it can be shown that for every $\theta \in \mathbf{R}^n$ there exists $w(\cdot) \in \mathcal{S}^L \cap \text{Int } \mathcal{K}^L$ such that $w(0) = \theta$. Recall that the Clarke directional derivative of a locally Lipschitz function $g : \mathbf{R}^n \rightarrow \mathbf{R}$ at $y_0 \in \mathbf{R}^n$ in the direction θ is defined by

$$g^0(y_0)(\theta) = \limsup_{y \rightarrow y_0, h \rightarrow 0^+} \frac{g(y + h\theta) - g(y)}{h}.$$

Let $V_x^0(0, x_0)(\theta)$ be defined as above for $g(\cdot) = V(0, \cdot)$, $y_0 = x_0$ and let $w(\cdot) \in \mathcal{S}^L \cap \text{Int } \mathcal{K}^L$. Consider any sequence $h_i \rightarrow 0^+$. By the variational inclusion from [48] there exist $x_i(\cdot) \in \mathcal{S}(x_0 + h_i w(0))$ such that $\frac{x_i(\cdot) - \bar{x}(\cdot)}{h_i}$ converges uniformly to $w(\cdot)$ when $i \rightarrow \infty$. Then for all large i , $x_i(\cdot) \in \mathcal{S}_K^{h_i}(x_0 + h_i w(0))$. Since V is nondecreasing along viable trajectories, it follows that $\varphi(x_i(1)) \geq V(0, x_0 + h_i w(0))$. Therefore, from the optimality of $\bar{x}(\cdot)$ we deduce that $\varphi^0(\bar{x}(1))(w(1)) +$

$V_x^0(0, x_0)(-w(0)) \geq 0$. Denoting by $\text{cl } \mathcal{S}^L$ the closure of \mathcal{S}^L in the space \mathcal{C} , we obtain

$$\varphi^0(\bar{x}(1))(w(1)) + V_x^0(0, x_0)(-w(0)) \geq 0 \quad \forall w(\cdot) \in \text{cl } \mathcal{S}^L \cap \text{Int } \mathcal{K}^L. \tag{32}$$

Define the linear operator $\gamma : \mathcal{C} \rightarrow \mathbf{R}^n \times \mathbf{R}^n$ by $\gamma(x(\cdot)) = (x(0), x(1))$ for all $x(\cdot) \in \mathcal{C}$. For a subset $\mathcal{E} \subset \mathcal{C}$, let $[\mathcal{E}]^+$ denote its positive polar cone. Inequality (32) implies that for some $a \in -\partial_x V(0, x_0)$ and $b \in \partial\varphi(\bar{x}(1))$ we have $\gamma^*(a, b) \in [\text{cl } \mathcal{S}^L \cap \text{Int } \mathcal{K}^L]^+ = [\text{cl } \mathcal{S}^L]^+ + [\mathcal{K}^L]^+$. Hence for some $\beta \in [\mathcal{K}^L]^+$

$$\gamma^*(a, b) - \beta \in [\text{cl } \mathcal{S}^L]^+. \tag{33}$$

Using [74] it can be deduced that there exists $\psi(\cdot) \in NBV([0, 1]; \mathbf{R}^n)$ satisfying (16) for a positive (scalar) Radon measure μ on $[0, 1]$ and a Borel measurable $\nu(\cdot) : [0, 1] \rightarrow \mathbf{R}^n$ as in (15) such that for every $x(\cdot) \in \mathcal{C}$, $\beta(x) = \int_0^1 x(t) d\psi(t)$ (the Stieltjes integral) and $\|\beta\| = \|\psi\|_{TV}$ (see [27] for details). Observe that \mathcal{S}^L is the set of trajectories of a linear control system without state constraint. A direct analysis of (33) allows to conclude that $b = -p(1) - \psi(1)$ and $a = p(0)$ for some $p(\cdot)$ as in (13), (14). \square

Remark 4.2. The derived necessary optimality condition is normal. Assumption (21) yields $\mathcal{S}^L \cap \text{Int } \mathcal{K}^L \neq \emptyset$. Without assuming (21) this intersection may be empty. Still a necessary optimality condition can be obtained by applying the separation theorem to the convex sets \mathcal{S}^L and $\text{Int } \mathcal{K}^L$. The necessary condition is then abnormal ($\lambda = 0$) and $p(0) = 0$ (see [27, Proof of Theorem 3.4], where a similar result was derived for a differential inclusion under state and end point constraints).

• **Maximum principle of the Bolza problem.** In the next section we shall use the maximum principle of a Bolza optimal control problem under state and end point constraints that we recall now. For every $x(\cdot) \in \mathcal{S}_K(x_0)$ let us denote by $\mathcal{U}(x(\cdot))$ the set of all controls corresponding to $x(\cdot)$, that is $u(\cdot) \in \mathcal{U}(x(\cdot))$ if and only if $u : [0, 1] \rightarrow U$ is Lebesgue measurable and $x'(t) = f(t, x(t), u(t))$ a.e.

Let $K_1 \subset \mathbf{R}^n$ and $\ell : [0, 1] \times \mathbf{R}^n \times U \rightarrow \mathbf{R}$. Consider the minimization problem

$$\min\{\varphi(x(1)) + \int_0^1 \ell(s, x(s), u(s)) ds \mid x(\cdot) \in \mathcal{S}_K(x_0), u(\cdot) \in \mathcal{U}(x(\cdot)), x(1) \in K_1\}. \tag{34}$$

Denote by $M(n \times n)$ the set of $n \times n$ matrices and for every $\lambda \geq 0$, define the Hamiltonian $H_\lambda : [0, 1] \times \mathbf{R}^n \times \mathbf{R}^n \rightarrow \mathbf{R} \cup \{+\infty\}$ of the Bolza problem by

$$H_\lambda(t, x, p) = \sup_{u \in U} (\langle p, f(t, x, u) \rangle - \lambda \ell(t, x, u)).$$

Definition 4.3. A trajectory/control pair $(\bar{x}(\cdot), \bar{u}(\cdot))$ of (1), (2) with $\bar{x}(1) \in K_1$ satisfies the maximum principle (of problem (34)) if there exist $\lambda \in \{0, 1\}$, $\psi(\cdot) \in NBV([0, 1]; \mathbf{R}^n)$ and $p(\cdot) \in W^{1,1}([0, 1]; \mathbf{R}^n)$ not vanishing simultaneously such that for some $\pi_1 \in \mathbf{R}^n$ and some integrable $A(\cdot) : [0, 1] \rightarrow M(n \times n)$, $\pi(\cdot) : [0, 1] \rightarrow \mathbf{R}^n$

$$-p'(t) = A(t)^*(p(t) + \psi(t)) - \lambda\pi(t) \text{ a.e. in } [0, 1], \tag{35}$$

$$-p(1) - \psi(1) \in \lambda\pi_1 + N_{K_1}(\bar{x}(1)), \tag{36}$$

$$\langle p(t) + \psi(t), f(t, \bar{x}(t), \bar{u}(t)) \rangle - \lambda\ell(t, \bar{x}(t), \bar{u}(t)) = H_\lambda(t, \bar{x}(t), p(t) + \psi(t)) \text{ a.e.} \tag{37}$$

and (16) holds true for a positive (scalar) Radon measure μ on $[0, 1]$ and a Borel measurable $\nu(\cdot) : [0, 1] \rightarrow \mathbf{R}^n$ satisfying (15). The maximum principle is called normal if $\lambda = 1$.

Remark 4.4. If there exist $\varepsilon > 0$ and $k(\cdot) \in L^1(0, 1)$ such that $f(t, \cdot, u)$ and $\ell(t, \cdot, u)$ are $k(t)$ -Lipschitz on $B(\bar{x}(t), \varepsilon)$ for a.e. $t \in [0, 1]$ and all $u \in U$, then under some mild regularity assumptions on f and ℓ , every optimal trajectory/control pair $(\bar{x}(\cdot), \bar{u}(\cdot))$ satisfies the maximum principle of Definition 4.3 for some $A(t) \in \partial_x f(t, \bar{x}(t), \bar{u}(t))$ (generalized Jacobian of $f(t, \cdot, \bar{u}(t))$ at $\bar{x}(t)$), $\pi(t) \in \partial_x \ell(t, \bar{x}(t), \bar{u}(t))$ (generalized gradient of $\ell(t, \cdot, \bar{u}(t))$ at $\bar{x}(t)$), $\pi_1 \in \partial\varphi(\bar{x}(1))$ and some $(\lambda, p, \psi) \neq 0$ (see [82]).

Furthermore, some other maximum principles that differ only in the adjoint equation (having the Hamiltonian or the Euler-Lagrange form) can be rewritten with the adjoint equation like (35). For instance, if $p(\cdot) \in W^{1,1}([0, 1]; \mathbf{R}^n)$ and $k(\cdot) \in L^1([0, 1]; \mathbf{R}_+)$, are such that $|p'(t)| \leq k(t)|p(t) + \psi(t)| + \lambda k(t)$ a.e., then it is not difficult to find $A(t)$ and $\pi(t)$ not necessarily related to the generalized Jacobian of $f(t, \cdot, \bar{u}(t))$ and to the generalized gradient of $\ell(t, \cdot, \bar{u}(t))$ such that $A(\cdot)$ and $\pi(\cdot)$ are integrable and (35) holds true. This is particularly useful for deducing normality for other forms of the maximum principle.

• **Normality of the maximum principle.** We provide next a sufficient condition for normality of the maximum principle of the Bolza problem when $x_0 \in \text{Int } K$ under the following *outward* pointing condition

$$\begin{cases} \forall r > 0, \exists \sigma_r > 0 \text{ such that } \forall t \in [0, 1], \forall x \in \partial K \cap B(0, r), \\ \exists v_{t,x} \in \text{co } f(t, x, U) \text{ satisfying } \langle \nabla d_j(x), v_{t,x} \rangle \geq \sigma_r \quad \forall j \in J(x). \end{cases} \tag{38}$$

In Proposition 4.5 below, (38) can be assumed only for $x = \bar{x}(t)$ and all $t \in [0, 1]$.

Consider a trajectory/control pair $(\bar{x}(\cdot), \bar{u}(\cdot))$ of (1), (2) satisfying the maximum principle (of problem (34)) for some (λ, p, ψ) , π_1 , $\pi(\cdot)$ and $A(\cdot)$. The reachable set from zero at time $t \in [0, 1]$ of the linear control system

$$w'(t) = A(t)w(t) + v(t), \quad v(t) \in T_{\text{co}f(t, \bar{x}(t), U)}(f(t, \bar{x}(t), \bar{u}(t))) \text{ a.e.} \tag{39}$$

is a convex cone in \mathbf{R}^n defined by

$$\mathcal{R}(t) = \{w(t) \mid w(\cdot) \text{ is a trajectory of (39) on } [0, t], w(0) = 0\}.$$

Proposition 4.5. *Assume (19), (20), (38). Let $x_0 \in \text{Int } K$ and let a trajectory/control pair $(\bar{x}(\cdot), \bar{u}(\cdot))$ of (1), (2) with $\bar{x}(1) \in K_1$ satisfy the maximum principle for some $(\lambda, p, \psi) \neq 0$ and $A(\cdot)$. Further assume that $\bar{x}([0, 1]) \cap \partial K \neq \emptyset$, $\text{Int } C_{K_1}(\bar{x}(1)) \cap \text{Int } T_K(\bar{x}(1)) \neq \emptyset$ and for $t_0 = \min\{t \in [0, 1] \mid \bar{x}(t) \in \partial K\}$ we have $\text{Int } T_K(\bar{x}(t_0)) \subset \mathcal{R}(t_0)$. Then $\lambda = 1$.*

Proof. As in [53], normality will follow if there exists a solution $w(\cdot)$ to (8) satisfying $0 \neq w(1) \in \text{Int } C_{K_1}(\bar{x}(1))$. Fix any $0 \neq w_1 \in \text{Int } C_{K_1}(\bar{x}(1)) \cap \text{Int } T_K(\bar{x}(1))$. Combining constructions of [53, Corollary 6.4] and [11, Proof of Theorem 3.2] (made backward in time) we obtain a solution $w(\cdot)$ to (39) defined on $[t_0, 1]$ and satisfying $w(1) = w_1$ and $w(t) \in \text{Int } T_K(\bar{x}(t))$ for all $t \in [t_0, 1]$. Because $\text{Int } T_K(\bar{x}(t_0)) \subset \mathcal{R}(t_0)$, $w(\cdot)$ can be extended on the time interval $[0, t_0]$ by a solution to (39) in such way that $w(0) = 0$. Since $\bar{x}([0, t_0]) \subset \text{Int } K$, the proof is complete. \square

When the end point is free, i.e. $K_1 = \mathbf{R}^n$, some sufficient *inward* pointing conditions for normality can be found in [11, 53, 54]. In [73] for a free end point optimal control problem the normal maximum principle was derived by the penalization of a state constraint satisfying the inward pointing condition.

5. Regularity of Optimal Trajectories and Controls

Consider again the Bolza problem (34). We discuss here regularity (with respect to the time) of trajectories and controls satisfying the normal maximum principle.

Let the Hamiltonian H_1 be defined as in the previous section. Recall that $H_1(t, x, \cdot)$ is convex and for every $q \in \mathbf{R}^n$ and $u \in U$ satisfying $H_1(t, x, q) = \langle q, f(t, x, u) \rangle - \ell(t, x, u)$, we have $f(t, x, u) \in \partial_p H_1(t, x, q)$, where $\partial_p H_1(t, x, q)$ denotes the subdifferential of $H_1(t, x, \cdot)$ at q . In Proposition 5.1 below we consider only Lipschitz continuous optimal trajectories. A sufficient condition for the existence of a Lipschitz continuous optimal trajectory for the Bolza problem can be found for instance in [21, 54]. We also impose some global assumptions on H_1 . However most of them can be localized for $H_1(t, \bar{x}(t), \cdot)$, where $\bar{x}(\cdot)$ is a trajectory of the control system (1), (2) under investigation. Define

$$\mathcal{M} := \{(t, x, f(t, x, u), r) \mid t \in [0, 1], x \in K, u \in U, r \geq \ell(t, x, u)\}.$$

Proposition 5.1 ([53]). *Assume (19), that \mathcal{M} is closed, that H_1 is continuous and that $H_1(t, x, \cdot)$ is differentiable for all $(t, x) \in [0, 1] \times K$. Let $(\bar{x}(\cdot), \bar{u}(\cdot))$ satisfy the normal maximum principle for some $p(\cdot), \psi(\cdot)$. If $\bar{x}(\cdot)$ is Lipschitz, then $\bar{x}(\cdot) \in C^1([0, 1])$, the mapping $(0, 1) \ni t \mapsto \frac{\partial H_1}{\partial p}(t, \bar{x}(t), p(t) + \psi(t))$ is continuous and $\bar{x}'(t) = \frac{\partial H_1}{\partial p}(t, \bar{x}(t), p(t) + \psi(t))$ for every $t \in (0, 1)$. Furthermore, $\psi(\cdot)$ is continuous on $(0, 1)$ provided for every $t \in (0, 1)$, $x \in \partial K$ and $p, q \in \mathbf{R}^n$*

the following implication holds true

$$\left(\begin{array}{l} p - q \in N_K(x) \\ H_1(t, x, p) = H_1(t, x, q) \\ \frac{\partial H_1}{\partial p}(t, x, p) = \frac{\partial H_1}{\partial p}(t, x, q) \end{array} \right) \implies p = q. \tag{40}$$

Observe that (40) is satisfied, in particular, when $\frac{\partial H_1}{\partial p}(t, x, \cdot)$ is strictly monotone in the directions normal to K at every $x \in \partial K$, i.e. when for every $t \in (0, 1)$ and all $p \neq q \in \mathbf{R}^n$ satisfying $p - q \in N_K(x)$ we have

$$\left\langle \frac{\partial H_1}{\partial p}(t, x, p) - \frac{\partial H_1}{\partial p}(t, x, q), p - q \right\rangle > 0.$$

If for all t, x , the Hamiltonian $H_1(t, x, \cdot)$ is twice differentiable and has strictly positive second derivative, then the last inequality is satisfied for all $p \neq q \in \mathbf{R}^n$.

Note that if $\frac{\partial H_1}{\partial p}$ is continuous, then every $\bar{x}(\cdot)$ satisfying the normal maximum principle (for some control $\bar{u}(\cdot)$) is Lipschitz continuous. The next theorem provides a sufficient condition for the absolute continuity of the mapping $\psi(\cdot)$ on $(0, 1)$.

Theorem 5.2 ([53]). *Assume (19), that \mathcal{M} is closed, H_1 is continuous, $H_1(t, x, \cdot)$ is differentiable for all $(t, x) \in [0, 1] \times K$ and $\frac{\partial H_1}{\partial p}$ is locally Lipschitz on $[0, 1] \times K \times \mathbf{R}^n$. Further assume that for every $t \in (0, 1)$, $x \in \partial K$ and $p, q \in \mathbf{R}^n$ the implication (40) holds true and that for every $r > 0$ there exist $k_r > 0$ and $\bar{\varepsilon} > 0$ such that for all $t \in [0, 1]$, $x \in \partial K \cap B(0, r)$ and $p, q \in B(0, r)$ we have*

$$(p - q \in N_K(x) \cap B(0, \bar{\varepsilon})) \implies \left\langle \frac{\partial H_1}{\partial p}(t, x, p) - \frac{\partial H_1}{\partial p}(t, x, q), p - q \right\rangle \geq k_r |p - q|^2.$$

If $(\bar{x}(\cdot), \bar{u}(\cdot))$ satisfy the normal maximum principle for some $p(\cdot)$, $\psi(\cdot)$, then $\psi(\cdot)$ is absolutely continuous on $(0, 1)$ and $\bar{x}'(\cdot)$ is absolutely continuous on $[0, 1]$.

Moreover, if $p(\cdot)$ is Lipschitz, then $\psi(\cdot)$ is Lipschitz on $(0, 1)$ and $\bar{x}'(\cdot)$ is Lipschitz on $[0, 1]$.

The coercivity assumption of the above theorem is automatically satisfied for all $p, q \in B(0, r)$ if $H_1(t, x, \cdot)$ is twice differentiable, $\frac{\partial^2 H_1}{\partial p^2}$ is continuous and $\frac{\partial^2 H_1}{\partial p^2}(t, x, \cdot) > 0$ for all $(t, x) \in [0, 1] \times \partial K$. Observe that $p(\cdot)$ is Lipschitz whenever $A(\cdot)$ and $\pi(\cdot)$ of the maximum principle are essentially bounded.

The proof of the above theorem relies on an induction argument developed in [62] for linear control systems, convex Lagrangian ℓ and convex state constraints. Some sufficient conditions for Hölder continuity of derivatives of optimal trajectories can be found in [12] and [78].

Corollary 5.3. *Under all the assumptions of Theorem 5.2, suppose that H_1 is locally Lipschitz. If $(\bar{x}(\cdot), \bar{u}(\cdot))$ satisfy the normal maximum principle for some $p(\cdot), \psi(\cdot)$, then there exists an absolutely continuous function $\phi : [0, 1] \rightarrow \mathbf{R}$ such that $\phi(t) = \ell(t, \bar{x}(t), \bar{u}(t))$ a.e. Moreover if $p(\cdot)$ is Lipschitz, then $\phi(\cdot)$ is Lipschitz.*

By Theorem 5.2, $\phi(\cdot)$ defined by $\phi(t) = \langle p(t) + \psi(t), \bar{x}'(t) \rangle - H_1(t, \bar{x}(t), p(t) + \psi(t))$ for $t \in (0, 1)$ and $\phi(0) = \phi(0+), \phi(1) = \phi(1-)$ is absolutely continuous. Furthermore, by (37), $\phi(t) = \ell(t, \bar{x}(t), \bar{u}(t))$ a.e. in $[0, 1]$ implying the above Corollary.

Regularity of $\psi(\cdot), p(\cdot)$ and $\bar{x}'(\cdot)$ helps to study regularity of optimal controls with respect to the time.

Proposition 5.4. *Assume that U is a closed convex subset of \mathbf{R}^m , that f, ℓ are defined on $[0, 1] \times \mathbf{R}^n \times \mathbf{R}^m$ and are continuous, and that $f(t, x, \cdot), \ell(t, x, \cdot)$ are differentiable for all $(t, x) \in [0, 1] \times K$. Define $\mathcal{H} : [0, 1] \times \mathbf{R}^n \times \mathbf{R}^m \times \mathbf{R}^m \rightarrow \mathbf{R}$ by*

$$\mathcal{H}(t, x, p, u) := \langle p, f(t, x, u) \rangle - \ell(t, x, u).$$

If for some $\Phi : \mathbf{R}_+ \times \mathbf{R}_+ \rightarrow \mathbf{R}$ satisfying $\lim_{k \rightarrow +\infty} \frac{\Phi(k, r)}{k} = +\infty$ for every $r > 0$,

- $$\left\{ \begin{array}{l} i) \ell(t, x, u) \geq \Phi(|f(t, x, u)|, r) \quad \forall (t, x) \in [0, 1] \times (K \cap B(0, r)), r > 0, u \in U; \\ ii) \lim_{|u| \rightarrow +\infty} (\inf_{(t, x) \in [0, 1] \times (K \cap B(0, r))} \ell(t, x, u)) = +\infty \quad \forall r > 0; \\ iii) \forall (t, x, p) \in [0, 1] \times K \times \mathbf{R}^n, \forall u_1, u_2 \in U \text{ with } u_1 \neq u_2, \\ \quad \langle \frac{\partial \mathcal{H}}{\partial u}(t, x, p, u_1) - \frac{\partial \mathcal{H}}{\partial u}(t, x, p, u_2), u_2 - u_1 \rangle > 0, \end{array} \right.$$

then for every $(t, x, p) \in [0, 1] \times K \times \mathbf{R}^n$, there exists a unique $v(t, x, p) \in U$ such that $H_1(t, x, p) = \mathcal{H}(t, x, p, v(t, x, p))$. Furthermore $H_1(\cdot, \cdot, \cdot)$ and $v(\cdot, \cdot, \cdot)$ are continuous.

Corollary 5.5. *Under all the assumptions of Proposition 5.4, let $(\bar{x}(\cdot), \bar{u}(\cdot))$ satisfy the normal maximum principle for some $p(\cdot), \psi(\cdot)$. If $\psi(\cdot)$ is continuous on $(0, 1)$, then there exists a continuous mapping $u_0(\cdot) : [0, 1] \rightarrow U$ such that $u_0(\cdot) = \bar{u}(\cdot)$ a.e. in $[0, 1]$. Consequently $\bar{x}(\cdot)$ is Lipschitz.*

Furthermore, if $v(\cdot, \cdot, \cdot)$ is locally Lipschitz on $[0, 1] \times K \times \mathbf{R}^n$ and $\psi(\cdot)$ is absolutely continuous on $(0, 1)$, then $u_0(\cdot)$ is absolutely continuous and if $p(\cdot)$ is Lipschitz and $\psi(\cdot)$ is Lipschitz on $(0, 1)$, then $u_0(\cdot)$ is Lipschitz.

Proof. By Proposition 5.4, for every $t \in (0, 1)$ there exists a unique $u_0(t) := v(t, \bar{x}(t), p(t) + \psi(t)) \in U$ with $\mathcal{H}(t, \bar{x}(t), p(t) + \psi(t), u_0(t)) = H_1(t, \bar{x}(t), p(t) + \psi(t))$. Let $\tilde{\psi}(\cdot) \in \mathcal{C}$ be such that $\tilde{\psi}(\cdot) = \psi(\cdot)$ on $(0, 1)$. Then $u_0(\cdot) := v(\cdot, \bar{x}(\cdot), p(\cdot) + \tilde{\psi}(\cdot))$ is continuous on $[0, 1]$. By (37), $u_0(t) = \bar{u}(t)$ for a.e. $t \in [0, 1]$. The remaining statements follow from the very definition of $u_0(\cdot)$. \square

For f affine with respect to controls, the local Lipschitz continuity of $v(\cdot, \cdot, \cdot)$ follows, for instance, from the assumptions of [78]. For such control systems this

question is related to Lipschitz continuity of a conjugate function. Assume that $f(t, x, u) = a(t, x) + g(t, x)u$ for some continuous $a : [0, 1] \times \mathbf{R}^n \rightarrow \mathbf{R}^n$ and $g : [0, 1] \times \mathbf{R}^n \rightarrow M(n \times m)$, where $u \in \mathbf{R}^m$ and let $\ell : [0, 1] \times \mathbf{R}^n \times \mathbf{R}^m \rightarrow \mathbf{R}$ be continuous and convex with respect to the last variable. Consider a closed convex subset $U \subset \mathbf{R}^m$ and assume that $\ell(t, x, \cdot)$ is differentiable and for all $u_1, u_2 \in U$ with $u_1 \neq u_2$

$$\left\langle \frac{\partial \ell}{\partial u}(t, x, u_1) - \frac{\partial \ell}{\partial u}(t, x, u_2), u_1 - u_2 \right\rangle > 0.$$

Then all the conclusions of Proposition 5.4 hold true whenever there exists $\theta : \mathbf{R}_+ \times \mathbf{R}_+ \rightarrow \mathbf{R}$ such that for every $r > 0$, $\lim_{k \rightarrow +\infty} \frac{\theta(k, r)}{k} = +\infty$ and $\ell(t, x, u) \geq \theta(|u|, r)$ for all $(t, x) \in [0, 1] \times (K \cap B(0, r))$ and $u \in U$. Let $\iota_U(\cdot)$ be the indicator function of U and denote by $\ell^F(t, x, \cdot)$ the Fenchel conjugate of $\ell(t, x, \cdot) + \iota_U(\cdot)$. Then, by the uniqueness of $v(t, x, p)$, the function $y \mapsto \ell^F(t, x, y)$ is differentiable and $v(t, x, p) = \frac{\partial \ell^F}{\partial y}(t, x, g(t, x)^* p)$. If $\frac{\partial \ell^F}{\partial y}(\cdot, \cdot, \cdot)$ is locally Lipschitz and $g(\cdot, \cdot)$ is locally Lipschitz, then also $v(\cdot, \cdot, \cdot)$ is locally Lipschitz.

When the mapping

$$[0, 1] \times K \times \mathbf{R}^n \ni (t, x, q) \rightsquigarrow \Upsilon(t, x, q) := \{u \in U \mid \mathcal{H}(t, x, q, u) = H_1(t, x, q)\}$$

is multivalued, then it may happen that several controls give rise to the same trajectory and an optimal control may be discontinuous. If Υ enjoys some regularity properties, then, taking its selections, it is still possible to deduce the existence of regular optimal controls from the regularity of $p(\cdot) + \psi(\cdot)$ and $\bar{x}(\cdot)$. In general however we can not expect Lipschitz and even continuous optimal controls in the nonlinear case even under assumptions like in Theorem 5.2 guaranteeing Lipschitz continuity of derivatives of optimal trajectories.

Acknowledgement

The author is grateful to Piernicola Bettioli, Piermarco Cannarsa, Pierre Cardaliaguet, Asen Dontchev and Sylvain Sorin for many constructive remarks concerning the preliminary version of this manuscript.

References

- [1] Arutyunov A. V. & Aseev S. M. (1997) *Investigation of the degeneracy phenomenon of the maximum principle for optimal control problems with state constraints*, SIAM J. Control Optim., **35**, 930–952.
- [2] Aubin J.-P. (1984) *Lipschitz behavior of solutions to convex minimization problems*, Mathematics of Oper. Res., **9**, 87–111.
- [3] Aubin J.-P. (1991) VIABILITY THEORY, Birkhäuser, Boston.

- [4] Aubin J.-P. & Cellina A. (1984) DIFFERENTIAL INCLUSIONS, Grundlehren der Mathematischen Wissenschaften, Springer-Verlag, Berlin.
- [5] Aubin J.-P. & Frankowska H. (1990) SET-VALUED ANALYSIS, Birkhäuser, Boston.
- [6] Bardi M. & Capuzzo Dolcetta I. (1997) OPTIMAL CONTROL AND VISCOSITY SOLUTIONS OF HAMILTON - JACOBI - BELLMAN EQUATIONS, Birkhäuser, Boston.
- [7] Barron E. N. & Jensen R. (1990) *Semicontinuous viscosity solutions of Hamilton-Jacobi equations with convex Hamiltonians*, Comm. Partial Diff. Eq., **15**, 1713–1742.
- [8] Bettiol P., Bressan A. & Vinter R. B. (2009) *On trajectories satisfying a state constraint: $W^{1,1}$ -estimates and counter-examples*, Preprint.
- [9] Bettiol P., Cardaliaguet P. & Quincampoix M. (2006) *Zero-sum state constrained differential games: Existence of value for Bolza problem*, International Journal of Game Theory, **34**, 495–527.
- [10] Bettiol P. & Frankowska H. (2007) *Regularity of solution maps of differential inclusions for systems under state constraints*, Set-Valued Analysis, **15**, 21–45.
- [11] Bettiol P. & Frankowska H. (2007) *Normality of the maximum principle for non convex constrained Bolza problems*, J. Diff. Eqs., **243**, 256–269.
- [12] Bettiol P. & Frankowska H. (2008) *Hölder continuity of adjoint states and optimal controls for state constrained problems*, Applied Math. and Optim., **57**, 125–147.
- [13] Bettiol P. & Frankowska H. (2010) *On the solution map to control systems with multiple state constraints*, Preprint.
- [14] Bettiol P. & Vinter R. B. (2010) *Sensitivity interpretations of the co-state variable for optimal control problems with state constraints*, SIAM J. Control Optim., **48**, 3297–3317.
- [15] Bonnans J. F. & Hermant A. (2009) *Second-order analysis for optimal control problems with pure and mixed state constraints*, Annales de l'I.H.P. - Nonlinear Analysis, **26**, 561–598.
- [16] Byrnes Ch. (1998) *The solution of nonlinear Lagrange and Bolza problems via Riccati partial differential equations*, J. Math. Systems, Estimation and Control, **8**, 1–54.
- [17] Byrnes Ch. & Frankowska H. (1992) *Unicité des solutions optimales et absence de chocs pour les équations d'Hamilton-Jacobi-Bellman et de Riccati*, Comptes-Rendus de l'Académie des Sciences, **315**, Série 1, Paris, 427–431.
- [18] Byrnes Ch. & Frankowska H. (1998) *Uniqueness of optimal trajectories and the nonexistence of shocks for Hamilton-Jacobi-Bellman and Riccati partial differential equations*, LECTURE NOTES IN NONLINEAR ANALYSIS, **2**, J. Schauder Center for Nonlinear Studies, 89–112.
- [19] Cannarsa P. & Frankowska H. (1991) *Some characterizations of optimal trajectories in control theory*, SIAM J. Control Optim., **29**, 1322–1347.
- [20] Cannarsa P., Frankowska H. & Marchini E. (2009) *Existence and Lipschitz regularity of solutions to Bolza problems in optimal control*, Trans. Amer. Math. Soc., **361**, 4491–4517.

- [21] Cannarsa P., Frankowska H. & Marchini E. (2009) *On Bolza optimal control problems with constraints*, Discrete and Continuous Dynamical Systems, **11**, 629–653.
- [22] Cannarsa P., Frankowska H. & Sinestrari C. (2000) *Optimality conditions and synthesis for the minimum time problem*, J. Set-Valued Analysis, **8**, 127–148.
- [23] Cannarsa P., Gozzi F. & Soner H. M. (1991) *A boundary value problem for Hamilton-Jacobi equations in Hilbert spaces*, Applied Math. and Optim., **24**, 197–220.
- [24] Cannarsa P. & Sinestrari C. (2004) SEMICONCAVE FUNCTIONS, HAMILTON-JACOBI EQUATIONS, AND OPTIMAL CONTROL, Birkhäuser, Boston.
- [25] Caroff N. & Frankowska H. (1992) *Optimality and characteristics of Hamilton-Jacobi-Bellman equations*, International Series of Numerical Mathematics, **107**, Birkhäuser Verlag, Basel, 169–180.
- [26] Caroff N. & Frankowska H. (1996) *Conjugate points and shocks in nonlinear optimal control*, Trans. Amer. Math. Soc., **348**, 3133–3153.
- [27] Cernea A. & Frankowska H. (2006) *A connection between the maximum principle and dynamic programming for constrained control problems*, SIAM J. Control Optim., **44**, 673–703.
- [28] Clarke F. (1983) OPTIMIZATION AND NONSMOOTH ANALYSIS, Wiley-Interscience, New York.
- [29] Clarke F., Rifford L. & Stern R. J. (2002) *Feedback in state constrained optimal control*, ESAIM Control Optim. Calc. Var., **7**, 97–133.
- [30] Clarke F. & Vinter R. B. (1987) *The relationship between the maximum principle and dynamic programming*, SIAM J. Control Optim., **25**, 1291–1311.
- [31] Clarke F. & Vinter R. B. (1990) *Regularity properties of optimal controls*, SIAM J. Control Optim., **28**, 980–997.
- [32] Crandall M. G. & Lions P. L. (1983) *Viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., **277**, 1–42.
- [33] Crandall M. G., Evans C. L. & Lions P. L. (1984) *Some properties of viscosity solutions of Hamilton-Jacobi equations*, Trans. Amer. Math. Soc., **282**, 487–502.
- [34] Chrysoschoos I. & Vinter R. B. (2003) *Optimal control problems on manifolds: a dynamic programming approach*, J. Math. Anal. Appl., **287**, 118–140.
- [35] Dal Maso G. & Frankowska H. (2003) *Autonomous integral functionals with discontinuous nonconvex integrands: Lipschitz regularity of minimizers, DuBois-Reymond necessary conditions, and Hamilton-Jacobi equations*, Appl. Math. Optim., **48**, 39–66.
- [36] Da Prato G. & Frankowska H. (2004) *Invariance of stochastic control systems with deterministic arguments*, J. Diff. Eqs., **200**, 18–52.
- [37] Delfour M. & Zolesio J.-P. (1994) *Shape analysis via oriented distance functions*, J. Func. Analysis, **123**, 129–201.
- [38] Delfour M. & Zolesio J.-P. (2004) *Oriented distance function and its evolution equation for initial sets with thin boundary*, SIAM J. Control Optim., **42**, 2286–2304.

- [39] Dikusar V. V. & Milyutin A. A. (1989) QUALITATIVE AND NUMERICAL METHODS IN THE MAXIMUM PRINCIPLE, Nauka.
- [40] Dontchev A. L. & Hager W. W. (1993) *Lipschitzian stability in nonlinear control and optimization*, SIAM J. Control Optim., **31**, 569–603.
- [41] Dontchev A. L. & Hager W. W. (1998) *A new approach to Lipschitz continuity in state constrained optimal control*, Systems and Control Letters, **35**, 137–143.
- [42] Dontchev A. L. & Rockafellar R. T. (2009) IMPLICIT FUNCTIONS AND SOLUTION MAPPINGS, Springer Mathematics Monographs, Springer, Dodrecht.
- [43] Dubovitskii A. Y. & Milyutin A. A. (1965) *Extremal problems with constraints*, USSR Comput. Math. and Math. Physics, **5**, 1–80.
- [44] Evans L. C. (1998) PARTIAL DIFFERENTIAL EQUATIONS, American Mathematical Society, Providence.
- [45] Filippov A. F. (1967) *Classical solutions of differential equations with multivalued right-hand side*, SIAM J. Control Optim., **5**, 609–621.
- [46] Fleming W. H. & Soner H. M. (2005) CONTROLLED MARKOV PROCESSES AND VISCOSITY SOLUTIONS, 2nd edition, Springer-Verlag.
- [47] Forcellini F. & Rampazzo F. (1999) *On non-convex differential inclusions whose state is constrained in the closure of an open set*, J. Differential Integral Equations, **12**, 471–497.
- [48] Frankowska H. (1987) *The maximum principle for an optimal solution to a differential inclusion with end points constraints*, SIAM J. Control Optim., **25**, 145–157.
- [49] Frankowska H. (1990) *Some inverse mapping theorems*, Annales de l’I.H.P. - Nonlinear Analysis, **7**, 183–234.
- [50] Frankowska H. (1991) *Lower semicontinuous solutions to Hamilton-Jacobi-Bellman equations*, Proceedings of 30th CDC IEEE Conference, Brighton, December 11–13.
- [51] Frankowska H. (1993) *Lower semicontinuous solutions of Hamilton-Jacobi-Bellman equation*, SIAM J. Control Optim., **31**, 257–272.
- [52] Frankowska H. (2005) *Optimal synthesis via superdifferentials of value function*, Control and Cybernetics, **34**, 787–803.
- [53] Frankowska H. (2006) *Regularity of minimizers and of adjoint states for optimal control problems under state constraints*, J. Convex Analysis, **13**, 299–328.
- [54] Frankowska H. & Marchini E. (2006) *Existence and Lipschitzianity of optimal trajectories for the Bolza optimal control problem*, Calculus of Variations and PDE, **27**, 467–492.
- [55] Frankowska H. & Ochal A. (2005) *On singularities of value function for Bolza optimal control problem*, JMAA, **306**, 714–729.
- [56] Frankowska H. & Plaskacz S. (1996) *A measurable upper semicontinuous viability theorem for tubes*, J. of Nonlinear Analysis, TMA, **26**, 565–582.
- [57] Frankowska H. & Plaskacz S. (2000) *Semicontinuous solutions of Hamilton-Jacobi-Bellman equations with degenerate state constraints*, J. Math. Anal. Appl., **251**, 818–838.

- [58] Frankowska H., Plaskacz S. & Rzezuchowski T. (1995) *Measurable viability theorems and Hamilton-Jacobi-Bellman equation*, J. Diff. Eqs., **116**, 265–305.
- [59] Galbraith G. & Vinter R. B. (2003) *Lipschitz continuity of optimal controls for state constrained problems*, SIAM J. Control Optim., **42**, 1727–1744.
- [60] Gamkrelidze R. V. (1960) *Optimal processes with bounded phase coordinates*, Izv. Akad. Nauk, USSR, Sec. Mat., **24**, 315–356.
- [61] Gavriel C. & Vinter R. B. (2010) *Regularity of minimizers for second order variational problems*, Preprint.
- [62] Hager W. W. (1979) *Lipschitz continuity for constrained processes*, SIAM J. Control Optim., **17**, 321–338.
- [63] Hartl R. F., Sethi S. P. & Vickson R. G. (1995) *A survey of the maximum principles for optimal control problems with state constraints*, SIAM Review, **37**, 181–218.
- [64] Ioffe A. D. (2000) *Metric regularity and subdifferential calculus*, Math. Surveys, **55**, 501–558.
- [65] Kuratowski K. (1966) *TOPOLOGY*, Academic Press/PWN.
- [66] Loewen P. & Rockafellar R. T. (1991) *The adjoint arc in nonsmooth optimization*, Trans. Amer. Math. Soc., **325**, 39–72.
- [67] Malanowski K. M. (1978) *On the regularity of solutions to optimal control problems for systems linear with respect to control variable*, Arch. Auto i Telemekh., **23**, 227–241.
- [68] Malanowski K. M. (2003) *On normality of Lagrange multipliers for state constrained optimal control problems*, Optimization, **52**, 75–91.
- [69] Malanowski K. M. (2007) *Sufficient optimality conditions in stability analysis for state-constrained optimal control*, Appl. Math. Optim., **55**, 255–271.
- [70] Malanowski K. M. (2007) *Stability analysis for nonlinear optimal control problems subject to state constraints*, SIAM J. Optim., **18**, 926–945.
- [71] Milyutin A. A. (2000) *On a certain family of optimal control problems with phase constraint*, J. Math. Sc., **100**, 2564–2571.
- [72] Pesch H. J. & Plail M. (2009) *The maximum principle of optimal control: a history of ingenious ideas and missed opportunities*, Control and Cybernetics, **38**.
- [73] Rampazzo F. & Vinter R. B. (1999) *A theorem on existence of neighbouring trajectories satisfying a state constraint, with applications to optimal control*, IMA J. Math. Control Inform., **16**, 335–351.
- [74] Rockafellar R. T. (1971) *Integrals which are convex functionals*, Pacific J. Math., **39**, 439–469.
- [75] Rockafellar R. T. & Wets R. J.-B. (1997) *VARIATIONAL ANALYSIS*, Springer-Verlag, Berlin.
- [76] Sarychev A. & Torres D. F. M. (2000) *Lipschitzian regularity of the minimizers for optimal control problems with control-affine dynamics*, Appl. Math. Optim., **41**, 237–254.

- [77] Schättler H. (2006) *Local fields of extremals for optimal control problems with state constraints of relative degree 1*, J. Dyn. Control Syst., **12**, 563–599.
- [78] Shvartsman I. A. & Vinter R. B. (2006) *Regularity properties of optimal controls for problems with time-varying state and control constraints*, Nonlinear Analysis TMA, **65**, 448–474.
- [79] Soner H. M. (1986) *Optimal control with state-space constraints*, SIAM J. Control Optim., **24**, 552–561.
- [80] Subbotin A. I. (1995) GENERALIZED SOLUTIONS OF FIRST ORDER PDES: THE DYNAMICAL OPTIMIZATION PERSPECTIVE, Birkhäuser, Boston, Basel, Berlin.
- [81] Subbotina N. N. (1989) *The maximum principle and the superdifferential of the value function*, Problems Control Inform. Theory, **18**, 151–160.
- [82] Vinter R. B. (2000) OPTIMAL CONTROL, Birkhäuser, Boston, Basel, Berlin.

Submodular Functions: Optimization and Approximation

Satoru Iwata*

Abstract

Submodular functions are discrete analogue of convex functions, arising in various fields of applied mathematics including game theory, information theory, and queueing theory. This survey aims at providing an overview on fundamental properties of submodular functions and recent algorithmic developments of their optimization and approximation.

For submodular function minimization, the ellipsoid method had long been the only polynomial algorithm until combinatorial strongly polynomial algorithms appeared a decade ago. On the other hand, for submodular function maximization, which is NP-hard and known to refuse any polynomial algorithms, constant factor approximation algorithms have been developed with applications to combinatorial auction, machine learning, and social networks. In addition, an efficient method has been developed for approximating submodular functions everywhere, which leads to a generic framework of designing approximation algorithms for combinatorial optimization problems with submodular costs. In some specific cases, however, one can devise better approximation algorithms.

Mathematics Subject Classification (2010). Primary 90C27; Secondary 68W25.

Keywords. Submodular functions, discrete optimization, approximation algorithms.

1. Introduction

Let V be a finite set. A set function $f : 2^V \rightarrow \mathbf{R}$ is said to be submodular if it satisfies

$$f(X) + f(Y) \geq f(X \cup Y) + f(X \cap Y), \quad \forall X, Y \subseteq V.$$

It is called monotone if $X \subseteq Y$ implies $f(X) \leq f(Y)$. Throughout this paper,

*Research Institute for Mathematical Sciences, Kyoto University, Kyoto 606-8502, Japan.
E-mail: iwata@kurims.kyoto-u.ac.jp.

we assume that a submodular function f satisfies $f(\emptyset) = 0$ and an evaluation oracle that computes the function value for each input in time EO is available.

Submodular functions arise in combinatorial optimization and various other fields of applied mathematics such as game theory [75], information theory [23] and queueing theory [12, 74]. Examples include the cut capacity functions of networks, the rank functions of matroids, and the entropy functions of multiple information sources.

Submodular functions are discrete analogue of convex functions. This analogy was exhibited by the discrete separation theorem of Frank [19] and the Fenchel-type duality theorem of Fujishige [25]. A more direct connection was established by Lovász [53], who clarified that the submodularity of a set function can be characterized by the convexity of a continuous function obtained by extending the set function in an appropriate manner. This observation together with valuated matroids invented by Dress and Wenzel [8] motivated Murota [58, 59, 60] to develop theory of discrete convex analysis.

Most efficiently solvable combinatorial optimization problems are related to submodular functions. The maximum flow and minimum spanning tree problems are two fundamental examples. The maximum-flow minimum-cut theorem says that the maximum flow value is equal to the minimum cut capacity. Since the cut capacity function is submodular, the maximum flow problem can be viewed as a special case of submodular function minimization. The greedy algorithm for the minimum spanning tree problem is extended and understood in the context of finding a minimum weight base in a matroid. The validity of this matroid greedy algorithm comes from the submodularity of the rank function.

The convex hull of the characteristic vectors of bases in a matroid is described by a set of linear inequalities that reflects the rank function. Although this polyhedral description requires an exponential number of inequality constraints, the greedy algorithm efficiently solves the linear optimization problem over the polytope. The greedy algorithm is further extended to solve a linear optimization problem over polyhedra associated with general submodular functions.

Grötschel, Lovász, and Schrijver [34] established a general principle that the optimization and separation problems are polynomially equivalent via the ellipsoid method, which had been used to give the first polynomial algorithm for linear programming by Khachiyan [48]. As a consequence, submodular functions can be minimized by the ellipsoid method in polynomial time, provided that an oracle for evaluating the function value is available. A strongly polynomial version was also developed in [35]. In spite of its polynomial time complexity, however, the ellipsoid method is not so efficient in practice.

Combinatorial strongly polynomial algorithms have been developed for general submodular function minimization independently by Schrijver [72] and by Iwata, Fleischer, and Fujishige [44]. Both of these algorithms build on earlier works of Cunningham [5, 6]. Since then several improved algorithms have been

presented [16, 41, 42, 46, 67]. The current best strongly polynomial bound, due to Orlin [67], is $O(n^5EO + n^6)$, where n is the cardinality of the ground set V .

In contrast, submodular function maximization is known to be NP-hard. It is also known that there is no polynomial algorithm that can maximize general submodular functions. This negative result is valid independently of the usual assumption that $P \neq NP$. There are quite a few interesting approximation results for submodular function maximization. The first such result is due to Nemhauser, Wolsey, and Fisher [65], who presented a constant factor approximation algorithm for maximizing a monotone submodular function under a cardinality constraint. Replacing the cardinality constraint by a matroidal constraint turns out to admit an approximation algorithm with the same constant factor [4]. The algorithm, however, relies on more sophisticated techniques such as pipage rounding originated by Ageev and Sviridenko [1].

For maximizing general nonnegative submodular functions, Feige, Mirrokni, and Vondrák [14] have presented a deterministic $1/3$ -approximation algorithm and a randomized $2/5$ -approximation algorithm based on local search techniques. They have shown that there is no polynomial algorithm with approximation factor better than $1/2$. Extensions with matroidal and/or knapsack constraints are designed subsequently [51, 77].

It is a natural attempt to replace linear functions in combinatorial optimization problems with submodular functions to obtain a more general results applicable to a wide variety of problems. Perhaps, the most successful classical result of this type is the theory of submodular flows introduced by Edmonds and Giles [10]. The submodular flow problem is obtained from the minimum cost flow problem by replacing the linear function in the conservation law with a submodular function, or even more generally a crossing submodular function. This problem generalizes the minimum cost flow, matroid intersection, and some graph connectivity problems such as the shortest dijoin problem [18, 54]. Algorithmic techniques for network flow problems such as the scaling methods [11] and the push/relabel framework [33] have been extended to submodular flows [7, 28, 40].

Svitkina and Fleischer [76] have started a systematic study of approximation algorithms for submodular cost minimization problems. For submodular sparsest cut and submodular load balancing problems, they have developed randomized $O(\sqrt{n/\log n})$ -approximation algorithms and shown that this factor is in fact best possible. This is in contrast to the corresponding original problems, which admit logarithmic or constant factor approximation algorithms. A recent work of Goel, Karande, Tripathi, and Wang [29] provides matching lower and upper bounds for the approximability of submodular cost versions of efficiently solvable problems such as the shortest path, minimum spanning tree, and minimum weight perfect matching problems. These results demonstrate that submodular functions are so general that replacing a linear cost by

a submodular function often makes it much harder to solve the problem even approximately.

A generic method for design of approximation algorithms for combinatorial optimization problems with submodular costs is to approximate the cost function f by a well-behaved submodular function \tilde{f} that is computable in polynomial time. If $\tilde{f}(S) \leq f(S) \leq \alpha\tilde{f}(S)$ holds for every $S \subseteq V$ and there exists a μ -approximation algorithm for the problem with cost function replaced by \tilde{f} , then we are able to design an algorithm with approximation ratio $\alpha\mu$. Goemans, Harvey, Iwata, and Mirrokni [30] have presented nearly $O(\sqrt{n})$ lower and upper bounds on α for monotone submodular functions.

In some cases, however, combinatorial optimization problems with submodular costs may admit much better approximation results. For instance, 2-approximation algorithms have been devised for the submodular vertex cover problem [29, 45]. This is further extended to the set cover problem with submodular cost functions. The submodular partition problem that generalizes the multi-cut problem admits constant factor approximation algorithms [66, 80].

These are the aspects of submodular functions this paper will survey. It is far from being comprehensive. The readers are referred to related chapters of Fujishige [27], Korte and Vygen [49], and Schrijver [73] for general background on submodular functions in combinatorial optimization. See also Frank [20, 21, 22] for applications of submodular functions in graph theory.

Throughout this paper, let \mathbf{R}^V denote the set of all the real valued functions $x : V \rightarrow \mathbf{R}$, which forms a linear space of dimension $n = |V|$. We identify a vector $x \in \mathbf{R}^V$ with a set function defined by $x(Y) = \sum_{v \in Y} x(v)$. For a subset $S \subseteq V$, we denote by χ_S the characteristic vector in \mathbf{R}^V , i.e., $\chi_S(v) = 1$ if $v \in S$, and $\chi_S(v) = 0$ otherwise.

2. Examples of Submodular Functions

In this section, we describe four examples of submodular functions.

Cut Capacity Functions. Let $G = (V, A)$ be a (directed or undirected) graph with an nonnegative arc capacity function $c : A \rightarrow \mathbf{R}_+$. For each vertex subset $X \subseteq V$, we denote by $\kappa(X)$ the sum of the arc capacities $c(a)$ of all the arcs a connecting X to $V \setminus X$. Then κ forms a submodular function on the subsets of V . If G is undirected, then κ is symmetric, i.e., $\kappa(X) = \kappa(V \setminus X)$ holds for every $X \subseteq V$.

Set Cover Function. Let $G = (U, V; E)$ be a bipartite graph with bipartition of the vertex set into U and V . For each $X \subseteq V$, let $\Gamma(X)$ denote the set of vertices in U adjacent to X . Then the function γ defined by $\gamma(X) = |\Gamma(X)|$ forms a monotone submodular function. The theorems of König and Hall show

that the size $\tau(G)$ of a maximum matching in G is given by

$$\tau(G) = \min\{\gamma(X) - |X| \mid X \subseteq V\} + |V|.$$

Thus the maximum bipartite matching problem can be viewed as a special case of submodular function minimization.

The bipartite graph can be used to represent set covers. Let $\mathcal{S} = \{S_1, \dots, S_n\}$ be a collection of subsets of U indexed by $V = \{1, \dots, n\}$. Then we can construct a bipartite graph $G = (U, V; E)$ whose edge set E consists of pairs (i, j) such that $i \in S_j$. Then for each subset $X \subseteq V$, we have $\Gamma(X) = \bigcup\{S_j \mid j \in X\}$. In particular, $X \subseteq V$ is called a set cover if $\Gamma(X) = U$. The maximum coverage problem asks to find a subset $X \subseteq V$ of given cardinality k so that $\gamma(X)$ is maximized. This is a special case of maximizing a monotone submodular function under a cardinality constraint.

Matroid Rank Functions. The concept of matroids was introduced by Whitney [79] as a combinatorial abstraction of linear independence. Let V be a finite set and \mathcal{I} be a family of subsets of V . A pair (V, \mathcal{I}) is a matroid if it satisfies a certain system of axioms. A member in \mathcal{I} is called an independent set. The rank function ρ of the matroid is defined by $\rho(X) = \max\{|J| \mid J \subseteq X, J \in \mathcal{I}\}$. Then ρ is a monotone submodular function that satisfies $\rho(\emptyset) = 0$ and $\rho(X) \leq |X|$ for $X \subseteq V$. Conversely, such an integer valued set function defines a matroid by $\mathcal{I} = \{J \mid J \subseteq V, \rho(J) = |J|\}$.

The convex hull of the characteristic vectors of the independent sets in \mathbf{R}^V coincides with

$$\text{MP}(\rho) = \{z \mid z \in \mathbf{R}_+^V, \forall X \subseteq V, z(X) \leq \rho(X)\},$$

which is called the matroid polyhedron. Testing if a given vector $z \in \mathbf{R}_+^V$ is in $\text{MP}(\rho)$ can be reduced to minimizing the submodular function $f(X) = \rho(X) - z(X)$. Cunningham [5] presented a combinatorial strongly polynomial algorithm for this special type of submodular function minimization.

Entropy Functions. Let V be a set of discrete memoryless information sources (random variables). For each nonempty subset X of V , let $h(X)$ denote the Shannon entropy of the corresponding joint distribution. In addition, we assign $h(\emptyset) = 0$. Then the set function h is a submodular function, which follows from the nonnegativity of conditional mutual information.

Let K be a positive definite symmetric matrix whose row/column set is indexed by V . For each $X \subseteq V$, let $K[X]$ denote the principal submatrix of K indexed by X . The set function f defined by $f(\emptyset) = 0$ and $f(X) = \log \det K[X]$ for nonempty X is a submodular function. The submodularity of this function f , known as Ky Fan's inequality, is a refinement of Hadamard's inequality. It can be interpreted as the submodularity of the entropy function of a multivariate normal distribution with covariance matrix K .

3. Associated Polyhedra and Discrete Convexity

For a submodular function f on the subsets of V , we consider the submodular polyhedron $P(f)$ and the base polyhedron $B(f)$ defined by

$$\begin{aligned} P(f) &= \{x \mid x \in \mathbf{R}^V, \forall Y \subseteq V, x(Y) \leq f(Y)\}, \\ B(f) &= \{x \mid x \in P(f), x(V) = f(V)\}. \end{aligned}$$

A vector in $B(f)$ is called a base. In particular, an extreme point of $B(f)$ is called an extreme base. The base polyhedron $B(f)$ is the set of maximal vectors in $P(f)$.

An extreme base can be computed by the greedy algorithm of Edmonds [9] and Shapley [75] as follows.

Let $L = (v_1, \dots, v_n)$ be a linear ordering of V . For any $v_j \in V$, we denote $L(v_j) = \{v_1, \dots, v_j\}$. The greedy algorithm with respect to L generates an extreme base $y \in B(f)$ by

$$y(u) := f(L(u)) - f(L(u) \setminus \{u\}). \quad (1)$$

Conversely, any extreme base can be obtained in this way with an appropriate linear ordering.

Given a nonnegative vector $p \in \mathbf{R}_+^V$, consider a linear ordering $L = (v_1, \dots, v_n)$ such that $p(v_1) \geq p(v_2) \geq \dots \geq p(v_n)$. The greedy algorithm with respect to L yields an optimal solution to the problem of maximizing the inner product $\langle p, x \rangle = \sum_{v \in V} p(v)x(v)$ in $P(f)$.

Let $p_1 > p_2 > \dots > p_k$ be the distinct values of p . For $j = 1, \dots, k$, we denote $V_j = \{v \mid p(v) \geq p_j\}$. We now define $\hat{f}(p)$ by

$$\hat{f}(p) = \sum_{j=1}^k (p_j - p_{j+1}) f(V_j),$$

where $p_{k+1} = 0$. Then the function \hat{f} satisfies

$$\hat{f}(p) = \max\{\langle p, x \rangle \mid x \in P(f)\}, \quad (2)$$

which follows from the validity of the greedy algorithm.

Note that the above definition of \hat{f} is free from the submodularity of f . For a set function f in general, we define \hat{f} in the same way. Then $\hat{f}(\chi_S) = f(S)$ holds for any $S \subseteq V$. Hence we may regard \hat{f} as an extension of f .

The restriction of \hat{f} to the hypercube $[0, 1]^V$ can be interpreted as follows. A linear ordering L corresponds to the simplex whose extreme points are given by the characteristic vectors of $L(v)$ for $v \in V$ and the empty set. Since there are $n!$ linear orderings of V , the hypercube $[0, 1]^V$ can be partitioned into $n!$ congruent simplices obtained by this way. Determine the function values of \hat{f}

in each simplex by the linear interpolation of the values at the extreme points. The resulting function \hat{f} is a continuous function on the hypercube.

The following theorem exhibits a close connection between submodularity and convexity. The proof relies on the validity of the greedy algorithm.

Theorem 3.1 (Lovász [53]). *A set function f is submodular if and only if \hat{f} is convex.*

4. Submodular Function Minimization

This section is devoted to combinatorial algorithms for minimizing submodular functions. More detailed descriptions and comparisons are given in [43, 55].

For any vector $x \in \mathbf{R}^V$, we denote $x^-(v) := \min\{x(v), 0\}$. The following min-max theorem plays a crucial role in submodular function minimization.

Theorem 4.1. *For a submodular function f , we have*

$$\min_{Y \subseteq V} \hat{f}(Y) = \max\{x^-(V) \mid x \in B(f)\}.$$

Moreover, if f is an integer valued function, then the maximum in the right-hand side is attained by an integer vector x .

This theorem is immediate from the vector reduction theorem on polymatroids due to Edmonds [9]. Note that $x^-(V) \leq x(Y) \leq f(Y)$ holds for any pair of $x \in B(f)$ and $Y \subseteq V$. The theorem shows that these inequalities are tight for appropriately chosen x and Y .

Theorem 4.1 seems to provide a good characterization of the minimum value of f . In fact, if we have a pair of $W \subseteq V$ and $x \in B(f)$ with $f(W) = x^-(V)$, then it follows from Theorem 4.1 that W attains the minimum value of f . This suggests a natural way to find the minimum by moving $x \in B(f)$ so that $x^-(V)$ increases. However, it is not easy to verify that the vector x in our hand stays in $B(f)$. A direct way to check this by the definition requires an exponential number of steps. On the other hand, an extreme base y of $B(f)$ can be verified by a linear ordering of V generating y . According to Caratheodory's theorem, an arbitrary point in a bounded polyhedron can be expressed as a convex combination of its extreme points. Keeping $x \in B(f)$ as a convex combination $x = \sum_{i \in I} \lambda_i y_i$ of extreme bases y_i , we are able to verify $x \in B(f)$ efficiently, provided that I is not too large. A base $x \in B(f)$ expressed by this way provides a compact certificate of $f(W)$ being the minimum value if $x^-(V) = f(W)$ holds.

This approach was introduced by Cunningham [5] in the separation problem for matroid polyhedra. Bixby, Cunningham, and Topkis [3] employed this approach to develop a combinatorial algorithm for minimizing a submodular function by a finite number of steps. Furthermore, Cunningham [6] improved this algorithm to the first combinatorial pseudopolynomial algorithm for computing the minimum value of an integer valued submodular function. In general,

a pseudopolynomial algorithm runs in time polynomial in the number of inputs and the maximum absolute value of the inputs. The running time bound of Cunningham's algorithm is $O(n^6 EOM \log nM)$, where M is the maximum absolute value of f .

Combinatorial strongly polynomial algorithms have been developed by Iwata, Fleischer, and Fujishige (IFF) [44] and by Schrijver [72]. Both of these algorithms build on works of Cunningham [5, 6].

The IFF algorithm employs a scaling scheme developed for the submodular flow problem [17, 40]. In contrast, Schrijver [72] directly achieves a strongly polynomial bound by introducing a novel subroutine in the framework of lexicographic augmentation. Subsequently, Fleischer and Iwata [15, 16] have described a push/relabel algorithm using Schrijver's subroutine to improve the running time bound. It has turned out however that Schrijver's algorithm has the same running time bound [78]. Combining the scaling scheme with the push/relabel technique yields a faster combinatorial algorithm [42], which currently achieves the best weakly polynomial running time bound $O((n^4 EO + n^5) \log M)$ for submodular function minimization. The current best strongly polynomial bound is $O(n^5 EO + n^6)$ due to Orlin [67]. This algorithm adopts a modified push/relabel technique and utilizes a system of linear equations whose coefficient matrix is an M-matrix, whereas Schrijver's subroutine solves a system of linear equations with a triangular coefficient matrix. Combining the techniques of [42] and [67] together with the use of a quadratic potential function, Iwata and Orlin [46] have presented a combinatorial approach that nearly matches the best weakly and strongly polynomial bounds.

All of these combinatorial algorithms perform multiplications and divisions, although the problem of submodular function minimization does not involve such arithmetic operations. Schrijver [72] has asked if one can minimize a submodular function in strongly polynomial time using only additions, subtractions, comparisons, and the oracle calls for function values. It turns out that the IFF strongly polynomial algorithm can be converted to such a fully combinatorial algorithm [41]. The subsequent algorithms developed in [42, 46] can also be implemented in a fully combinatorial manner. The current best running time bound of a fully combinatorial algorithm is $O((n^7 EO + n^8) \log n)$ given in [46]. The existence of a fully combinatorial algorithm is used by Nagano [64] to show strong polynomiality of the line search problem in submodular polyhedra with the aid of the parametric search technique of Megiddo [56, 57].

A particularly nice feature of submodular function minimization is that the set of all the optimal solutions can be expressed in a compact manner. If X and Y both minimize a submodular function f , then it follows from the submodularity that both $X \cap Y$ and $X \cup Y$ minimize f as well. Thus the set of all the minimizers of f forms a distributive lattice. In particular, there is a unique minimal/maximal minimizer of f . According to Birkhoff's representation theorem, any distributive lattice can be expressed as the set of ideals of a poset.

To be more concrete, one can obtain the poset that represents the set of all the minimizers as follows.

Let $x = \sum_{i \in I} \lambda_i y_i$ be a base that attains the maximum in the right-hand side of the min-max relation in Theorem 4.1, where y_i is an extreme base and the coefficient λ_i in the convex combination is positive for each $i \in I$. The strongly polynomial algorithms presented in [16, 67, 72] produce such a base x represented as a convex combination of extreme bases without any extra efforts. For each extreme base y_i , one can construct a poset associated with y_i in $O(n^2 \text{EO} + n^3)$ time [3]. Take the superposition of those posets to obtain a digraph $G_x = (V, E_x)$. Then a subset $Y \subseteq V$ is a minimizer of f if and only if $x(Y) = x^-(V)$ and there is no arc leaving Y in G_x . In particular, the set of vertices reachable from $N = \{v \mid x(v) < 0\}$ in G_x is the unique minimal minimizer of f . The unique maximal minimizer is the set of vertices from which $P = \{v \mid x(v) > 0\}$ is not reachable in G_x .

The minimum-norm base of f is a base $x \in \mathbf{B}(f)$ that minimizes $\sum_{v \in V} x(v)^2$. The following theorem suggests that finding the minimum-norm base provides another possible approach to submodular function minimization.

Theorem 4.2 (Fujishige [24, 26]). *Let x^* be the minimum-norm base of f . Then $S = \{v \mid x^*(v) < 0\}$ is the unique minimal minimizer, and $T = \{v \mid x^*(v) \leq 0\}$ is the unique maximal minimizer.*

Fujishige [27, §7.1 (a)] describes an algorithm for finding the minimum-norm base, which works well in practice. The complexity analysis of this algorithm remains open.

Submodular function minimization finds a variety of applications in evacuation planning [38], wireless communication [31], and computational group theory [71]. A certain type of submodular functions that arise in multiclass queueing systems allow a much more efficient algorithm based on computational geometry [39].

Grötschel, Lovász, and Schrijver [35] have discussed minimizing submodular functions among odd subsets. Goemans and Ramakrishnan [32] have presented extensions, including an efficient method to find the second smallest value of a submodular function.

5. Symmetric Submodular Function Minimization

A set function f is said to be symmetric if $f(X) = f(V \setminus X)$ holds for every $X \subseteq V$. A symmetric submodular function f satisfies $f(X) \geq f(\emptyset)$ for any $X \subseteq V$. Hence it is trivial to compute the minimum value of f . Symmetric submodular function minimization is the problem of finding the minimum value of $f(X)$ among proper nonempty subsets X .

Examples of symmetric submodular functions are the cut capacity function of an undirected network and the connectivity function of a matroid. Symmetric submodular function minimization for the cut capacity function corresponds to the minimum cut problem. Based on the max-flow min-cut theorem one can efficiently find a minimum cut in a graph via maximum flow algorithms. Instead, Nagamochi and Ibaraki [62] introduced a novel algorithm that finds a minimum cut directly and more efficiently. The algorithm exploits the maximum-adjacency ordering of vertices in each iteration. As a generalization of this minimum cut algorithm, Queyranne [69] presented a fully combinatorial strongly polynomial algorithm for symmetric submodular function minimization.

Queyranne's algorithm in fact deals with an arbitrary submodular function f to find a proper nonempty subset X that minimizes $f(X) + f(V \setminus X)$. It adopts a novel procedure **Pendant-Pair** that provides an ordering of V as follows. First, select an arbitrary element as v_1 . Subsequently, for $j = 1$ to $n - 1$, given $W_j = \{v_1, \dots, v_j\}$, the procedure selects an element $u \in V \setminus W_j$ that minimizes $f(W_j \cup \{u\}) - f(\{u\})$ as v_{j+1} . The pair (v_{n-1}, v_n) is called a pendant pair.

Theorem 5.1 (Queyranne [69]). *For any subset X that separates the pendant pair v_{n-1} and v_n , we have $f(X) + f(V \setminus X) \geq f(\{v_n\}) + f(V \setminus \{v_n\})$. \square*

Theorem 5.1 suggests a way to find the minimum value of $f(X) + f(V \setminus X)$. Let (u, v) be the pendant pair obtained by applying **Pendant-Pair** to f . Consider a submodular function f' on the ground set $V' := V \setminus \{v\}$ defined by

$$f'(X) = \begin{cases} f(X) & (u \notin X) \\ f(X \cup \{v\}) & (u \in X). \end{cases}$$

Then the minimum value of $f(X) + f(V \setminus X)$ is equal to $f(\{v\}) + f(V \setminus \{v\})$ or to the minimum value of $f'(X) + f'(V' \setminus X)$, which can be computed recursively. Thus we obtain an algorithm to find the minimum value of $f(X) + f(V \setminus X)$ by applying **Pendant-Pair** $O(n)$ times. Since one application of **Pendant-Pair** takes $O(n^2\text{EO})$ time, the total running time bound is $O(n^3\text{EO})$.

This algorithm is further generalized in two different directions by Nagamochi and Ibaraki [63] and by Rizzi [70].

Nagamochi [61] has presented another efficient minimum cut algorithm as well as its generalization to symmetric submodular function minimization. The minimum cut algorithm uses the minimum degree ordering of vertices instead of the maximum adjacency ordering. Nagamochi's algorithm for symmetric submodular functions works as follows.

Let f be a symmetric submodular function. The algorithm adopts a procedure **Flat-Pair** that provides an ordering v_1, v_2, \dots, v_n of V as follows. For $j = 0$ to $n - 1$, given $W_j = \{v_1, \dots, v_j\}$, where $W_0 = \emptyset$, the procedure selects an element $u \in V \setminus W_j$ that minimizes $f(W_j \cup \{u\}) + f(\{u\})$ as v_{j+1} . The pair (v_{n-1}, v_n) is called a flat pair.

Theorem 5.2 (Nagamochi [61]). *For any subset X that separates the flat pair v_{n-1} and v_n , we have $f(X) \geq \min\{f(\{v\}) \mid v \in X\}$.*

A subset $X \subseteq V$ is called an extreme set if it satisfies $f(Z) > f(X)$ for every proper nonempty set $Z \subsetneq X$. Then the family of extreme sets forms a laminar, i.e., for a pair of extreme sets X and Y , we have $X \subseteq Y$, $X \supseteq Y$, or $X \cap Y = \emptyset$. Theorem 5.2 suggests a way to enumerate all the extreme sets. All the singletons are extreme sets. Put $\theta(u) = f(\{u\})$ for each $u \in V$. Let (u, v) be the flat pair obtained by applying Flat-Pair to f . Then any set X that separates u and v is not an extreme set unless X is a singleton. Shrink the flat pair (u, v) into a new element w . The value of $\theta(w)$ is given by the minimum of $\theta(u)$, $\theta(v)$, and $f(\{u, v\})$. If $\theta(w) < \min\{\theta(u), \theta(v)\}$, then w corresponds to an extreme set. We repeat this process until V becomes a singleton. Thus we obtain all the extreme sets, among which there is a minimizer of f .

Since one application of Flat-Pair takes $O(n^2\text{EO})$ time, the total running time is $O(n^3\text{EO})$. This is the same as that of Queyranne's algorithm, whereas Nagamochi's algorithm provides not only a minimizer but also all the extreme sets.

6. Submodular Function Maximization

Submodular function maximization has also been studied extensively. It finds interesting applications in marketing strategy through social networks [36, 47] and selecting features or sensors for observation [50].

The first guaranteed approximation algorithm was developed by Nemhauser, Wolsey, and Fisher [65] for maximizing a monotone submodular function under a cardinality constraint.

Let $f : 2^V \rightarrow \mathbf{R}$ be a monotone submodular function. The algorithm determines a linear ordering (v_1, \dots, v_n) of V as follows. For $j = 0$ to $n - 1$, given $W_j = \{v_1, \dots, v_j\}$, where $W_0 = \emptyset$, select $u \in V \setminus W_j$ that maximizes $f(W_j \cup \{u\})$ as v_{j+1} . The linear ordering thus obtained has the following property.

Theorem 6.1 (Nemhauser, Wolsey, and Fisher [65]). *For any subset $S \subseteq V$ of cardinality k , we have $f(W_k) \geq (1 - 1/e)f(S)$.*

Theorem 6.1 suggests that W_k serves as an approximate solution with ratio at least $1 - 1/e$. On the other hand, Feige [13] has shown that there is no polynomial algorithm with approximation factor better than $1 - 1/e$ for the maximum coverage problem, assuming $\text{P} \neq \text{NP}$.

Maximizing non-monotone submodular function is also known to be NP-hard, as it contains the max cut problem. Constant factor approximation algorithms for this fundamental setting have been presented only recently by Feige, Mirrokni, and Vondrák [14]. More specifically, they provide a deterministic local search algorithm with an approximation ratio $1/3$, and its randomized version with approximation ratio $2/5$.

The deterministic local search algorithm works as follows. The algorithm starts with a singleton $S = \{v\}$ with maximum value. Then the algorithm repeatedly adds an element to S or deletes an element from S if this increases the value of f by more than a factor of $1 + \epsilon/n^2$. The algorithm terminates if no such element is left, and returns the maximum of $f(S)$ and $f(V \setminus S)$. Feige, Mirrokni, and Vondrák [14] showed that this local search algorithm terminates after $O(\frac{1}{\epsilon} n^3 \log n)$ function evaluation and the obtained solution achieves the value at least $(\frac{1}{3} - \frac{\epsilon}{n})$ times the optimal value. Their randomized version improves this ratio to $\frac{2}{5} - \frac{\epsilon}{n}$.

Submodular function maximization under a matroid constraint is of interest. Let $\mathbf{M} = (V, \mathcal{I})$ be a matroid with ground set V and independent set family \mathcal{I} . Suppose that f is a monotone submodular function over the subsets of V . Then the problem is to find an independent subset $I \in \mathcal{I}$ that maximizes the value $f(I)$. This maximization problem generalizes the one with a cardinality constraint. Calinescu, Chekuri, Pál, and Vondrák [4] have devised a $(1 - 1/e)$ -approximation algorithm for this general problem.

The key concept in this algorithm is a multilinear extension \bar{f} of the submodular function f . For any point in $p \in [0, 1]^V$, consider a random set R_p that contains $v \in V$ with probability $p(v)$. Then $\bar{f}(p)$ is defined to be the expectation of $f(R_p)$, i.e.,

$$\bar{f}(p) = \mathbb{E}[f(R_p)] = \sum_{X \subseteq V} f(X) \prod_{v \in X} p(v) \prod_{u \in V \setminus X} (1 - p(u)).$$

Then the multilinear extension \bar{f} satisfies $\frac{\partial \bar{f}}{\partial p(v)} \geq 0$ for any $v \in V$ and $\frac{\partial^2 \bar{f}}{\partial p(u) \partial p(v)} \leq 0$ for any $u, v \in V$. Note that \bar{f} is not a concave function. It is concave in a nonnegative direction, and convex in the direction of $\chi_u - \chi_v$ for $u, v \in V$.

The algorithm consists of two phases: continuous greedy algorithm and pipage rounding. The continuous greedy algorithm aims at finding a good approximate solution for maximizing \bar{f} in the matroid polyhedron $\text{MP}(\rho)$. The output y is shown to be a $(1 - 1/e)$ -approximate solution. Then the pipage rounding scheme finds a base B whose value is at least as large as $\bar{f}(y)$ in expectation. Thus the entire algorithm serves as a randomized $(1 - 1/e)$ -approximation algorithm for maximizing a monotone submodular function among the independent set.

A motivating example of this problem is the social welfare maximization in combinatorial auction. Suppose we are given a set U of m items and n players. Each player j has a utility function $f_j : 2^U \rightarrow \mathbf{R}$ that is submodular. Then the goal is to find a partition of U into disjoint subsets S_1, \dots, S_n so as to maximize the social welfare $\sum_{j=1}^n f_j(S_j)$. This can be formulated as maximizing a submodular function over a matroid as follows. Let U_1, \dots, U_n be disjoint copies of U and V denote their union. Each utility function f_j can be regarded as a set function on U_j . Consider a partition matroid on the ground set V , in which

a subset $X \subseteq V$ as independent if no distinct copies of the same item belong to X . This way the problem is to maximize the submodular function value $f(X) = \sum_{j=1}^n f_j(X \cap U_j)$ among the independent sets.

7. Submodular Function Approximation

This section is devoted to the problem of approximating submodular functions everywhere. Let f be a nonnegative submodular function given by an evaluation oracle. The goal is to construct a function \tilde{f} such that $\tilde{f}(S) \leq f(S) \leq \alpha \tilde{f}(S)$ holds for every $S \subseteq V$ by a polynomial number of oracle calls and arithmetic operations. The constructed approximate function \tilde{f} should be evaluated in polynomial time for any input. It is shown in [30, 76] that this requires $\alpha = \Omega(\sqrt{n/\log n})$. On the other hand, Goemans, Harvey, Iwata and Mirrokni [30] have developed algorithms with $\alpha = \sqrt{n+1}$ for matroid rank functions and $\alpha = O(\log n)$ for monotone submodular functions in general.

For a monotone submodular function f , let $Q(f)$ be a polyhedron defined by

$$Q(f) = \left\{ x \mid \sum_{v \in X} |x(v)| \leq f(X), \forall X \subseteq V \right\},$$

which is called the symmetrized polymatroid. Since $Q(f)$ is a centrally symmetric convex body, it follows from John’s theorem that the maximum volume ellipsoid E contained in $Q(f)$ satisfies $E \subseteq Q(f) \subseteq \sqrt{n}E$. Because of the symmetry of $Q(f)$, the maximum volume ellipsoid E must be axis-aligned. Hence E can be represented as

$$E = \left\{ x \mid \sum_{v \in V} \frac{x(v)^2}{d(v)} \leq 1 \right\},$$

where $d(v) > 0$ for each $v \in V$. If this ellipsoid were known, we could construct a submodular function \tilde{f} by

$$\tilde{f}(S) = \sqrt{\sum_{v \in S} d(v)}.$$

Then this would imply $\tilde{f}(S) \leq f(S) \leq \sqrt{n} \tilde{f}(S)$ for every $S \subseteq V$. Instead, the algorithm approximately finds the maximum volume ellipsoid. The algorithm keeps an axis-aligned ellipsoid $E \subseteq Q(f)$ and repeatedly checks if it satisfies $Q(f) \subseteq \sqrt{n+1} E$ approximately with an approximation ratio β . If the answer is ‘yes,’ then the algorithm adopts the current axis-aligned ellipsoid. Otherwise, it uses the certificate to update the axis-aligned ellipsoid so that the volume will increase by at least a factor of $1 + 4/n^2$. Starting with a certain initial ellipsoid, the algorithm terminates after $O(n^3 \log n)$ iterations, and the obtained ellipsoid E satisfies $E \subseteq Q(f) \subseteq \frac{\sqrt{n+1}}{\beta} E$.

The decision problem in each iteration is reduced to a separable convex quadratic maximization in the polymatroid, which has an optimal solution at an extreme point. In the special case of matroid rank functions, all the extreme points are 0-1 vectors. One can exploit this fact to solve the problem exactly, i.e., $\beta = 1$. Thus we obtain $\alpha = \sqrt{n+1}$ for matroid rank functions. For general monotone submodular functions, an $O(1/\log n)$ approximation algorithm is presented in [30], which leads to $\alpha = O(\sqrt{n \log n})$.

Svitkina and Fleischer [76] have introduced the submodular load balancing problem. Let f_1, \dots, f_m be monotone submodular functions on the ground set V . The goal is to find a partition of V into disjoint subsets V_1, \dots, V_m that minimizes $\max_j f_j(V_j)$. Svitkina and Fleischer [76] have shown that any algorithm with polynomial number of oracle calls could not achieve the approximation ratio $o(\sqrt{\frac{n}{\log n}})$. They have also presented a randomized approximation algorithm matching this lower bound.

Alternatively, one can use the approximate submodular functions $\tilde{f}_1, \dots, \tilde{f}_m$ to design a deterministic algorithm [30]. Suppose that each \tilde{f}_j is in the form of

$$\tilde{f}_j(S) = \sqrt{\sum_{v \in S} d_j(v)}.$$

Minimizing $\max_j \tilde{f}_j(V_j)$ means minimizing $\max_j \sum_{v \in V_j} d_j(v)$. This problem is equivalent to minimizing the makespan in nonpreemptive scheduling on parallel machines, for which Lenstra, Shmoys, and Tardos [52] have given a deterministic 2-approximation algorithm. Adopting the output of this algorithm, we obtain a $\sqrt{2}\alpha$ -approximate solution to the original problem. Thus we have a deterministic $O(\sqrt{n \log n})$ -approximation algorithm for the submodular load balancing problem.

8. Submodular Cost Set Cover

Let U be a finite set of cardinality m and $\mathcal{S} = \{S_1, \dots, S_n\}$ be a collection of its subsets indexed by $N = \{1, \dots, n\}$. We say that a subset $X \subseteq N$ is a set cover if $U = \bigcup \{S_i \mid i \in X\}$. Given a nonnegative cost function $c : N \rightarrow \mathbf{R}_+$, the set cover problem asks for finding a set cover $X \subseteq N$ that minimizes the cost $c(X) = \sum_{i \in X} c(i)$. This problem is known to be solved approximately in polynomial time within a factor of $O(\ln m)$ or the maximum frequency $\eta = \max_{u \in U} |N_u|$, where $N_u = \{i \mid u \in S_i\}$. Given a nonnegative submodular function $f : 2^N \rightarrow \mathbf{R}_+$, the submodular cost set cover problem asks for finding a set cover $X \subseteq N$ that minimizes $f(X)$.

The $O(\log m)$ -approximation for the set cover problem is achieved by a greedy algorithm. However, this approach does not extend to the submodular cost set cover problem. In fact, it is shown in [45] that no polynomial algorithm

can achieve an approximation factor of $o(m/\log^2 m)$ even for the special case of the submodular edge cover problem.

In contrast, the η -approximation results have been extended to the submodular cost set cover problem. Consider the following convex programming relaxation of the submodular cost set cover problem:

$$\begin{aligned}
 \text{(SCP)} \quad & \text{Minimize} \quad \widehat{f}(x) \\
 & \text{subject to} \quad \sum_{i \in N_u} x(i) \geq 1 \quad (u \in U) \\
 & \quad \quad \quad x(i) \geq 0 \quad (i \in N).
 \end{aligned}$$

This problem can be solved in polynomial time with the aid of the ellipsoid method.

Let $x^* \in \mathbf{R}^N$ be an optimal solution to (SCP). Then $T = \{i \mid x^*(i) \geq 1/\eta\}$ is a set cover. Let T° denote the unique minimal minimizer of f among all the subsets Z with $T \subseteq Z \subseteq N$. Note that T° can be obtained by executing submodular function minimization. Then T° is an η -approximate solution for the submodular cost set cover problem. This extends the rounding algorithm due to Hochbaum [37].

Alternatively, one can extend the primal-dual approximation algorithm due to Bar-Yehuda and Even [2]. The dual problem to (SCP) is given as follows.

$$\begin{aligned}
 \text{(DCP)} \quad & \text{Maximize} \quad \sum_{u \in U} y(u) \\
 & \text{subject to} \quad z \in P(f), \\
 & \quad \quad \quad \sum_{u \in S_i} y(u) = z(i) \quad (i \in N), \\
 & \quad \quad \quad y(u) \geq 0 \quad (u \in U).
 \end{aligned}$$

The primal-dual algorithm keeps a feasible solution (y, z) of (DCP) and a subset $T \subseteq N$ that is z -tight. The algorithm starts with $y := 0, z := 0$ and $T := \emptyset$. Since f is a nonnegative submodular function with $f(\emptyset) = 0$, this gives a feasible solution of (DCP) and we have $z(T) = f(T)$. While T is not a set cover, there must be an element $u \in U$ which is not covered by T . The algorithm augments $y(u)$ and $z(i)$ for $i \in N_u$ as much as possible without violating the constraints in (DCP). Then the algorithm updates T to be the unique maximal set with $z(T) = f(T)$. The algorithm iterates this procedure until T becomes a set cover. The number of iterations is at most n and the resulting T is an η -approximate solution to the submodular cost set cover problem.

A very special case with $\eta = 2$ can be formulated in terms of graphs. Let $G = (V, E)$ be a graph with vertex set V and edge set E . A vertex subset $X \subseteq V$ is called a vertex cover in G if every edge in E is incident to a vertex in X . Given a nonnegative submodular function $f : 2^V \rightarrow \mathbf{R}_+$, the submodular vertex cover problem asks for finding a vertex cover $X \subseteq V$ that minimizes the cost $f(X)$.

A natural approach to this problem is to consider the following relaxation problem:

$$\begin{aligned}
 \text{(CPR)} \quad & \text{Minimize} \quad \widehat{f}(x) \\
 & \text{subject to} \quad x(u) + x(v) \geq 1 \quad ((u, v) \in E) \\
 & \quad \quad \quad x(v) \geq 0 \quad (v \in V).
 \end{aligned}$$

This convex programming relaxation problem can be solved by the ellipsoid method in polynomial time. It is shown in [45] that the relaxation problem has a half-integral optimal solution. Moreover, such a half-integral optimal solution can be found by a combinatorial algorithm for minimizing a submodular function over a distributive lattice.

Let x^* a half-integral optimal solution to (CRP). Then $X^* = \{v \mid x^*(v) \leq \frac{1}{2}\}$ is a vertex cover. Moreover, $f(X^*) \leq 2f(X)$ holds for any vertex cover X in G . This provides a rounding 2-approximation algorithm for the submodular vertex cover problem.

9. Submodular Partition

Let f be a nonnegative submodular function on the subsets of V . The submodular partition problem asks for finding a partition of V into k disjoint nonempty subsets V_1, \dots, V_k minimizing $\sum_{j=1}^k f(V_j)$. This problem contains the multi-cut problem of graphs, which is known to be NP-hard if k is a part of the input.

A natural approach to the submodular partition problem is a greedy splitting algorithm that works as follows. The algorithm starts with a trivial partition that consists of only one component V . In each iteration, given a partition of V into j disjoint subsets, the algorithm computes a partition of each component W into S and $W \setminus S$ minimizing $f(S) + f(W \setminus S)$. This can be done by Queyranne's algorithm for symmetric submodular function minimization. Then the algorithm compares the minimum values among components and adopts the smallest one to obtain the partition into $j+1$ disjoint subsets. At the beginning of this iteration, the algorithm has already computed a minimum partition of each of $j-1$ old components, and hence it suffices to compute the minimum partition within the new two components. Therefore, the entire algorithm consists of $O(k)$ applications of symmetric submodular function minimization.

Zhao, Nagamochi, and Ibaraki [80] have presented and analyzed this greedy splitting algorithm. They have shown that the approximation ratio is $k-1$ for general nonnegative submodular functions and $2 - \frac{2}{k}$ for monotone submodular functions. The same performance guarantee $2 - \frac{2}{k}$ for symmetric submodular functions were suggested earlier by Queyranne.

They have also considered a generalization called a multiway partition problem. In this problem, a subset T is specified, and the goal is to find a partition that minimized $\sum_{j=1}^k f(V_j)$ with an additional constraint that $V_j \cap T \neq \emptyset$ for each $j = 1, \dots, k$. This problem contains the multiway cut problem of graphs,

which is also known to be NP-hard. On the other hand, if $T = V$, the multiway partition problem reduces to the above submodular partition problem.

In order to deal with the additional condition, they modified the greedy splitting algorithm as follows. In each iteration, the algorithm finds a minimum partition in each component W with an additional constraint that the both parts S and $W \setminus S$ must intersect with T . This can be done by applying general submodular function minimization $|W \cap T|$ times. Thus the modified algorithm runs in $O(k|T|\text{SFM})$ time, where **SFM** designates the time for submodular function minimization. Zhao, Nagamochi, and Ibaraki [80] have shown that the same performance guarantee as the submodular partition problem is extended to this general setting.

A recent paper of Okumoto, Fukunaga, and Nagamochi [66] presents improved approximation algorithms for the submodular partition problem for general nonnegative submodular functions with fixed k . In particular, they have devised an efficient exact algorithm for $k = 3$.

References

- [1] A. A. Ageev and M. I. Sviridenko: Pipage rounding: A new method of constructing algorithms with proven performance guarantee. *J. Combin. Optim.*, 8 (2004), 307–328.
- [2] R. Bar-Yehuda and S. Even: A linear time approximation algorithm for the weighted vertex cover problem. *J. Algorithms*, 2 (1981), 198–203.
- [3] R. E. Bixby, W. H. Cunningham, and D. M. Topkis: Partial order of a polymatroid extreme point. *Math. Oper. Res.*, 10 (1985), 367–378.
- [4] G. Calinescu, C. Chekuri, M. Pál, and J. Vondrák: Maximizing a submodular set function subject to a matroid constraint. *SIAM J. Comput.*, to appear.
- [5] W. H. Cunningham: Testing membership in matroid polyhedra. *J. Combin. Theory*, Ser. B, 36 (1984), 161–188.
- [6] W. H. Cunningham: On submodular function minimization. *Combinatorica*, 5 (1985), 185–192.
- [7] W. H. Cunningham and A. Frank: A primal-dual algorithm for submodular flows. *Math. Oper. Res.*, 10 (1985), 251–262.
- [8] A. W. M. Dress and W. Wenzel: Valuated matroids. *Adv. Math.*, 93 (1992), 214–250.
- [9] J. Edmonds: Submodular functions, matroids, and certain polyhedra. *Combinatorial Structures and Their Applications*, R. Guy, H. Hanani, N. Sauer, and J. Schönheim, eds., Gordon and Breach, 1970, 69–87.
- [10] J. Edmonds and R. Giles: A min-max relation for submodular functions on graphs. *Ann. Discrete Math.*, 1 (1977), 185–204.
- [11] J. Edmonds and R. M. Karp: Theoretical improvements in algorithmic efficiency for network flow problems. *J. ACM*, 19 (1972), 248–264.

-
- [12] A. Federgruen and H. Groenevelt: Characterization and optimization of achievable performance in general queueing systems. *Oper Res.*, 36 (1988), 733–741.
 - [13] U. Feige: A threshold of $\ln n$ for approximating set cover. *J. ACM*, 45 (1998), 634–652.
 - [14] U. Feige, V. Mirrokni, and J. Vondrák: Maximizing non-monotone submodular functions. *Proc. 48th IEEE Annual Symposium on Foundation of Computer Science* (2007), 461–471.
 - [15] L. Fleischer and S. Iwata: Improved algorithms for submodular function minimization and submodular flow. *Proceedings of the 32nd ACM Symposium on Theory of Computing* (2000), 107–116.
 - [16] L. Fleischer and S. Iwata: A push-relabel framework for submodular function minimization and applications to parametric optimization. *Discrete Appl. Math.*, 131 (2003), 311–322.
 - [17] L. Fleischer, S. Iwata, and S. T. McCormick: A faster capacity scaling algorithm for minimum cost submodular flow. *Math. Programming*, 92 (2002), 119–139.
 - [18] A. Frank: How to make a digraph strongly connected. *Combinatorica*, 1 (1981), 145–153.
 - [19] A. Frank: An algorithm for submodular functions on graphs. *Ann. Discrete Math.*, 16 (1982), 97–120.
 - [20] A. Frank: Submodular functions in graph theory. *Discrete Math.*, 111 (1993), 231–241.
 - [21] A. Frank: Applications of submodular functions. *Surveys in Combinatorics*, K. Walker, ed., Cambridge University Press, 1993, 85–136.
 - [22] A. Frank: Applications of relaxed submodularity. *Doc. Math.*, Extra Volume ICM III (1998), 355–354.
 - [23] S. Fujishige: Polymatroidal dependence structure of a set of random variables. *Inform. Contr.*, 39 (1978), 55–72.
 - [24] S. Fujishige: Lexicographically optimal base of a polymatroid with respect to a weight vector. *Math. Oper. Res.*, 5 (1980), 186–196.
 - [25] S. Fujishige: Theory of submodular programs — A Fenchel-type min-max theorem and subgradients of submodular functions. *Math. Programming*, 29 (1984), 142–155.
 - [26] S. Fujishige: Submodular systems and related topics. *Math. Programming Stud.*, 22 (1984), 113–131.
 - [27] S. Fujishige: *Submodular Functions and Optimization*, Elsevier, 2005.
 - [28] S. Fujishige and X. Zhang: New algorithms for the intersection problem of submodular systems. *Japan. J. Indust. Appl. Math.*, 9 (1992), 369–382.
 - [29] G. Goel, C. Karande, P. Tripathi, and L. Wang: Approximability of combinatorial problems with multi-agent submodular cost functions. *Proc. 50th Annual IEEE Symposium on Foundations of Computer Science* (2009), 755–764.
 - [30] M. X. Goemans, N. J. A. Harvey, S. Iwata and V. Mirrokni: Approximating submodular functions everywhere. *Proc. 20th ACM-SIAM Symposium on Discrete Algorithms* (2009), 535–544.

-
- [31] M. X. Goemans, S. Iwata, and R. Zenklusen: An algorithmic framework for wireless information flow. *Proc. 47th Allerton Conference on Communication, Control, and Computing* (2009), to appear.
- [32] M. X. Goemans and V. S. Ramakrishnan: Minimizing submodular functions over families of subsets. *Combinatorica*, 15 (1995), 499–513.
- [33] A. V. Goldberg and R. E. Tarjan: A new approach to the maximum flow problem. *J. ACM*, 35 (1988), 921–940.
- [34] M. Grötschel, L. Lovász, and A. Schrijver: The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1 (1981), 169–197.
- [35] M. Grötschel, L. Lovász, and A. Schrijver: *Geometric Algorithms and Combinatorial Optimization*, Springer-Verlag, 1988.
- [36] J. Hartline, V. S. Mirrokni, and M. Sundarajan: Optimal marketing strategies over social networks. *Proc. 17th International World Wide Web Conference* (2008), 189–198.
- [37] D. S. Hochbaum: Approximation algorithms for the set covering and vertex cover problems. *SIAM J. Comput.*, 11 (1982), 555–556.
- [38] B. Hoppe and É. Tardos: The quickest transshipment problem. *Math. Oper. Res.*, 25 (2000), 36–62.
- [39] T. Itoko and S. Iwata: Computational geometric approach to submodular function minimization for multiclass queueing systems. *Proc. 12th Conference on Integer Programming and Combinatorial Optimization* (2007), LNCS 4513, Springer-Verlag, 267–279.
- [40] S. Iwata: A capacity scaling algorithm for convex cost submodular flows. *Math. Programming*, 76 (1997), 299–308.
- [41] S. Iwata: A fully combinatorial algorithm for submodular function minimization. *J. Combin. Theory, Ser. B*, 84 (2002), 203–212.
- [42] S. Iwata: A faster scaling algorithm for minimizing submodular functions. *SIAM J. Comput.*, 32 (2003), 833–840.
- [43] S. Iwata: Submodular function minimization. *Math. Programming*, 112 (2008), 45–64.
- [44] S. Iwata, L. Fleischer, and S. Fujishige: A combinatorial strongly polynomial algorithm for minimizing submodular functions. *J. ACM*, 48 (2001), 761–777.
- [45] S. Iwata and K. Nagano: Submodular function minimization under covering constraints. *Proc. 50th Annual IEEE Symposium on Foundation of Computer Science* (2009), 671–680.
- [46] S. Iwata and J. B. Orlin: A simple combinatorial algorithm for submodular function minimization. *Proc. 20th ACM-SIAM Symposium on Discrete Algorithms* (2009), 1230–1237.
- [47] D. Kempe, J. Kleinberg, and É. Tardos: Maximizing the spread of inference through a social network. *Proc. 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (2003), 137–146.
- [48] L. G. Khachiyan: A polynomial algorithm in linear programming. *Soviet Math. Dokl.*, 20 (1979), 191–194.

- [49] B. Korte and J. Vygen: *Combinatorial Optimization — Theory and Algorithms*, Springer-Verlag, Berlin, 2000.
- [50] A. Krause and C. Guestrin: Near-optimal observation selection using submodular functions. *Proc. of 22nd Conference on Artificial Intelligence* (2007), 1650–1654.
- [51] J. Lee, V. Mirrokni, V. Nagarajan, and M. Sviridenko: Maximizing nonmonotone submodular functions under matroid or knapsack constraints. *SIAM J. Discrete Math.*, 23 (2010), 2053–2078.
- [52] J. K. Lenstra, D. Shmoys, and É. Tardos: Approximation algorithms for scheduling unrelated parallel machines. *Math. Programming*, 46 (1990), 259–271.
- [53] L. Lovász: Submodular functions and convexity. *Mathematical Programming — The State of the Art*, A. Bachem, M. Grötschel and B. Korte, eds., Springer-Verlag, 1983, 235–257.
- [54] C. L. Lucchesi and D. H. Younger: A minimax theorem for directed graphs. *J. London Math. Soc.*, 17 (1978), 369–374.
- [55] S. T. McCormick: Submodular function minimization. *Discrete Optimization*, K. Aardal, G. Nemhauser, and R. Weismantel, eds., Handbooks in Operations Research, Vol. 12, Elsevier, 2005, 321–391.
- [56] N. Megiddo: Combinatorial optimization with rational objective functions. *Math. Oper. Res.*, 4 (1979), 414–424.
- [57] N. Megiddo: Applying parallel computation algorithms in the design of serial algorithms. *J. ACM*, 30 (1983), 852–865.
- [58] K. Murota: Convexity and Steinitz’s exchange property. *Adv. Math.*, 124 (1996), 272–311.
- [59] K. Murota: Discrete convex analysis. *Math. Programming*, 83 (1998), 313–371.
- [60] K. Murota: *Discrete Convex Analysis*, SIAM, 2003.
- [61] H. Nagamochi: Minimum degree orderings. *Algorithmica*, 56 (2010), 17–34.
- [62] H. Nagamochi and T. Ibaraki: Computing edge-connectivity of multigraphs and capacitated graphs. *SIAM J. Discrete Math.*, 5 (1992), 54–66.
- [63] H. Nagamochi and T. Ibaraki: A note on minimizing submodular functions. *Inform. Process. Lett.*, 67 (1998), 239–244.
- [64] K. Nagano: A strongly polynomial algorithm for line search in submodular polyhedra. *Discrete Optim.*, 4 (2007), 349–359.
- [65] G. L. Nemhauser, L. A. Wolsey, and M. L. Fisher: An analysis of approximations for maximizing submodular set functions I. *Math. Programming*, 14 (1978), 265–294.
- [66] K. Okumoto, T. Fukunaga, and H. Nagamochi: Divide-and-conquer algorithm for partitioning hypergraphs and submodular systems. *Proc. 20th International Symposium on Algorithms and Computation* (2009), LNCS 5878, Springer-Verlag, 55–64.
- [67] J. B. Orlin: A faster strongly polynomial time algorithm for submodular function minimization. *Math. Programming*, 118 (2009), 237–251.
- [68] M. Queyranne: Structure of a simple scheduling polyhedra. *Math. Programming*, 58 (1993), 263–285.

-
- [69] M. Queyranne: Minimizing symmetric submodular functions. *Math. Programming*, 82 (1998), 3–12.
- [70] R. Rizzi: On minimizing symmetric set functions. 20 (2000), 445–450.
- [71] A. Roig, E. Ventura, and P. Weil: On the complexity of the Whitehead minimization problem. *Int. J. Algebra Comput.*, 8 (2007), 1611–1634.
- [72] A. Schrijver: A combinatorial algorithm minimizing submodular functions in strongly polynomial time. *J. Combin. Theory, Ser. B*, 80 (2000), 346–355.
- [73] A. Schrijver: *Combinatorial Optimization — Polyhedra and Efficiency*, Springer-Verlag, Berlin, 2003.
- [74] J. G. Shanthikumar and D. D. Yao: Multiclass queueing systems: polymatroidal structure and optimal scheduling control. *Oper. Res.*, 40 (1992), S293–S299.
- [75] L. S. Shapley: Cores of convex games. *Int. J. Game Theory*, 1 (1971), 11–26.
- [76] Z. Svitkina and L. Fleischer: Submodular approximation: Sampling based algorithms and lower bounds. *Proc. 49th IEEE Annual Symposium on Foundation of Computer Science* (2008), 697–706.
- [77] J. Vondrák: Symmetry and approximability of submodular maximization problems. *Proc. 50th IEEE Annual Symposium on Foundation of Computer Science* (2009), 651–670.
- [78] J. Vygen: A note on Schrijver’s submodular function minimization algorithm. *J. Combin. Theory, Ser. B*, 88 (2003), 399–402.
- [79] H. Whitney: On the abstract properties of linear dependence. *Amer. J. Math.*, 57 (1935), 509–533.
- [80] L. Zhao, H. Nagamochi, and T. Ibaraki: Greedy splitting algorithms for approximating multiway partition problems. *Math. Programming*, 102 (2005), 167–183.

Recent Advances in Structural Optimization

Yurii Nesterov*

Abstract

In this paper we present the main directions of research in Structural Convex Optimization. In this field, we use additional information on the structure of specific problem instances for accelerating standard Black-Box methods. We show that the proper use of problem structure can provably accelerate these methods by the order of magnitudes. As examples, we consider polynomial-time interior-point methods, smoothing technique, minimization of composite functions and some other approaches.

Mathematics Subject Classification (2010). Primary 90C25; Secondary 90C06.

Keywords. Convex optimization, structural optimization, complexity estimates, worst-case analysis, polynomial-time methods, interior-point methods, smoothing technique.

1. Introduction

Optimization problems are usually related to some models of the real-life situations. On the other hand, in order to develop a method for solving such problems, a numerical analyst starts from creating an appropriate model of a particular class of optimization problems. For this, there exist several reasons. Firstly, it is natural to use the developed scheme for solving many optimization problem with similar characteristics. Secondly, the model of the problem provides us with useful properties and inequalities helping to approach the optimal solution. Finally, fixation of the model allows us to perform a worst-case complexity analysis and to develop the optimal schemes.

The progress in performance of methods in Convex Optimization during the last three decades is closely related to evolution of the above models and

*Catholic University of Louvain (UCL), Department INMA/CORE
CORE, 34 voie du Roman Pays, 1348 Louvain-la-Neuve, Belgium.
E-mail: Yurii.Nesterov@uclouvain.be.

to better understanding of significance of the help we can offer to numerical schemes by opening an access to the structure of problem instances.

At the very first stage of the development of our field, the standard model of optimization problem was quite poor. It was not even clear that such a notion is necessary or useful. The tradition was to fix the analytic form of the problem and the classes of functional components. For example, for “unconstrained” convex optimization problem, the standard form was as follows:

$$\min_{x \in Q} f(x), \quad (1)$$

where $Q \subset R^n$ is a closed bounded convex set ($\|x\| \leq R$ for all $x \in Q$), and f is a closed convex function. If we assume that

$$\|\nabla f(x)\| \leq L \quad \forall x \in Q, \quad (2)$$

then we get the problem class \mathcal{C}_1 , which is formed by the problems of unconstrained “nonsmooth” minimization. Assuming that

$$\|\nabla f(x) - \nabla f(y)\| \leq M\|x - y\| \quad \forall x, y \in Q, \quad (3)$$

we get the problem class \mathcal{C}_2 of smooth optimization problems. Thus, the model of the problem was represented as the set of useful properties and inequalities which can be somehow employed by optimization scheme. For example, if $f \in \mathcal{C}_1$, then by its convexity we know that

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle \quad \forall x, y \in Q. \quad (4)$$

Therefore, evaluating this function at points $\{x_i\}_{i=1}^k$, we can form its model as follows:

$$\mathcal{L}_k(x) \stackrel{\text{def}}{=} \max_{1 \leq i \leq k} [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle] \leq f(x), \quad x \in Q. \quad (5)$$

In this methodology, the elements of function $\mathcal{L}_k(x)$ and the bound (2) represent the *full available information* on our objective function. In other words, the designed numerical methods are obliged to work with these objects only. If $f \in \mathcal{C}_2$, then we have also

$$f(y) \stackrel{(3)}{\leq} f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}L\|y - x\|^2, \quad \forall x, y \in Q.$$

This inequality together with (5) helps to increase the rate of convergence of corresponding optimization schemes.

Thus, it was natural that the Complexity Theory of Convex Optimization, developed by A.Nemirovskii and D.Yudin [7] in the late 70’s, was based on the *Black-Box Concept*. It was assumed that the only information the optimization methods can learn about the particular problem instance is the values and derivatives of these components at some test points. This data can be reported

by a special unit called *oracle* which is *local*. This means that it is not changing if the function is modified far enough from the test point. At the time of its development, this concept fitted very well the existing computational practice, where the interface between the general optimization packages and the problem's data was established by Fortran subroutines created independently by the users.

Black-Box framework allows to speak about the lower performance bounds for different problem classes in terms of *informational complexity*. That is the lower estimate for the number of calls of oracle which is necessary for any optimization method in order to guarantee delivering an ϵ -solution to any problem from the problem class. In this performance measure we do not include at all the complexity of auxiliary computations of the scheme.

In the table below, the first column indicates the problem class, the second one gives an upper bound for allowed number of calls of oracle in the optimization scheme¹, and the last column gives the lower bound for analytical complexity of the problem class, which depends on the absolute accuracy ϵ and the class parameters.

Problem class	Limit for calls	Lower bound
$\mathcal{C}_1 : \ \nabla f(\cdot)\ \leq L$	$\leq O(n)$	$O\left(\frac{L^2 R^2}{\epsilon^2}\right)$
$\mathcal{C}_2 : \ \nabla^2 f(\cdot)\ \leq M$	$\leq O(n)$	$O\left(\frac{M^{1/2} R}{\epsilon^{1/2}}\right)$
$\mathcal{C}_3 : \ \nabla f(\cdot)\ \leq L$	$\geq O(n)$	$O\left(n \ln \frac{LR}{\epsilon}\right)$

(6)

It is important that these bounds are *exact*. This means that there exist methods, which have efficiency estimates on corresponding problem classes proportional to the lower bounds. The corresponding *optimal methods* were developed in [6, 7, 16, 19, 20]. For further references, we present a simplified version of the optimal method [7] as applied to the problem (1) with $f \in \mathcal{C}_2$:²

Choose a starting point $y_0 \in Q$ and set $x_{-1} = y_0$. For $k \geq 0$ iterate:

$$x_k = \arg \min_{x \in Q} \left[f(y_k) + \langle \nabla f(y_k), x - y_k \rangle + \frac{M}{2} \|x - y_k\|^2 \right], \tag{7}$$

$$y_{k+1} = x_k + \frac{k}{k+3}(x_k - x_{k-1}).$$

¹If this upper bound is smaller than $O(n)$, then the dimension of the problem is really very big, and we cannot afford the method to perform this amount of calls.

²In method (11)-(13) from [7], we can set $a_k = 1 + k/2$ since in the proof we need only to ensure $a_{k+1}^2 - a_k^2 \leq a_{k+1}$.

As we see, the complexity of each iteration of this scheme is comparable with that of the simplest gradient method. However, the rate of convergence of method (7) is much faster.

After a certain period of time, it became clear that, despite to its mathematical excellence, Complexity Theory of Convex Optimization has a hidden drawback. Indeed, in order to apply convex optimization methods, we need to be *sure* that functional components of our problem are convex. However, we can check convexity only by analyzing the *structure* of these functions:³ If our function is obtained from the *basic* convex functions by *convex* operations (summation, maximum, etc.), we conclude that it is convex. If not, then we have to apply general nonlinear optimization methods which usually do not have theoretical guarantees for the global performance.

Thus, the functional components of the problem are not in the black box in the moment we check their convexity and choose minimization scheme. However, we put them into the black box for numerical methods. This is the main conceptual contradiction of the standard Convex Optimization Theory.

As we have already mention, the progress in Convex Optimization was mainly related to discovering the different possibilities for opening the Black Box for numerical methods. In this paper we present some of these approaches and discuss the corresponding improvements of complexity estimates as compared to the standard Black-Box framework. The order of discussion of these approaches has certain logic. However, it does not reflect the chronology of the development.

2. Primal-dual Subgradient Methods

In our first approach, we do not accelerate the Black-Box methods. We just look inside the oracle and show how this information can be used for constructing an approximate solution to the dual problems [14].

Let the norm $\|\cdot\|$ be Euclidean. We can form a *linear model* of function $f \in \mathcal{C}_1$ as follows:

$$l_k(x) = \frac{1}{k+1} \sum_{i=0}^k [f(x_i) + \langle \nabla f(x_i), x - x_i \rangle].$$

This model can be used in the following optimization method [14]:

$$x_{k+1} = \arg \min_{x \in Q} [l_k(x) + \frac{1}{2} \gamma_k \|x - x_0\|^2], \quad (8)$$

where $\gamma_k > 0$ are certain parameters. This is a usual Black-Box subgradient method for solving problem (1). If $\gamma_k = \frac{L}{R\sqrt{k+1}}$ and $\hat{x}_k = \frac{1}{k+1} \sum_{i=0}^k x_i$, then

$$f(\hat{x}_k) - f(x^*) \leq \frac{2LR}{\sqrt{k+1}}, \quad k \geq 0, \quad (9)$$

³Numerical verification of convexity is an extremely difficult problem.

where x^* is the solution of problem (1). Thus, method (8) is optimal for our class of problems. In order to understand what is the dual problem, we need to look inside the oracle.

1. Discrete minimax. Consider the following variant of problem (1):

$$\min_{x \in Q} [f(x) = \max_{1 \leq j \leq m} f_j(x)], \tag{10}$$

where $f_j \in \mathcal{C}_1$, $j = 1, \dots, m$. Denote $\Delta_m = \{y \in R_+^m : \sum_{j=1}^m y^{(j)} = 1\}$. Then

$$\begin{aligned} f^* &= \min_{x \in Q} \max_{1 \leq j \leq m} f_j(x) = \min_{x \in Q} \max_{y \in \Delta_m} \sum_{j=1}^m y^{(j)} f_j(x) \\ &= \max_{y \in \Delta_m} \left[\phi(y) \stackrel{\text{def}}{=} \min_{x \in Q} \sum_{j=1}^m y^{(j)} f_j(x) \right]. \end{aligned}$$

Thus, the dual problem is

$$f^* = \max_{y \in \Delta_m} \phi(y). \tag{11}$$

Note that the computation of the value of dual function $\phi(y)$ may be difficult since it requires to solve a nonlinear optimization problem.

Denote by e_j the j th coordinate vector in R^m . Let us look at the following variant of method (8) with $\gamma_k = \frac{L}{R\sqrt{k+1}}$.

<p>Initialization: Set $l_0(x) \equiv 0$, $m_0 = 0 \in Z^m$.</p>	
<p>Iteration ($k \geq 0$):</p> <ol style="list-style-type: none"> 1. Choose any j_k^*: $f_{j_k^*}(x_k) = f(x_k)$. 2. Set $l_{k+1}(x) = \frac{k}{k+1}l_k(x) + \frac{1}{k+1}[f(x_k) + \langle \nabla f_{j_k^*}(x_k), x - x_k \rangle]$. 3. Compute $x_{k+1} = \arg \min_{x \in Q} \{l_{k+1}(x) + \frac{1}{2}\gamma_k\ x - x_0\ ^2\}$. 4. Update $m_{k+1} = m_k + e_{j_k^*}$. 	(12)
<p>Output: $\hat{x}_{k+1} = \frac{1}{k+1} \sum_{i=0}^k x_i$, $\hat{y}_{k+1} = \frac{1}{k+1} m_{k+1}$.</p>	

Thus, the entries of vector \hat{y}_k are the *frequencies* of appearing the corresponding functional components as the biggest ones of the objective function. For the output of this process we have the following guarantee:

$$f(\hat{x}_k) - \phi(\hat{y}_k) \leq \frac{2LR}{\sqrt{k+1}}, \quad k \geq 0. \quad (13)$$

2. Primal-dual problem. Let f be a closed convex function defined on R^n . Consider the conjugate function:

$$f_*(s) = \sup_{x \in R^n} [\langle s, x \rangle - f(x)].$$

Then $f(x) = \max_s [\langle s, x \rangle - f_*(s)]$, $x \in R^n$. Denote by $s(x)$ an optimal solution of the latter problem. Note that

$$\begin{aligned} f^* &= \min_{x \in Q} \max_{s \in \text{dom} f_*} [\langle s, x \rangle - f_*(s)] = \max_{s \in \text{dom} f_*} \min_{x \in Q} [\langle s, x \rangle - f_*(s)] \\ &= \max_{s \in \text{dom} f_*} \left[\psi(s) \stackrel{\text{def}}{=} -\xi_Q(s) - f_*(s) \right], \end{aligned}$$

where $\xi_Q(u) = \max_{x \in Q} \langle u, x \rangle$. Thus, the problem dual to (1) is as follows:

$$f^* = \max_{s \in \text{dom} f_*} \psi(s). \quad (14)$$

It appears, that the method (8) is able to approximate the optimal solution to this problem. Indeed, let $\{x_k\}$ be formed by (8) with $\gamma_k = \frac{L}{R\sqrt{k+1}}$. Define

$$\hat{s}_k = \frac{1}{k+1} \sum_{i=0}^k s(x_i). \quad \text{Then}$$

$$f(\hat{x}_k) - \phi(\hat{y}_k) \leq \frac{2LR}{\sqrt{k+1}}, \quad k \geq 0.$$

Again, we find an approximate solution to the dual problem without computing the values of the dual function (this may be difficult).

3. Polynomial-time Interior-point Methods

Thus, we have seen that a proper use of structure of the oracle can help in generating an approximate solution to the dual problem. Is it possible to use this structure for accelerating the Black-Box schemes? Intuitively we always hope that this is true. Unfortunately, structure is a very fuzzy notion, which is quite difficult to formalize. One possible way to describe the structure is to fix the *analytical type* of functional components. For example, we can consider the problems with linear constraints only. It can help, but this approach is very fragile: If we add just a single constraint of another type, then we get a new problem class, and all theory must be redone from scratch.

On the other hand, it is clear that having the structure at hand we can play a lot with the *analytical form* of the problem. We can rewrite the problem in many equivalent settings using non-trivial transformations of variables or constraints, introducing additional variables, etc. However, this would serve almost no purpose without fixing a clear final goal. So, let us try to understand what it could be.

As usual, it is better to look at classical examples. In many situations the sequential reformulations of the initial problem can be seen as a part of numerical scheme. We start from a complicated problem \mathcal{P} and, step by step, change its structure towards to the moment we get a trivial problem (or, a problem which we know how to solve):

$$\mathcal{P} \longrightarrow \dots \longrightarrow (f^*, x^*).$$

A good example of such a strategy is the standard approach for solving system of linear equations

$$Ax = b.$$

We can proceed as follows:

1. Check if A is symmetric and positive definite. Sometimes this is clear from the origin of the matrix.
2. Compute Cholesky factorization of this matrix:

$$A = LL^T,$$

where L is a lower-triangular matrix. Form two auxiliary systems

$$Ly = b, \quad L^T x = y.$$

3. Solve these system by sequential exclusion of variables.

Imagine for a moment that we do not know how to solve the system of linear equations. In order to *discover* the above scheme we should apply the following

GOLDEN RULES	
<ol style="list-style-type: none"> 1. Find a class of problems which can be solved very efficiently.^a 2. Describe the transformation rules for converting the initial problem into desired form. 3. Describe the class of problems for which these transformation rules are applicable. 	(15)
<hr style="width: 20%; margin-left: 0;"/> <p>^aIn our example, it is the class of linear systems with triangular matrices.</p>	

In Convex Optimization, these rules were used already several times for breaking down the limitations of Complexity Theory.

Historically, the first example of that type was the theory of polynomial-time interior-point methods (IPM) based on *self-concordant barriers* [15]. The first step in the development of this theory was discovery of *unconstrained* minimization problems which can be solved efficiently by the Newton method. We say that a closed convex function f is *self-concordant* on its open domain $\text{dom } f \subset R^n$ if

$$D^3 f(x)[h]^3 \leq 2D^2 f(x)[h]^2 \quad \forall x \in \text{dom } f, h \in R^n,$$

where $D^k f(x)[h]^k$ is the k th differential of function f at x along direction h . It appears that the properties of these functions fit very well the Newton scheme.

Let us use the Hessian $\nabla^2 f(x)$ of such a function for defining a *local norm* around x :

$$\|h\|_x = \langle \nabla^2 f(x)h, h \rangle^{1/2}, \quad \|s\|_x^* = \langle s, [\nabla^2 f(x)]^{-1}s \rangle^{1/2}.$$

(It is possible to prove that if $\text{dom } f$ is bounded, then the Hessian is nondegenerate at any point of the domain.) Then we can define the Dikin ellipsoid at x as follows:

$$W_r(x) = \{y \in R^n : \|y - x\|_x \leq r\}.$$

It appears that for any $r < 1$ we have $W_r(x) \subset \text{dom } f$ for any feasible x . Moreover, inside this ellipsoid we can predict very well the variation of the Hessian:

$$(1-r)^2 \nabla^2 f(x) \preceq \nabla^2 f(y) \preceq \frac{1}{(1-r)^2} \nabla^2 f(x), \quad \forall y \in W_r(x), x \in \text{dom } f.$$

This property results in the following behavior of the Damped Newton Method:

$$x_{k+1} = x_k - \frac{[\nabla^2 f(x)]^{-1} \nabla f(x)}{1 + \|\nabla f(x)\|_x^*}.$$

If $\|\nabla f(x)\|_x \geq \beta$ for some $\beta \in (0, 1)$, then $f(x_{k+1}) \leq f(x_k) - [\beta - \ln(1 + \beta)]$, else

$$\|\nabla f(x_{k+1})\|_{x_{k+1}}^* \leq 2 \left(\|\nabla f(x_k)\|_{x_k}^* \right)^2. \quad (16)$$

Thus, we have now an affine-invariant description of the region of quadratic convergence of the Newton method:

$$\{x \in \text{dom } f : \|\nabla f(x)\|_x < \frac{1}{2}\}.$$

Now we can try to use our achievement for solving more complicated problems, the problems of *constrained* minimization.

Consider the following *standard* minimization problem:

$$\min_{x \in Q} \langle c, x \rangle, \quad (17)$$

where Q is a bounded closed convex set. Let us assume that $Q = \text{Cl}(\text{dom } f)$ for some self-concordant function f . Then we can try to solve (17) by a *path-following method*. Define the central path $x(t)$, $t > 0$, as follows:

$$tc + \nabla f(x(t)) = 0. \quad (18)$$

This is the first-order optimality condition for the unique minimum of self-concordant function

$$\psi_t(x) = t\langle c, x \rangle + f(x),$$

for which we already have a convenient description of the region of quadratic convergence of the Newton scheme. How quickly we can increase the penalty parameter keeping the possibility to come in a close neighborhood of the new point at the central path by a quadratically convergent Newton scheme? For that we need to ensure

$$\begin{aligned} \frac{1}{2} > \|\nabla \psi_{t+\Delta}(x(t))\|_{x(t)}^* &= \|(t+\Delta)c + \nabla f(x(t))\|_{x(t)}^* \\ &\stackrel{(18)}{=} \Delta \|c\|_{x(t)}^* \stackrel{(18)}{=} \frac{\Delta}{t} \|\nabla f(x(t))\|_{x(t)}^*. \end{aligned} \quad (19)$$

Thus, in order to increase t in a linear rate, we need to assume uniform boundedness of the local norm of the gradient of f . This is the reason for working with the following barrier function.

Definition 1. Function f is called a ν -self-concordant barrier for convex set Q if it is self-concordant on $\text{int } Q$ and

$$\langle \nabla f(x), [\nabla^2 f(x)]^{-1} \nabla f(x) \rangle \leq \nu, \quad x \in \text{dom } f.$$

The value ν is called the *parameter* of the barrier f .

Self-concordant barriers have many useful properties (see [15], [8]). One of them is related to asphericity of the set Q with respect to the point $x(0)$, which is called the *analytic center* of Q :

$$W_1(x(0)) \subseteq Q \subseteq W_{\nu+2\sqrt{\nu}}(x(0)). \quad (20)$$

Note that the value of the barrier parameter ν can be much smaller than the dimension of the space of variables.

As we can see from the reasoning (19), we can solve the standard minimization problems with complexity $O(\sqrt{\nu} \ln \frac{\nu}{\epsilon})$, where ϵ is the desired accuracy of the solution. How wide is the class of problems to which we can apply this machinery?

It appears that in principle we cover *all* convex optimization problems. Indeed, for any closed convex set Q we can define the following *universal barrier*:

$$f_Q(x) = \kappa \cdot \ln \text{Vol } P(x), \quad P(x) = \{s : \langle s, y - x \rangle \leq 1 \ \forall y \in Q\}.$$

Then, for certain value of $\kappa > 0$, this function is $O(n)$ -self-concordant barrier for Q . Hence, in principle, we can solve all convex optimization problems with complexity $O(\sqrt{n} \ln \frac{n}{\epsilon})$. Of course, in the framework of Back-Box methods this is just impossible. Hence, we conclude that something should violate the Black-Box assumptions. And indeed, this is the process of *creating* the self-concordant barriers. There exists a kind of calculus for doing this. However, it needs a direct access to the structure of the problem, possibility to introduce additional variables, etc. As a result, we are able to apply linearly convergent methods practically to all convex optimization problem with known structure. It is interesting that the standard classification of the problems in accordance to the level of smoothness of functional components is useless here. We need only a possibility to construct self-concordant barriers for their epigraphs. However, note that each iteration of the path-following schemes is quite heavy. This is the reason for development of the cheap gradient schemes, which we describe in the remaining sections.

4. Smoothing Technique

The second example of using the rules (15) needs more explanations. By certain circumstances, these results were discovered with a delay of twenty years. Perhaps they were too simple. Or maybe they are in a seemingly very sharp contradiction with the rigorously proved lower bounds of Complexity Theory.

Anyway, now everything looks almost evident. Indeed, in accordance to Rule 1 in (15), we need to find a class of very easy problems. And this class can be discovered *directly* in the table (6)! To see that, let us compare the complexity of the classes \mathcal{C}_1 and \mathcal{C}_2 for the accuracy of 1% ($\epsilon = 10^{-2}$). Note that in this case, the accuracy-dependent factors in the efficiency estimates vary from ten to ten thousands. So, the natural question is:

Can the easy problems from \mathcal{C}_2 help us somehow in finding an approximate solution to the difficult problems from \mathcal{C}_1 ?

And the evident answer is: Yes, of course! It is a simple exercise in Calculus to show that we can always approximate a Lipschitz-continuous nonsmooth convex function on a bounded convex set with a uniform accuracy $\epsilon > 0$ by a smooth convex function with Lipschitz-continuous gradient. We pay for the accuracy of approximation by a large Lipschitz constant M for the gradient, which should be of the order $O(\frac{1}{\epsilon})$. Putting this bound for M in the efficiency estimate of \mathcal{C}_2 in (6), we can see that in principle, it is possible to minimize

nonsmooth convex functions by the oracle-based gradient methods with analytical complexity $O(\frac{1}{\epsilon})$. But what about the Complexity Theory? It seems that it was *proved* that such efficiency is just impossible.

It is interesting that in fact we do not get any contradiction. Indeed, in order to minimize a smooth approximation of nonsmooth function by an oracle-based scheme, we need to change the initial oracle. Therefore, from mathematical point of view, we violate the Black-Box assumption. On the other hand, in the majority of practical applications this change is not difficult. Usually we can work directly with the structure of our problem, at least in the cases when it is created by ourselves.

Thus, the basis of the *smoothing technique* [9, 10] is formed by two ingredients: the above observation, and a trivial but systematic way for approximating a nonsmooth function by a smooth one. This can be done for convex functions represented explicitly in a max-form:

$$f(x) = \max_{u \in Q_d} \{\langle Ax - b, u \rangle - \phi(u)\},$$

where Q_d is a bounded and convex dual feasible set and $\phi(u)$ is a concave function. Then, choosing a nonnegative strongly convex function $d(u)$, we can define a smooth function

$$f_\mu(x) = \max_{u \in Q_d} \{\langle Ax - b, u \rangle - \phi(u) - \mu \cdot d(u)\}, \quad \mu > 0, \quad (21)$$

which approximates the initial objective. Indeed, denoting $D_d = \max_{u \in Q_d} d(u)$, we get

$$f(x) \geq f_\mu(x) \geq f(x) - \mu D_d.$$

At the same time, the gradient of function f_μ is Lipschitz-continuous with Lipschitz constant of the order of $O(\frac{1}{\mu})$ (see [9]) for details).

Thus, we can see that for an *implementable* definition (21), we get a possibility to solve problem (1) in $O(\frac{1}{\epsilon})$ iterations of the fast gradient method (7). In order to see the magnitude of improvement, let us look at the following example:

$$\min_{x \in \Delta_n} \left[f(x) \stackrel{\text{def}}{=} \max_{1 \leq j \leq m} \langle a_j, x \rangle \right], \quad (22)$$

where $\Delta_n \in R^n$ is a standard simplex. Then the properly implemented smoothing technique ensures the following rate of convergence:

$$f(x_N) - f^* \leq \frac{4\sqrt{\ln n \cdot \ln m}}{N} \cdot \max_{i,j} |a_j^{(i)}|.$$

If we apply to problem (22) the standard subgradient methods (e.g. [14]), we can guarantee only

$$f(x_N) - f^* \leq \frac{\sqrt{\ln n}}{\sqrt{N+1}} \cdot \max_{i,j} |a_j^{(i)}|.$$

Thus, up to a logarithmic factor, for obtaining the same accuracy, the methods based on smoothing technique need only a square root of iterations of the usual subgradient scheme. Taking into account, that usually the subgradient methods are allowed to run many thousands or even millions of iterations, the gain of the smoothing technique in computational time can be enormously big.⁴

It is interesting, that for problem (22) the computation of the smooth approximation is very cheap. Indeed, let us use for smoothing the *entropy function*:

$$d(u) = \ln m + \sum_{i=1}^n u^{(i)} \ln u^{(i)}, \quad u \in \Delta_m.$$

Then the smooth approximation (21) of the objective function in (22) has the following compact representation:

$$f_\mu(x) = \mu \ln \left[\frac{1}{m} \sum_{j=1}^m e^{\langle a_j, x \rangle / \mu} \right].$$

Thus, the complexity of the oracle for $f(x)$ and $f_\mu(x)$ is similar. Note that, as in the polynomial-time IPM theory, we apply the standard oracle-based method ((7) in this case) to a function which does not satisfy the Black-Box assumptions.

5. Conclusion

Let us briefly look at one more example of acceleration strategy in Structural Optimization.

Consider the problem of minimizing the *composite* objective function:

$$\min_{x \in R^n} [f(x) + \Psi(x)], \quad (23)$$

where the function f is a convex differentiable function on $\text{dom } \Psi$ with Lipschitz-continuous gradient, and function Ψ is an *arbitrary* closed convex function. Since Ψ can be even discontinuous, in general this problem is very difficult. However, if we assume that function Ψ is *simple*, then the situation is changing. Indeed, suppose that for any $\bar{y} \in \text{dom } \Psi$ we are able to solve explicitly the following auxiliary optimization problem:

$$\min_{x \in \text{dom } \Psi} [f(\bar{y}) + \langle \nabla f(\bar{y}), x - \bar{y} \rangle + \frac{M}{2} \|x - \bar{y}\|^2 + \Psi(x)] \quad (24)$$

⁴It is easy to see that the standard subgradient methods for nonsmooth convex minimization need indeed $O(\frac{1}{\varepsilon^2})$ operations to converge. Consider a univariate function $f(x) = |x|$, $x \in R$. Let us look at the subgradient process:

$$x_{k+1} = x_k - h_k f'(x_k), \quad x_0 = 1, \quad h_k = \frac{1}{\sqrt{k+1}} + \frac{1}{\sqrt{k+2}}, \quad k \geq 0.$$

It is easy to see that $|x_k| = \frac{1}{\sqrt{k+1}}$. However, the step-size sequence is optimal [6].

(compare with (7)). Then it becomes possible to develop for problem (23) the fast gradient methods (similar to (7)), which have the rate of convergence of the order $O(\frac{1}{k^2})$ (see [11] for details; similar technique was developed in [3]). Note that the formulation (23) can be also seen as a part of Structural Optimization since we use the knowledge of the structure of its objective function directly in the optimization methods.

In this paper, we have considered several examples of significant acceleration of the usual oracle-based methods. Note that the achieved progress is visible only because of the supporting complexity analysis. It is interesting that all these methods have some prototypes proposed much earlier:

- Optimal method (7) is very similar to the *heavy point* method:

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}),$$

where α and β are some fixed positive coefficients (see [17] for historical details).

- Polynomial-time IPM are very similar to *some* variants of the classical barrier methods [4].
- The idea to apply smoothing for solving minimax problems is also not new (see [18] and the references therein).

At certain moments of time, these ideas were quite new and attractive. However, they did not result in a significant change in computational practice since they were not provided with a convincing complexity analysis. Indeed, many other schemes have similar theoretical justifications and it was not clear at all why these particular suggestions deserve more attention. Moreover, even now, when we know that the modified variants of some old methods give excellent complexity results, we cannot say too much about the theoretical efficiency of the original schemes.

Thus, we have seen that in Convex Optimization the complexity analysis plays an important role in *selecting* the promising optimization methods among hundreds of others. Of course, it is based on investigation of the worst-case situation. However, even this limited help is important for choosing the perspective directions for further research. This is true especially now, when the development of Structural Optimization makes the problem settings and corresponding efficiency estimates more and more interesting and diverse.

The size of this paper does not allow us to discuss other interesting setting of Structural Convex Optimization (e.g. optimization in relative scale [12, 13]). However, we hope that even the presented examples can help the reader to find new and interesting research directions in this promising field (see, for example, [1, 2, 5]).

References

- [1] A. d'Aspremont, O. Banerjee, and L. El Ghaoui. First-Order Methods for Sparse Covariance Selection. *SIAM Journal on Matrix Analysis and its Applications*, **30**(1), 56–66, (2008).
- [2] O. Banerjee, L. El Ghaoui, and A. d'Aspremont. Model Selection Through Sparse Maximum Likelihood Estimation. *Journal of Machine Learning Research*, **9**, 485–516 (2008).
- [3] A. Beck and M. Teboulle. A Fast Iterative Shrinkage-Threshold Algorithm Linear Inverse Problems. Research Report, Technion (2008).
- [4] A.V. Fiacco and G.P. McCormick. Nonlinear Programming: Sequential Unconstrained Minimization Technique. *John Wiley*, New York, 1968.
- [5] S. Hoda, A. Gilpin, and J. Pena. Smoothing techniques for computing Nash equilibria of sequential games. Research Report. Carnegie Mellon University, (2008).
- [6] Nemirovsky A, Yudin D. Problems complexity and method efficiency in Optimization. 1983. Wiley-Interscience, New York.
- [7] Yu. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $O(\frac{1}{k^2})$. *Doklady AN SSSR* (translated as Soviet Mathematics Doklady), **269**(3), 543–547 (1983).
- [8] Yu. Nesterov. Introductory Lectures on Convex Optimization. *Kluwer*, Boston, 2004.
- [9] Yu. Nesterov. Smooth minimization of non-smooth functions. CORE Discussion Paper 2003/12 (2003). Published in *Mathematical Programming*, **103** (1), 127–152 (2005).
- [10] Yu. Nesterov. Excessive gap technique in nonsmooth convex minimization. *SIAM Journal on Optimization*, **16** (1), 235–249 (2005).
- [11] Yu. Nesterov. Gradient methods for minimizing composite objective function. *CORE Discussion Paper 2007/76*, (2007).
- [12] Yu. Nesterov. Rounding of convex sets and efficient gradient methods for linear programming problems. *Optimization Methods and Software*, **23**(1), 109–135 (2008).
- [13] Yu. Nesterov. Unconstrained convex minimization in relative scale. *Mathematics of Operations Research*, **34**(1), 180–193 (2009).
- [14] Yu. Nesterov. Primal-dual subgradient methods for convex problems. *Mathematical Programming*, **120**(1), 261–283 (2009).
- [15] Yu. Nesterov, A. Nemirovskii. Interior point polynomial methods in convex programming: Theory and Applications, SIAM, Philadelphia, 1994.
- [16] B. Polyak. A general method of solving extremum problems. *Soviet Mat. Dokl.* **8**, 593–597 (1967)

- [17] B. Polyak. Introduction to Optimization. *Optimization Software*, New York, 1987.
- [18] R. Polyak. Smooth Optimization Methods for Minimax Problems. *SIAM J. Control and Optimization*, **26**(6), 1274–1286 (1988).
- [19] N.Z. Shor. Minimization Methods for Nondifferentiable Functions. *Springer-Verlag*, Berlin, 1985.
- [20] S.P. Tarasov, L.G. Khachiyan, and I.I. Erlikh. The Method of Inscribed Ellipsoids. *Soviet Mathematics Doklady*, **37**, 226–230 (1988).

Computational Complexity of Stochastic Programming: Monte Carlo Sampling Approach

Alexander Shapiro*

Abstract

For a long time modeling approaches to stochastic programming were dominated by scenario generation methods. Consequently the main computational effort went into development of decomposition type algorithms for solving constructed large scale (linear) optimization problems. A different point of view emerged recently where computational complexity of stochastic programming problems was investigated from the point of view of randomization methods based on Monte Carlo sampling techniques. In that approach the number of scenarios is irrelevant and can be infinite. On the other hand, from that point of view there is a principle difference between computational complexity of two and multistage stochastic programming problems – certain classes of two stage stochastic programming problems can be solved with a reasonable accuracy and reasonable computational effort, while (even linear) multistage stochastic programming problems seem to be computationally intractable in general.

Mathematics Subject Classification (2010). Primary 90C15; Secondary 90C60.

Keywords. Stochastic programming, Monte Carlo sampling, sample average approximation, dynamic programming, asymptotics, computational complexity, stochastic approximation.

1. Introduction

Origins of Stochastic Programming are going back to more than 50 years ago to papers of Beale [2] and Dantzig [4]. The essential idea of that approach is

*This research was partly supported by the NSF award DMS-0914785 and ONR award N000140811104.

Georgia Institute of Technology, Atlanta, Georgia 30332, USA.
E-mail: ashapiro@isye.gatech.edu.

that decision variables are divided into groups of “here-and-now” decision variables which should be made before a realization of the uncertain data becomes available, and “wait-and-see” decision variables made after observing data and which are functions of the data. Furthermore, the uncertain parameters are modeled as random variables, with a specified probability distribution, and consequently the optimization problem is formulated in terms of minimizing the expected values of the uncertain costs.

Two-stage stochastic linear programming problems can be written in the form

$$\text{Min}_{x \in \mathcal{X}} \langle c, x \rangle + \mathbb{E}[Q(x, \xi)], \tag{1.1}$$

where $\mathcal{X} = \{x \in \mathbb{R}^n : Ax \leq b\}$ and $Q(x, \xi)$ is the optimal value of the second stage problem

$$\text{Min}_{y \in \mathbb{R}^m} \langle q, y \rangle \text{ subject to } Tx + Wy \leq h. \tag{1.2}$$

Some/all of the parameters, summarized in data vector $\xi := (q, h, T, W)$, of the second stage problem (1.2) are unknown (uncertain) at the first stage when a “here-and-now” decision x should be made, while second stage decisions $y = y(\xi)$ are made after observing the data and are functions of the data parameters. Parameters of the second stage problem are modeled as random variables and the expectation in (1.1) is taken with respect to a specified distribution of the random vector ξ .

This can be extended to the following multistage setting of T -stage stochastic programming problems

$$\text{Min}_{x_1 \in \mathcal{X}_1} f_1(x_1) + \mathbb{E} \left[\inf_{x_2 \in \mathcal{X}_2(x_1, \xi_2)} f_2(x_2, \xi_2) + \mathbb{E} \left[\dots + \mathbb{E} \left[\inf_{x_T \in \mathcal{X}_T(x_{T-1}, \xi_T)} f_T(x_T, \xi_T) \right] \right] \right], \tag{1.3}$$

driven by the random data process $\xi_1, \xi_2, \dots, \xi_T$. Here $x_t \in \mathbb{R}^{n_t}$, $t = 1, \dots, T$, are decision variables, $f_t : \mathbb{R}^{n_t} \times \mathbb{R}^{d_t} \rightarrow \mathbb{R}$ are continuous functions and $\mathcal{X}_t : \mathbb{R}^{n_{t-1}} \times \mathbb{R}^{d_t} \rightrightarrows \mathbb{R}^{n_t}$, $t = 2, \dots, T$, are measurable closed valued multifunctions. The first stage data, i.e., the vector ξ_1 , the function $f_1 : \mathbb{R}^{n_1} \rightarrow \mathbb{R}$, and the set $\mathcal{X}_1 \subset \mathbb{R}^{n_1}$ are deterministic (not random). It is said that the multistage problem is *linear* if the objective functions and the constraint functions are linear. That is,

$$f_t(x_t, \xi_t) := \langle c_t, x_t \rangle, \quad \mathcal{X}_1 := \{x_1 : A_1 x_1 \leq b_1\}, \tag{1.4}$$

$$\mathcal{X}_t(x_{t-1}, \xi_t) := \{x_t : B_t x_{t-1} + A_t x_t \leq b_t\}, \quad t = 2, \dots, T, \tag{1.5}$$

where $\xi_1 := (c_1, A_1, b_1)$ is known at the first stage (and hence is nonrandom), and $\xi_t := (c_t, B_t, A_t, b_t) \in \mathbb{R}^{d_t}$, $t = 2, \dots, T$, are data vectors.

For a long time approaches to modeling and solving stochastic programming problems were dominated by scenario generation methods. In such an approach a finite number of scenarios, representing what may happen in the future with

assigned probability weights, were generated and consequently the constructed optimization problem was solved by decomposition type methods. If one takes the position that generated scenarios represent reality in a reasonably accurate way, then there is no dramatic difference between two and multistage stochastic programming. An argument is that considering many scenarios is certainly better than solving the problem for just one scenario which would be a deterministic optimization approach. Everybody would agree, however, that what will really happen in the future will be different with probability one (w.p.1) from the set of generated scenarios. This raises the question of what does it mean to solve a stochastic programming problem? In that respect we may cite [3, p. 413]: "... it is absolutely unclear what the resulting solution [of a scenario based approximation of a multistage stochastic program] has to do with the problem we intend to solve. Strictly speaking, we even cannot treat this solution as a candidate solution, bad or good alike, to the original problem – the decision rules we end up with simply do not say what our decisions should be when the actual realizations of the uncertain data differ from the scenario realizations."

Somewhat different point of view emerged in a number of recent publications. It was shown theoretically and demonstrated in various numerical studies that certain classes of two stage stochastic programming problems can be solved with reasonable accuracy and reasonable computational effort by employing Monte Carlo sampling techniques. From that point of view the number of scenarios is irrelevant and can be astronomically large or even infinite. On the other hand, it turns out that computational complexity of multistage stochastic programming problems is conceptually different and scenario generation methods typically fail to solve multistage stochastic problems in a reasonable sense to a "true" optimality. It also could be pointed out the criticism of the modeling assumption of knowing the "true" probability distribution of the involved random data. We will not discuss this aspect of the stochastic programming approach here.

We will use the following notation and terminology through the paper. Notation " $:=$ " means "equal by definition"; by $\Delta_n := \{x \in \mathbb{R}_+^n : \sum_{i=1}^n x_i = 1\}$ we denote the n -dimensional simplex; \mathbb{S}^m denotes the linear space of $m \times m$ symmetric matrices; $\langle x, y \rangle$ denotes the standard scalar product of two vectors $x, y \in \mathbb{R}^n$ and $\langle x, y \rangle := \text{Tr}(xy)$ for $x, y \in \mathbb{S}^m$; unless stated otherwise $\|x\| = \sqrt{\langle x, x \rangle}$ denotes the Euclidean norm of vector x ; $C^* = \{y : \langle y, x \rangle \geq 0, \forall x \in C\}$ denotes the (positive) dual of cone $C \subset \mathbb{R}^n$; by " \preceq_C " we denote partial order induced by a closed convex cone C in a finite dimensional vector space, i.e., $x \preceq_C y$ means that $y - x \in C$; $\text{int}(C)$ denotes the interior of set $C \subset \mathbb{R}^n$; $\text{dist}(x, C) := \inf_{y \in C} \|x - y\|$ denotes the distance from point $x \in \mathbb{R}^n$ to set C ; $\text{Prob}(A)$ denotes probability of event A ; $\Delta(\xi)$ denotes measure of mass one at point ξ ; " $\xrightarrow{\mathcal{D}}$ " denotes convergence in distribution; $\mathcal{N}(\mu, \sigma^2)$ denotes normal distribution with mean μ and variance σ^2 ; $M_Y(t) := \mathbb{E}[\exp(tY)]$ is the moment generating function of random variable Y ; $\mathbb{E}[X|Y]$ denotes condi-

tional expectation of random variable X given Y , and $\text{Var}[X]$ denotes variance of X .

2. Asymptotic Analysis

Consider the following stochastic optimization problem

$$\text{Min}_{x \in \mathcal{X}} \{f(x) := \mathbb{E}[F(x, \xi)]\}. \quad (2.1)$$

Here \mathcal{X} is a nonempty closed subset of \mathbb{R}^n , ξ is a random vector whose probability distribution P is supported on a set $\Xi \subset \mathbb{R}^d$, and $F : \mathcal{X} \times \Xi \rightarrow \mathbb{R}$. Unless stated otherwise all probabilistic statements will be made with respect to the distribution P . The two stage problem (1.1) is of that form with $F(x, \xi) := \langle c, x \rangle + Q(x, \xi)$. We assume that the expectation $f(x)$ is well defined and finite valued for every $x \in \mathbb{R}^n$. This, of course, implies that $F(x, \xi)$ is finite valued for almost every (a.e.) $\xi \in \Xi$. For the two stage problem (1.1) the later means that the second stage problem (1.2) is bounded from below and its feasible set is nonempty for a.e. realization of the random data.

Suppose that we have a sample ξ^1, \dots, ξ^N of N realizations of the random vector ξ . We assume that the sample is iid (independent identically distributed). By replacing the “true” distribution P with its empirical estimate $P_N := \frac{1}{N} \sum_{j=1}^N \Delta(\xi^j)$, we obtain the following approximation of the “true” problem (2.1):

$$\text{Min}_{x \in \mathcal{X}} \left\{ \hat{f}_N(x) := \frac{1}{N} \sum_{j=1}^N F(x, \xi^j) \right\}. \quad (2.2)$$

We denote by ϑ^* and $\hat{\vartheta}_N$ the optimal values of problems (2.1) and (2.2), respectively, and by \mathcal{S} and \mathcal{S}_N the respective sets of optimal solutions.

In the recent literature on stochastic programming, problem (2.2) is often referred to as the Sample Average Approximation (SAA) problem, and in machine learning as the empirical mean optimization. The sample ξ^1, \dots, ξ^N can be a result of two somewhat different procedures – it can be given by a historical data of observations of ξ , or it can be generated in the computer by Monte Carlo sampling techniques. We will be mainly interested here in the second case where we view the SAA problem (2.2) as an approach to solving the true problem (2.1) by randomization techniques.

By the Law of Large Numbers (LLN) we have that for any $x \in \mathcal{X}$, $\hat{f}_N(x)$ tends to $f(x)$ w.p.1 as $N \rightarrow \infty$. Moreover, let us assume the following.

- (A1) For any $x \in \mathcal{X}$ the function $F(\cdot, \xi)$ is continuous at x for a.e. $\xi \in \Xi$.
- (A2) There exists an integrable function $H(\xi)$ such that $|F(x, \xi)| \leq H(\xi)$ for all $x \in \mathcal{X}$ and $\xi \in \Xi$.

These assumptions imply that $f(x)$ is continuous on \mathcal{X} and $\hat{f}_N(x)$ converges w.p.1 to $f(x)$ uniformly on any compact subset of \mathcal{X} (uniform LLN). Assuming

further that \mathcal{X} is compact, it is not difficult to show that the optimal value $\hat{\vartheta}_N$ and an optimal solution \hat{x}_N of the SAA problem converge to their true counterparts w.p.1 as $N \rightarrow \infty$ (see, e.g., [20, section 5.1.1.]).

It is also possible to derive rates of convergence. Let us make the following stronger assumptions.

(A3) For some point $x^* \in \mathcal{X}$ the expectation $\mathbb{E}[F(x^*, \xi)^2]$ is finite.

(A4) There exists a measurable function $C(\xi)$ such that $\mathbb{E}[C(\xi)^2]$ is finite and

$$|F(x, \xi) - F(x', \xi)| \leq C(\xi)\|x - x'\|, \quad \forall x, x' \in \mathcal{X}, \forall \xi \in \Xi.$$

Suppose further that the set \mathcal{X} is compact and consider Banach space $C(\mathcal{X})$ of continuous functions $\phi : \mathcal{X} \rightarrow \mathbb{R}$. Then \hat{f}_N can be viewed as a random element of $C(\mathcal{X})$, and $N^{1/2}(\hat{f}_N - f)$ converges in distribution to a random element $Y \in C(\mathcal{X})$. This is the so-called functional Central Limit Theorem (CLT) (e.g., [1]). By employing further an infinite dimensional Delta Theorem it is possible to derive the following result (cf., [17]).

Theorem 2.1. *Suppose that the set \mathcal{X} is compact and assumptions (A3) and (A4) hold. Then $N^{1/2}(\hat{f}_N - f)$ converges in distribution to a random element $Y \in C(\mathcal{X})$ and*

$$\hat{\vartheta}_N = \inf_{x \in \mathcal{S}} \hat{f}_N(x) + o_p(N^{-1/2}), \tag{2.3}$$

$$N^{1/2}(\hat{\vartheta}_N - \vartheta^*) \xrightarrow{\mathcal{D}} \inf_{x \in \mathcal{S}} Y(x). \tag{2.4}$$

If, moreover, $\mathcal{S} = \{\bar{x}\}$ is a singleton, then

$$N^{1/2}(\hat{\vartheta}_N - \vartheta^*) \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma^2), \tag{2.5}$$

where $\sigma^2 := \text{Var}[F(\bar{x}, \xi)]$.

The above result shows that the optimal value of the SAA problem converges to the optimal value of the true problem at a stochastic rate of $O_p(N^{-1/2})$. In particular, if the true problem has unique optimal solution \bar{x} , then $\hat{\vartheta}_N = \hat{f}_N(\bar{x}) + o_p(N^{-1/2})$, i.e., $\hat{\vartheta}_N$ converges to ϑ^* at the same asymptotic rate as $\hat{f}_N(\bar{x})$ converges to $f(\bar{x})$.

It is not difficult to show that $\mathbb{E}[\hat{\vartheta}_N] \leq \vartheta^*$ and $\mathbb{E}[\hat{\vartheta}_{N+1}] \geq \mathbb{E}[\hat{\vartheta}_N]$ (cf., [10]), i.e., $\hat{\vartheta}_N$ is a biased down estimate of ϑ^* and the bias is monotonically decreasing with increase of the sample size N . Note that for any fixed $x \in \mathcal{X}$ we have that $\mathbb{E}[\hat{f}_N(x)] = f(x)$ and hence $\mathbb{E}[Y(x)] = 0$, where $Y(x)$ is the random function specified in Theorem 2.1. Therefore if $\mathcal{S} = \{\bar{x}\}$ is a singleton, then the asymptotic bias of $\hat{\vartheta}_N$ is of order $o(N^{-1/2})$. On the other hand, if the true problem has more than one optimal solution, then the expected value of $\inf_{x \in \mathcal{S}} Y(x)$ typically will be strictly negative and hence the asymptotic bias will be of order $O(N^{-1/2})$.

In some situations the feasible set of stochastic program is also given in a form of expected value constraints. That is, consider the following problem

$$\text{Min}_{x \in \mathcal{X}} \{f(x) := \mathbb{E}[F(x, \xi)]\} \text{ subject to } g(x) \preceq_C 0, \tag{2.6}$$

where $C \subset \mathbb{R}^m$ is a closed convex cone and $g(x) := \mathbb{E}[G(x, \xi)]$ with $G : \mathcal{X} \times \Xi \rightarrow \mathbb{R}^m$. Note that constraint $g(x) \preceq_C 0$ means that $-g(x) \in C$. We assume that for every $x \in \mathbb{R}^n$ the expectation $g(x)$ is well defined and finite valued. Here in addition to the data of problem (2.1) we have constraints $g(x) \preceq_C 0$. For example if $C := \mathbb{R}_+^m$, then these constraints become $g_i(x) \leq 0, i = 1, \dots, m$, where $g_i(x)$ is the i -th component of the mapping $g(x)$. If $C := \mathbb{S}_+^m$ is the cone of $m \times m$ positive semidefinite matrices and $G(x, \xi)$ is an affine in x mapping, then these constraints become constraints of semidefinite programming. Given random sample ξ^1, \dots, ξ^N , the expected value mapping $g(x)$ can be approximated by the sample average $\hat{g}_N(x) := \frac{1}{N} \sum_{j=1}^N G(x, \xi^j)$, and hence the following SAA problem can be constructed

$$\text{Min}_{x \in \mathcal{X}} \hat{f}_N(x) \text{ subject to } \hat{g}_N(x) \preceq_C 0. \tag{2.7}$$

We say that problem (2.6) is convex if the set \mathcal{X} is convex, and for every $\xi \in \Xi$ the function $F(\cdot, \xi)$ is convex and the mapping $G(\cdot, \xi)$ is convex with respect to the cone C , i.e.,

$$G(tx + (1 - t)y, \xi) \preceq_C tG(x, \xi) + (1 - t)G(y, \xi), \quad \forall x, y \in \mathbb{R}^n, \forall t \in [0, 1]. \tag{2.8}$$

Note that the above condition (2.8) is equivalent to the condition that $\langle \lambda, G(x, \xi) \rangle$ is convex in x for every $\lambda \in C^*$. Note also that convexity of $F(\cdot, \xi)$ and $G(\cdot, \xi)$ imply convexity of the respective expected value functions.

Consider the Lagrangian $L(x, \lambda, \xi) := F(x, \xi) + \langle \lambda, G(x, \xi) \rangle$, and its expectation $\ell(x, \lambda) := \mathbb{E}[L(x, \lambda, \xi)]$ and sample average $\hat{\ell}_N(x, \lambda) := \hat{f}_N(x) + \langle \lambda, \hat{g}_N(x) \rangle$, associated with problem (2.6). The Lagrangian dual of problem (2.6) is the problem

$$\text{Max}_{\lambda \in C^*} \left\{ \psi(\lambda) := \min_{x \in \mathcal{X}} \ell(x, \lambda) \right\}. \tag{2.9}$$

It is said that the Slater condition for problem (2.6) holds if there exists a point $x^* \in \mathcal{X}$ such that $g(x^*) \prec_C 0$, i.e., $-g(x^*) \in \text{int}(C)$. If the problem is convex and the Slater condition holds, then the optimal values of problems (2.6) and (2.9) are equal to each other and the dual problem (2.9) has a nonempty and bounded set of optimal solutions, denoted Λ .

We can now formulate an analogue of the asymptotic result of Theorem 2.1 for convex problems of the form (2.6) (cf., [20, section 5.1.4]). We will need the following analogues of assumptions (A3) and (A4).

(A5) For some point $x^* \in \mathcal{X}$ the expectation $\mathbb{E} [\|G(x^*, \xi)\|^2]$ is finite.

(A6) There exists a measurable function $C(\xi)$ such that $\mathbb{E}[C(\xi)^2]$ is finite and

$$\|G(x, \xi) - G(x', \xi)\| \leq C(\xi)\|x - x'\|, \quad \forall x, x' \in \mathcal{X}, \quad \forall \xi \in \Xi.$$

As before we denote by ϑ^* and $\hat{\vartheta}_N$ the optimal values of the true and SAA problems (problems (2.6) and (2.7)), respectively.

Theorem 2.2. *Suppose that: problem (2.6) is convex, Slater condition holds, the set \mathcal{X} is compact and assumptions (A3) – (A6) are satisfied. Then*

$$\hat{\vartheta}_N = \inf_{x \in \mathcal{S}} \sup_{\lambda \in \Lambda} \hat{\ell}_N(x, \lambda) + o_p(N^{-1/2}). \tag{2.10}$$

If, moreover, $\mathcal{S} = \{\bar{x}\}$ and $\Lambda = \{\bar{\lambda}\}$ are singletons, then

$$N^{1/2}(\hat{\vartheta}_N - \vartheta^*) \xrightarrow{D} \mathcal{N}(0, \sigma^2), \tag{2.11}$$

where $\sigma^2 := \text{Var}[L(\bar{x}, \bar{\lambda}, \xi)]$.

There is an interesting consequence of the above result. It was assumed that in the SAA problem (2.7) the *same* sample ξ^1, \dots, ξ^N was used in constructing approximations $\hat{f}_N(x)$ and $\hat{g}_N(x)$ of the objective and constraints functions, and the asymptotic result (2.11) is formulated for that case. That is, the asymptotic variance σ^2 is given by $\text{Var}[L(\bar{x}, \bar{\lambda}, \xi)] = \text{Var}[F(\bar{x}, \xi) + \sum_{i=1}^m \bar{\lambda}_i G_i(\bar{x}, \xi)]$. In the Monte Carlo sampling approach we have a choice of estimating the objective function and each component of the constraint mapping $g(x)$ by using *independently* generated samples. In that case similar result holds but with the asymptotic variance given by $\text{Var}[F(\bar{x}, \xi)] + \sum_{i=1}^m \text{Var}[\bar{\lambda}_i G_i(\bar{x}, \xi)]$. Since it could be expected that the components $G_i(\bar{x}, \xi)$, $i = 1, \dots, m$, of the constraint mapping are positively correlated with each other, in order to reduce variability of the SAA estimates it would be advantageous to use the independent samples strategy.

3. Multistage Problems

The above analysis is performed for stochastic programs of a static form (2.1) and can be applied to two stage programming problems. What can be said in that respect for dynamic programs formulated in a multistage form? A solution of the multistage program (1.3) is a policy $\bar{x}_t = \bar{x}_t(\xi_{[t]})$, $t = 1, \dots, T$, given by measurable functions of the data process $\xi_{[t]} := (\xi_1, \dots, \xi_t)$ available at the decision time $t = 2, \dots, T$, with \bar{x}_1 being deterministic. It is said that policy is feasible if it satisfies the feasibility constraints for a.e. realization of the data process, i.e., $\bar{x}_1 \in \mathcal{X}_1$ and $\bar{x}_t \in \mathcal{X}_t(\bar{x}_{t-1}, \xi_t)$, $t = 2, \dots, T$, w.p.1.

The following dynamic programming equations can be written for the multistage program (1.3) going backward in time

$$Q_t(x_{t-1}, \xi_{[t]}) = \inf_{x_t \in \mathcal{X}_t(x_{t-1}, \xi_t)} \{f_t(x_t, \xi_t) + Q_{t+1}(x_t, \xi_{[t]})\}, \quad t = T, \dots, 2, \tag{3.1}$$

where $Q_{T+1}(x_T, \xi_{[T]}) \equiv 0$ by definition and

$$Q_{t+1}(x_t, \xi_{[t]}) := \mathbb{E} \{ Q_{t+1}(x_t, \xi_{[t+1]}) | \xi_{[t]} \}, \quad t = T - 1, \dots, 2, \quad (3.2)$$

are the respective cost-to-go functions. Finally at the first stage the following problem should be solved

$$\text{Min}_{x_1 \in \mathcal{X}_1} f_1(x_1) + \mathbb{E}[Q_2(x_1, \xi_2)]. \quad (3.3)$$

A policy $\bar{x}_t = \bar{x}_t(\xi_{[t]})$, $t = 1, \dots, T$, is optimal if w.p.1 it holds that

$$\bar{x}_t \in \arg \min_{x_t \in \mathcal{X}_t(\bar{x}_{t-1}, \xi_t)} \{ f_t(x_t, \xi_t) + Q_{t+1}(x_t, \xi_{[t]}) \}, \quad t = T, \dots, 2, \quad (3.4)$$

and \bar{x}_1 is an optimal solution of the first stage problem (3.3).

Problem (3.3) looks similar to the stochastic program (2.1). The difference, however, is that for $T \geq 3$ the function $Q_2(x_1, \xi_2)$ is not given explicitly, or at least in a computationally accessible form, but in itself is a solution of a multistage stochastic programming problem. Therefore in order to solve (1.3) numerically one would need to approximate the data process ξ_1, \dots, ξ_T by generating a tree of scenarios. The Monte Carlo sampling approach can be employed in the following way. First, a random sample $\xi_2^1, \dots, \xi_2^{N_1}$ of N_1 realizations of the random vector ξ_2 is generated. For each ξ_2^j , $j = 1, \dots, N_1$, a random sample of size N_2 of ξ_3 , according to the distribution of ξ_3 conditional on $\xi_2 = \xi_2^j$, is generated and so forth for later stages. We refer to this procedure as *conditional sampling*. In that way the true distribution of the random data process is discretized with every generated path of the process taken with equal probability. We refer to each generated path as scenario and to the collection of all scenarios as scenario tree. Note that the total number of scenarios $N = \prod_{t=1}^{T-1} N_t$. We denote $\mathcal{N} := \{N_1, \dots, N_{T-1}\}$ and by ϑ^* and $\hat{\vartheta}_{\mathcal{N}}$ the optimal values of the true problem (1.3) and the constructed SAA problem, respectively.

Assume for the sake of simplicity that the data process is *stagewise independent*, i.e., random vector ξ_{t+1} is distributed independently of $\xi_{[t]}$, $t = 1, \dots, T-1$. Then the cost-to-go functions $Q_{t+1}(x_t)$, $t = 1, \dots, T-1$, do not depend on the random data process. Also in that case there are two possible approaches to conditional sampling, namely for each ξ_2^j , $j = 1, \dots, N_1$, it is possible to generate different samples of ξ_3 independent of each other, or it is possible to use the same sample $\xi_3^1, \dots, \xi_3^{N_2}$, and so forth for later stages. We consider the second approach, which preserves the stagewise independence in the generated scenario tree, with respective samples $\xi_t^1, \dots, \xi_t^{N_{t-1}}$, at stages $t = 2, \dots, T$.

We can write dynamic programming equations for the constructed SAA problem. Eventually the (true) first stage problem (3.3) will be approximated by the following SAA problem

$$\text{Min}_{x_1 \in \mathcal{X}_1} f_1(x_1) + \hat{Q}_{2, N_1}(x_1), \quad (3.5)$$

where $\hat{Q}_{2,N_1}(x_1) = \frac{1}{N_1} \sum_{j=1}^{N_1} \tilde{Q}_2(x_1, \xi_2^j, \tilde{\xi})$. Here $\tilde{\xi} = (\xi_3^1, \dots, \xi_3^{N_2}, \dots, \xi_T^1, \dots, \xi_T^{N_{T-1}})$ is random vector composed from the samples at stages $t \geq 3$ and $\tilde{Q}_2(x_1, \xi_2, \tilde{\xi})$ is the corresponding cost-to-go function of the SAA problem. Note that function $\tilde{Q}_2(x_1, \xi_2, \tilde{\xi})$ depends on the random samples used at stages $t = 3, \dots, T$, as well.

Suppose now that the sample size N_1 tends to infinity while sample sizes N_t , $t = 2, \dots, T - 1$, are fixed. Then by the LLN we have that $\hat{Q}_{2,N_1}(x_1)$ converges (pointwise) w.p.1 to the function $\mathfrak{E}_2(x_1, \tilde{\xi}) := \mathbb{E}[\tilde{Q}_2(x_1, \xi_2, \tilde{\xi}) | \tilde{\xi}]$. Consider the problem

$$\text{Min}_{x_1 \in \mathcal{X}_1} f_1(x_1) + \mathfrak{E}_2(x_1, \tilde{\xi}). \tag{3.6}$$

Conditional on $\tilde{\xi}$ we can view problem (3.5) as the SAA problem associated with the (static) expected value problem (3.6). Consequently asymptotic results of section 2 can be applied to the pair of problems (3.5) and (3.6).

Denote by $\vartheta^*(\tilde{\xi})$ the optimal value of problem (3.6), and recall that $\hat{\vartheta}_N$ denotes the optimal value of problem (3.5). We have that conditional on $\tilde{\xi}$, the SAA optimal value $\hat{\vartheta}_N$ is a biased down estimate of $\vartheta^*(\tilde{\xi})$. Since $\mathfrak{E}_2(x_1, \tilde{\xi})$ is an SAA estimate of $Q_2(x_1)$, we also have that $\mathbb{E}[\mathfrak{E}_2(x_1, \tilde{\xi})] \leq Q_2(x_1)$ for every $x_1 \in \mathcal{X}_1$. It follows that $\mathbb{E}[\vartheta^*(\tilde{\xi})] \leq \vartheta^*$. Consequently the bias of the SAA estimate $\hat{\vartheta}_N$, of the optimal value ϑ^* of the true multistage problem (1.3), will increase with increase of the number of stages. It is possible to show that for some models this bias growth exponentially with increase of the number T of stages (cf., [20, p.225]).

In order for the SAA problems to converge to the true problem all samples should be increased, i.e., all sample sizes N_t should tend to infinity. In the next section we will discuss estimates of sample sizes required to solve the true problem with a given accuracy.

4. Estimates of Stochastic Complexity

In order to solve a stochastic optimization problem of the form (2.1) one needs to evaluate expectations $\mathbb{E}[F(x, \xi)]$, given by multidimensional integrals. This, in turn, requires a discretization of (continuous) distribution of the random vector ξ . Suppose that the components of ξ are distributed independently of each other and that r points are used for discretization of the marginal distribution of each component. Then the total number of discretization points (scenarios) is r^d . That is, while the input data is proportional to rd and grows linearly with increase of the number d of random parameters, the number of scenarios increases exponentially. This indicates that deterministic optimization algorithms cannot cope with such stochastic optimization problems. And, indeed, it is shown in [5] that, under the assumption that the stochastic parameters are independently distributed, two-stage linear stochastic programming problems are #P-hard.

The Monte Carlo sampling approach of approximating the true problem (2.1) by the corresponding SAA problem (2.2) suggests a randomization approach to solving stochastic optimization problems. In a sense the sample size N , required to solve the true problem with a given accuracy, gives an estimate of computational complexity of the considered problem. Note that the SAA approach is not an algorithm, one still needs to solve the constructed SAA problem. Numerical experiments indicate that for various classes of problems, e.g., two stage linear stochastic programs, computational effort in solving SAA problems by efficient algorithms is more or less proportional to the sample size N . Theorem 2.1 suggests that the convergence of SAA estimates is rather slow. However, the convergence does not depend directly on dimension d of the random data vector, but rather on variability of the objective function.

We proceed now to estimation of the sample size required to solve the true problem with a given accuracy $\varepsilon > 0$. Recall that it is assumed that the expectation $f(x)$ is well defined and finite valued for all $x \in \mathcal{X}$. It is said that a point $\bar{x} \in \mathcal{X}$ is an ε -optimal solution of problem (2.1) if $f(\bar{x}) \leq \inf_{x \in \mathcal{X}} f(x) + \varepsilon$. We denote by \mathcal{S}^ε and $\hat{\mathcal{S}}_N^\varepsilon$ the sets of ε -optimal solutions of the true and SAA problems (2.1) and (2.2), respectively. Let us make the following assumptions.

(C1) There exist constants $\sigma > 0$ and $\tau > 0$ such that

$$M_{x,x'}(t) \leq \exp(\sigma^2 t^2 / 2), \quad \forall t \in [-\tau, \tau], \forall x, x' \in \mathcal{X}, \quad (4.1)$$

where $M_{x,x'}(t)$ is the moment generating function of the random variable $[F(x', \xi) - f(x')] - [F(x, \xi) - f(x)]$.

(C2) There exists a measurable function $\kappa : \Xi \rightarrow \mathbb{R}_+$ such that its moment generating function $M_\kappa(t)$ is finite valued for all t in a neighborhood of zero and

$$|F(x, \xi) - F(x', \xi)| \leq \kappa(\xi) \|x - x'\|, \quad \forall x, x' \in \mathcal{X}, \forall \xi \in \Xi. \quad (4.2)$$

By Cramér's Large Deviations Theorem it follows from assumption (C2) that for any $L > \mathbb{E}[\kappa(\xi)]$ there is a positive constant $\beta = \beta(L)$ such that

$$\text{Prob}(\hat{\kappa}_N > L) \leq \exp(-N\beta), \quad (4.3)$$

where $\hat{\kappa}_N := N^{-1} \sum_{j=1}^N \kappa(\xi^j)$. The following estimate of the sample size is obtained by applying (pointwise) upper bound of Cramér's Large Deviations Theorem and constructing a ν -net in \mathcal{X} with number of points less than $(\varrho D / \nu)^n$, where $D := \sup_{x, x' \in \mathcal{X}} \|x' - x\|$ is the diameter of the set \mathcal{X} and $\varrho > 0$ is an appropriate constant (cf., [18],[20, section 5.3.2]).

Theorem 4.1. *Suppose that the set \mathcal{X} has a finite diameter D and assumptions (C1) – (C2) hold with respective constants σ and τ , and let $\alpha \in (0, 1)$, $L >$*

$\mathbb{E}[\kappa(\xi)]$, $\beta = \beta(L)$ and $\varepsilon > 0$, $\delta > 0$ be constants such that $\varepsilon > \delta \geq 0$ and $\varepsilon - \delta \leq \tau\sigma^2$. Then for the sample size N satisfying

$$N \geq \beta^{-1} \ln(2/\alpha) \quad \text{and} \quad N \geq \frac{8\sigma^2}{(\varepsilon - \delta)^2} \left[n \ln \left(\frac{8\rho LD}{\varepsilon - \delta} \right) + \ln \left(\frac{2}{\alpha} \right) \right], \quad (4.4)$$

it follows that

$$\text{Prob}(\hat{\mathcal{S}}_N^\delta \subset \mathcal{S}^\varepsilon) \geq 1 - \alpha. \quad (4.5)$$

In particular, if in (4.2) the function $\kappa(\xi) \equiv L$, i.e., the Lipschitz constant of $F(\cdot, \xi)$ does not depend on ξ , then the first condition in the sample estimate (4.4) can be omitted and the constant σ^2 can be replaced by the estimate $2L^2D^2$. The assertion (4.5) of the above theorem means that if $\bar{x} \in \mathcal{X}$ is a δ -optimal solution of the SAA problem with the sample size N satisfying (4.4), then \bar{x} is an ε -optimal solution of the true problem with probability $\geq 1 - \alpha$. That is, by solving the SAA problem with accuracy $\delta < \varepsilon$, say $\delta := \varepsilon/2$, we are guaranteed with confidence $1 - \alpha$ that we solve the true problem with accuracy ε . Similar estimates of the sample size can be obtained by using theory of Vapnik-Chervonenkis (VC) dimension (cf., [21]).

The above estimate of the sample size is theoretical and typically is too conservative for practical applications. Nevertheless it leads to several important conclusions. From this point of view the number of scenarios in a formulation of the true problem is irrelevant and can be infinite, while the computational difficulty is influenced by variability of the objective function which, in a sense, measured by the constant σ^2 . It also suggests that the required sample size is proportional to ε^{-2} . Such dependence of the sample size on required accuracy is typical for Monte Carlo sampling methods and cannot be changed. Similar rates of convergence can be derived for the optimal value of the SAA problem. Central Limit Theorem type result of Theorem 2.1 confirms this from another point of view. In some situations quasi-Monte Carlo methods can enhance rates of convergence (cf., [6]), but in principle it is impossible to evaluate multidimensional integrals with a high accuracy. On the other hand dependence on the confidence $1 - \alpha$ is logarithmic, e.g., increasing the confidence say from 90% to 99.99% does not require a big change of the sample size.

For well conditioned problems it is possible to derive better rates of convergence. It is said that a γ -order growth condition, with $\gamma \geq 1$, holds for the true problem if its set \mathcal{S} of optimal solutions is nonempty and there is constant $c > 0$ such that

$$f(x) \geq \vartheta^* + c[\text{dist}(x, \mathcal{S})]^\gamma \quad (4.6)$$

for all $x \in \mathcal{X}$ in a neighborhood of \mathcal{S} . Of interest is the growth condition of order $\gamma = 1$ and $\gamma = 2$. If $\mathcal{S} = \{\bar{x}\}$ is a singleton and the first-order growth condition holds, the optimal solution \bar{x} is referred to as *sharp*. For convex problems satisfying the second order growth condition the sample size estimate becomes of order $O(\varepsilon^{-1})$. In convex case of sharp optimal solution \bar{x} the convergence is finite, in the sense that w.p.1 for N large enough the SAA problem has unique

optimal solution \bar{x} coinciding with the optimal solution of the true problem and, moreover, the probability of this event approaches one exponentially fast with increase of N (see [20, p.190] for a discussion and exact formulation).

5. Multistage Complexity

Consider the multistage setting of section 3. Recall that the optimal value ϑ^* of the multistage problem (1.3) is given by the optimal value of the problem (3.3) and $\mathcal{Q}_2(x_1) = \mathbb{E}[Q_2(x_1, \xi_2)]$. Similarly to the static case we say that $\bar{x}_1 \in \mathcal{X}_1$ is an ε -optimal solution of the first stage of the true problem (1.3) if $f_1(\bar{x}_1) + \mathcal{Q}_2(\bar{x}_1) \leq \vartheta^* + \varepsilon$. Suppose for the moment that $T = 3$. Then under regularity conditions similar to the static case it is possible to derive the following estimate of the sample sizes N_1 and N_2 needed to solve the first stage problem with a given accuracy $\varepsilon > 0$ and confidence $1 - \alpha$ while solving the SAA problem with accuracy, say, $\varepsilon/2$ (see [20, section 5.8.2] for technical details).

For constants $\varepsilon > 0$ and $\alpha \in (0, 1)$ and sample sizes N_1 and N_2 satisfying

$$\left[\frac{O(1)D_1L_1}{\varepsilon} \right]^{n_1} \exp \left\{ -\frac{O(1)N_1\varepsilon^2}{\sigma_1^2} \right\} + \left[\frac{O(1)D_2L_2}{\varepsilon} \right]^{n_2} \exp \left\{ -\frac{O(1)N_2\varepsilon^2}{\sigma_2^2} \right\} \leq \alpha, \quad (5.1)$$

we have that any $(\varepsilon/2)$ -optimal solution of the first stage of the SAA problem is an ε -optimal solution of the first stage of the true problem with probability at least $1 - \alpha$. Here $O(1)$ denotes a generic constant independent of the data and $\sigma_1, \sigma_2, D_1, D_2$ and L_1, L_2 are certain analogues of the constants of the estimate (4.4).

In particular suppose that $N_1 = N_2$ and let $n := \max\{n_1, n_2\}$, $L := \max\{L_1, L_2\}$, $D := \max\{D_1, D_2\}$. Then (5.1) becomes

$$N_1 \geq \frac{O(1)\sigma^2}{\varepsilon^2} \left[n \ln \left(\frac{O(1)LD}{\varepsilon} \right) + \ln \left(\frac{1}{\alpha} \right) \right]. \quad (5.2)$$

The above estimate looks similar to the estimate (4.4) of the two stage program. Note, however, that in the present case of three stage program the total number of scenarios of the SAA problem is $N = N_1^2$. This analysis can be extended to a larger number of stages with the conclusion that the total number of scenarios needed to solve the true problem with a given accuracy grows *exponentially* with increase of the number T of stages. Another way of putting this is that the number of scenarios needed to solve T -stage problem (1.3) would grow as $O(\varepsilon^{-2(T-1)})$ with decrease of the error level $\varepsilon > 0$. This indicates that from the point of view of the number of scenarios, complexity of multistage programming problems grows exponentially with the number of stages. Furthermore, as it was pointed in the Introduction, even if the SAA problem can be solved, its solution does not define a policy for the true problem and of use may be only

the computed first stage solution. There are even deeper reasons to believe that (even linear) multistage stochastic programs are computationally intractable (cf., [19]). This does not mean, of course, that some specific classes of multistage stochastic programs cannot be solved efficiently.

6. Approximations of Multistage Stochastic Programs

If multistage stochastic programming problems cannot be solve to optimality, one may think about approximations. There are several possible approaches to trying to solve multistage stochastic programs approximately. One approach is to reduce dynamic setting to a static case. Suppose that we can identify a parametric family of policies $\bar{x}_t(\xi_{[t]}, \theta_t)$, $t = 1, \dots, T$, depending on a finite number of parameters $\theta_t \in \Theta_t \subset \mathbb{R}^{q_t}$, and such that these policies are feasible for all parameter values. That is, for all $\theta_t \in \Theta_t$, $t = 1, \dots, T$, it holds that $\bar{x}_1(\theta_1) \in \mathcal{X}_1$ and $\bar{x}_t(\xi_{[t]}, \theta_t) \in \mathcal{X}_t(\bar{x}_{t-1}(\xi_{[t-1]}, \theta_{t-1}), \xi_t)$, $t = 2, \dots, T$, w.p.1. Consider the following stochastic program

$$\begin{aligned} \text{Min}_{\theta_1, \dots, \theta_T} \quad & f_1(\bar{x}_1(\theta_1)) + \mathbb{E} \left[\sum_{t=2}^T f_t(\bar{x}_t(\xi_{[t]}, \theta_t), \xi_t) \right] \\ \text{s.t.} \quad & \theta_t \in \Theta_t, \quad t = 1, \dots, T. \end{aligned} \tag{6.1}$$

The above problem (6.1) is a (static) stochastic problem of the form (2.1) and could be solved, say by the SAA method, provided that the sets Θ_t are defined in a computationally accessible way. Of course, quality of an obtained solution $\bar{x}_t(\xi_{[t]}, \theta_t^*)$, $t = 1, \dots, T$, viewed as a solution of the original multistage problem (1.3), depends on a successful choice of the parametric family.

Suppose, for example, that we have a finite family of feasible policies $\{x_t^k(\xi_{[t]}), t = 1, \dots, T\}$, $k = 1, \dots, K$. Suppose, further, that the multifunctions $\mathcal{X}_t(\cdot, \xi_t)$ are convex, i.e., the set \mathcal{X}_1 is convex and for a.e. ξ_t and all x_{t-1}, x'_{t-1} and $\tau \in [0, 1]$ it holds that

$$\tau \mathcal{X}_t(x_{t-1}, \xi_t) + (1 - \tau) \mathcal{X}_t(x'_{t-1}, \xi_t) \subset \mathcal{X}_t(\tau x_{t-1} + (1 - \tau)x'_{t-1}, \xi_t), \quad t = 2, \dots, T.$$

Then any convex combination

$$\bar{x}_t(\xi_{[t]}, \theta) := \sum_{k=1}^K \theta_k x_t^k(\xi_{[t]}), \quad t = 1, \dots, T,$$

where $\theta = (\theta_1, \dots, \theta_K) \in \Delta_K$ with Δ_K being K -dimensional simplex, of these policies is feasible. This approach with several examples was discussed in [8].

As another example consider linear multistage stochastic programs with fixed recourse. That is, assume the setting of (1.4)–(1.5) with only the right hand sides vectors b_t , $t = 2, \dots, T$, being random. Moreover, for the sake of

simplicity assume that the data process b_1, \dots, b_T , is stagewise independent with distribution of random vector b_t supported on set Ξ_t , $t = 2, \dots, T$. Motivated by its success in robust optimization it was suggested in [19] to use affine decision rules. That is, consider policies of the form

$$\bar{x}_t = \phi_t + \sum_{\tau=2}^t \Phi_{t\tau} b_\tau, \quad t = 2, \dots, T, \quad (6.2)$$

depending on parameters – vectors ϕ_t and matrices $\Phi_{t\tau}$. The feasibility constraints here take the form

$$\begin{aligned} A_1 x_1 &\leq b_1, \quad B_2 x_1 + A_2(\phi_2 + \Phi_{22} b_2) \leq b_2, \\ B_t(\phi_{t-1} + \sum_{\tau=2}^{t-1} \Phi_{t-1,\tau} b_\tau) + A_t(\phi_t + \sum_{\tau=2}^t \Phi_{t\tau} b_\tau) &\leq b_t \quad t = 3, \dots, T, \end{aligned} \quad (6.3)$$

and should hold for every $b_t \in \Xi_t$, $t = 2, \dots, T$ (we can pass here from the condition “for a.e.” to “for every” by continuity arguments). The system (6.3), defining feasible parameters of the policy (6.2), involves an infinite number of linear constraints. In case the sets Ξ_t are polyhedral, defined by a finite number of linear constraints, it is possible to handle the semi-infinite system (6.3) in a computationally efficient way (cf., [19]).

An alternative approach to solving multistage program (1.3) is to approximate dynamic programming equations (3.1). One such approach can be described as follows. Consider the linear setting (1.4)–(1.5) and assume that the stagewise independence condition holds. In that case the cost-to-go functions $\mathcal{Q}_t(x_{t-1})$, $t = 2, \dots, T$, are convex and do not depend on the random data. Consider the corresponding SAA problem based on (independent) samples $\xi_t^1, \dots, \xi_t^{N_t-1}$, $t = 2, \dots, T$. By the above analysis we have (under mild regularity conditions) that if all sample sizes are of the same order, say all $N_t = M$, $t = 1, \dots, T-1$, then in order to solve the first stage problem with accuracy $\varepsilon > 0$ we need M to be of order $O(\varepsilon^{-2})$. Of course, even for moderate values of M , say $M = 100$, the total number of scenarios $N = M^{T-1}$ quickly becomes astronomically large with increase of the number of stages. Therefore, instead of solving the corresponding linear programming problem involving all scenarios, one can try to approximate the cost-to-go functions of the SAA problem.

For a given set of samples of size $\mathcal{N} = (N_1, \dots, N_{T-1})$, let $\tilde{\mathcal{Q}}_{t,\mathcal{N}}(x_{t-1})$, $t = 2, \dots, T$, be cost-to-go functions of the SAA problem. These functions are convex piecewise linear and do not depend on realizations of scenarios from the SAA scenario tree. Suppose that we have a procedure for generating cutting (supporting) planes for the SAA cost-to-go functions. By taking maximum of respective collections of these cutting planes we can construct piecewise linear convex functions $\tilde{\mathcal{Q}}_t(x_{t-1})$ approximating the SAA cost-to-go functions from below, i.e., $\tilde{\mathcal{Q}}_{t,\mathcal{N}}(\cdot) \geq \mathcal{Q}_t(\cdot)$, $t = 2, \dots, T$. These functions $\tilde{\mathcal{Q}}_t(x_{t-1})$ and a feasible first stage solution \bar{x}_1 define the following policy:

$$\bar{x}_t \in \arg \min \{ \langle c_t, x_t \rangle + \tilde{\mathcal{Q}}_{t+1}(x_t) : A_t x_t \leq b_t - B_t \bar{x}_{t-1} \}, \quad t = 2, \dots, T, \quad (6.4)$$

with $\mathfrak{Q}_{T+1}(x_T) \equiv 0$ by definition. This policy can be applied to the true multistage problem and to its sample average approximation. In both cases the policy is feasible by the construction and hence its expected value gives an upper bound for the optimal value of the corresponding multistage program. The expected value of this policy can be estimated by sampling.

Since functions $\mathfrak{Q}_t(\cdot)$ are given as maximum of a finite number of affine functions, the optimization problems in the right hand side of (6.4) can be formulated as linear programming problems of reasonable sizes. It was suggested in [14] to generate trial decision points \bar{x}_t using randomly generated sample paths in a forward step procedure of the form (6.4) and consequently to add cutting planes, computed at these trial decision points, to approximations $\mathfrak{Q}_t(\cdot)$ in a backward step procedure. The required cutting planes are constructed by solving duals of the linear programming problems associated with right hand side of (6.4). This algorithm, called Stochastic Dual Dynamic Programming (SDDP), became popular in energy planning procedures. It is possible to show that under mild regularity conditions, functions $\mathfrak{Q}_t(\cdot)$ converge w.p.1 to their counterparts $\hat{Q}_{t,N}(\cdot)$ of the SAA problem, and hence policy (6.4) converges to an optimal policy of the SAA problem (cf., [15]). The convergence can be slow however.

For two stage programs the SDDP algorithm becomes Kelley's cutting plane algorithm, [7]. Worst case analysis of Kelley's algorithm is discussed in [13, pp. 158-160], with an example of a problem where an ε -optimal solution cannot be obtained by this algorithm in less than $(\frac{1}{2 \ln 2}) 1.15^{n-1} \ln(\varepsilon^{-1})$ calls of the oracle, i.e., the number of oracle calls grows exponentially with increase of the dimension n of the problem. It was also observed empirically that Kelley's algorithm could behave quite poorly in practice.

On the other hand, complexity of one run of the forward and backward steps of the SDDP algorithm grows linearly with increase of the number of stages and the algorithm produces a feasible and implementable policy.

7. Concluding Remarks

So far we discussed computational complexity from the point of view of the required number of scenarios. It should be remembered that a constructed SAA problem still needs to be solved by an appropriate deterministic algorithm. Consider for example the SAA problem associated with two stage linear problem (1.1). In order to compute a subgradient of the respective sample average function $\hat{Q}_N(x) = \frac{1}{N} \sum_{j=1}^N Q(x, \xi^j)$ at an iteration point of a subgradient type algorithmic procedure, one would need to solve N second stage problems together with their duals.

For *convex* (static) stochastic problems an alternative to the SAA approach is the Stochastic Approximation (SA) method going back to Robbins and Monro

[16]. The classical SA algorithm generates iterates for solving problem (2.1) by the formula

$$x_{j+1} = \Pi_{\mathcal{X}}(x_j - \gamma_j G(x_j, \xi^j)), \quad j = 1, \dots, \quad (7.1)$$

where $G(x, \xi) \in \partial_x F(x, \xi)$ is a subgradient of $F(x, \xi)$, $\Pi_{\mathcal{X}}$ is the metric projection onto the set \mathcal{X} and $\gamma_j > 0$ are chosen stepsizes. The standard choice of the stepsizes is $\gamma_j = \theta/j$ for some constant $\theta > 0$. For an *optimal* choice of the constant θ the estimates of rates of convergence of this method are similar to the respective estimates of the SAA method. However, the method is very sensitive to the choice of the constant θ and often does not work well in practice. It is somewhat surprising that a robust version of the SA algorithm, taking its origins in the mirror descent method of Nemirovski and Yudin [11], can significantly outperform SAA based algorithms for certain classes of convex stochastic problems (cf., [12]).

Theoretical estimates of the form (4.4), of the required sample size, are too conservative for practical applications. In that respect we may refer to [10] and [9] for a discussion of computational methods for evaluating quality of solutions of the first stage of two stage stochastic problems.

References

- [1] Araujo, A. and Giné, E., *The Central Limit Theorem for Real and Banach Valued Random Variables*. Wiley, New York, 1980.
- [2] Beale, E.M.L., On minimizing a convex function subject to linear inequalities, *Journal of the Royal Statistical Society, Series B*, **17** (1955), 173–184.
- [3] Ben-Tal, A., El Ghaoui, L. and Nemirovski, A., *Robust Optimization*, Princeton University Press, Princeton, 2009.
- [4] Dantzig, G.B., Linear programming under uncertainty, *Management Science*, **1** (1955), 197–206.
- [5] Dyer, M. and Stougie, L., Computational complexity of stochastic programming problems, *Mathematical Programming*, **106** (2006), 423–432.
- [6] Homem-de Mello, T., On rates of convergence for stochastic optimization problems under non-independent and identically distributed sampling, *SIAM J. Optim.*, **19** (2008), 524–551.
- [7] Kelley, J.E., The cutting-plane method for solving convex programs, *Journal of the Society for Industrial and Applied Mathematics*, **8** (1960), 703–712.
- [8] Koivu, M. and Pennanen, T., Galerkin methods in dynamic stochastic programming, *Optimization*, to appear.
- [9] Lan, G., Nemirovski, A. and Shapiro, A., Validation analysis of mirror descent stochastic approximation method, *E-print available at: <http://www.optimization-online.org>*, 2008.
- [10] Mak, W.K., Morton, D.P. and Wood, R.K., Monte Carlo bounding techniques for determining solution quality in stochastic programs, *Operations Research Letters*, **24** (1999), 47–56.

-
- [11] Nemirovski, A. and Yudin, D., *Problem Complexity and Method Efficiency in Optimization*, Wiley-Intersci. Ser. Discrete Math. 15, John Wiley, New York, 1983.
- [12] Nemirovski, A., Juditsky, A., Lan, G. and Shapiro, A., Robust stochastic approximation approach to stochastic programming, *SIAM Journal on Optimization*, **19** (2009), 1574–1609.
- [13] Nesterov, Yu., *Introductory Lectures on Convex Optimization*, Kluwer, Boston, 2004.
- [14] Pereira, M.V.F. and Pinto, L.M.V.G., Multi-stage stochastic optimization applied to energy planning, *Mathematical Programming*, **52** (1991), 359–375.
- [15] Philpott, A.B. and Guan, Z., On the convergence of stochastic dual dynamic programming and related methods, *Operations Research Letters*, **36** (2008), 450–455.
- [16] Robbins, H. and Monro, S., A stochastic approximation method. *Annals of Math. Stat.*, **22** (1951), 400–407.
- [17] Shapiro, A., Asymptotic analysis of stochastic programs. *Annals of Operations Research*, **30** (1991), 169–186.
- [18] Shapiro, A., Monte Carlo approach to stochastic programming. In B.A. Peters, J.S. Smith, D.J. Medeiros and M.W. Rohrer, editors, *Proceedings of the 2001 Winter Simulation Conference*, pp. 428–431, 2001.
- [19] Shapiro, A. and Nemirovski, A., On complexity of stochastic programming problems, in: *Continuous Optimization: Current Trends and Applications*, pp. 111–144, V. Jeyakumar and A.M. Rubinov (Eds.), Springer, 2005.
- [20] Shapiro, A., Dentcheva, D. and Ruszczyński, A., *Lectures on Stochastic Programming: Modeling and Theory*, SIAM, Philadelphia, 2009.
- [21] Vidyasagar, M., Randomized algorithms for robust controller synthesis using statistical learning theory, *Automatica*, **37** (2001), 1515–1528.

A Cutting Plane Theory for Mixed Integer Optimization

Robert Weismantel*

Abstract

From a practical perspective, mixed integer optimization represents a very powerful modeling paradigm. Its modeling power, however, comes with a price. The presence of both integer and continuous variables results in a significant increase in complexity over the pure integer case with respect to geometric, algebraic, combinatorial and algorithmic properties. Specifically, the theory of cutting planes for mixed integer linear optimization is not yet at a similar level of development as in the pure integer case. The goal of this paper is to discuss four research directions that are expected to contribute to the development of this field of optimization. In particular, we examine a new geometric approach based on lattice point free polyhedra and use it for developing a cutting plane theory for mixed integer sets. We expect that these novel developments will shed some light on the additional complexity that goes along with mixing discrete and continuous variables.

Mathematics Subject Classification (2000). Primary, 90C11; Secondary, 90C10.

Keywords. Mixed-integer, cutting plane, lattice point free convex sets

1. Mixed Integer Cutting Planes and Lattice Point Free Sets

The purpose of this section is to establish a link between the generation of cutting planes for mixed integer linear optimization problems and the theory of lattice point free polyhedra. When we refer to a mixed integer (linear) optimization problem, we mean an optimization problem of the form

$$\max a^T x + g^T y \text{ subject to } (x, y) \in P_I = \{ (x, y) \in \mathbb{Z}^n \times \mathbb{R}^d, Ax + Gy \geq b \},$$

*Department of Mathematics, IFOR, ETH Zentrum HG G12, CH-8092 Zürich, Switzerland. E-mail: robert.weismantel@ifor.math.ethz.ch.

where all input data a, g, A, G, b is assumed to be rational. The point of departure for studying mixed integer linear sets is the fact that for a polyhedron $P \subseteq \mathbb{R}^{n+d}$ with feasible mixed integer set $P_I = P \cap (\mathbb{Z}^n \times \mathbb{R}^d)$, the set $\text{conv}(P_I)$ is a polyhedron. Hence,

$$\begin{aligned} c^* &= \max \{a^T x + g^T y \mid (x, y) \in P_I\} \\ &= \max \{a^T x + g^T y \mid (x, y) \in \text{conv}(P_I)\}. \end{aligned}$$

Valid linear inequalities for $\text{conv}(P_I)$ are called cutting planes. In the pure integer case, i.e., the case of $d = 0$, geometric principles such as the rounding of hyperplanes [11] and the lift-and-project approach [6] lead directly to finite cutting plane procedures [14]. These questions remain challenging in the presence of both discrete and continuous variables. The purpose of this paper is to develop a link between the generation of cutting planes for mixed integer optimization and lattice point free convex sets.

Definition 1 (Lattice point free convex sets). *Let $L \subseteq \mathbb{R}^n$ be a convex set. If $\text{relint}(L) \cap \mathbb{Z}^n = \emptyset$, then L is called lattice point free.*

It turns out that lattice point free convex sets that are maximal with respect to inclusion are polyhedra. For rational polyhedra that are maximal with respect to inclusion it can be shown that each of their facets contains relative interior integer points [17]. Moreover, a maximally lattice point free rational polyhedron $L \subseteq \mathbb{R}^n$ is always full dimensional and can be represented as the Minkowski sum of a polytope \mathcal{P} and a linear space \mathcal{S} satisfying the following codimension property:

$$1 \leq \dim(\mathcal{P}) \leq n \text{ and } \dim(\mathcal{S}) = n - \dim(\mathcal{P}). \quad (1)$$

This fact suggests to associate a *split dimension* with every maximally lattice point free rational polyhedron.

Definition 2 (Split dimension, split polyhedron). *Let L be a lattice point free rational polyhedron. If L satisfies the codimension property Eq. (1), then it is called a split polyhedron. Its split dimension is denoted by $\text{dims}(L)$ and defined to be the dimension of the polytope \mathcal{P} in Eq. (1).*

Numerous papers are available that deal with properties of lattice-point-free convex sets that we will not discuss here. For a survey on the topic we refer to [17]. For several results we need the following size measure of lattice point free polyhedra. These measures are standard terminology in the theory of lattice point free convex sets [17].

Definition 3 (Lattice width, max-facet-width). *For a split polyhedron $L \subseteq \mathbb{R}^n$ and a vector $v \in \mathbb{Z}^n$, the width of L along v is defined to be the number*

$$w(L, v) := \max_{x \in L} v^T x - \min_{x \in L} v^T x.$$

The lattice width of L is

$$\min\{w(L, v) \mid v \in \mathbb{Z}^n \setminus \{0\}\}.$$

The max-facet-width $w_f(L)$ of L is defined to be the largest of the numbers $w(L, \pi)$ over all facet defining inequalities $\pi^T x \geq \pi_0$ for L using primitive integer data for the vector π .

Note that the max-facet-width measures how wide a split polyhedron is parallel to its facets.

A first simple link between cutting planes for a mixed integer linear program and lattice point free convex sets is made precise below. Here and throughout the paper $\text{proj}_x(F)$ denotes the projection of a set $F \subseteq \mathbb{R}^{n+d}$ to the space of x variables in dimension n .

Proposition 1. *Any valid inequality $a^T x + g^T y \geq \gamma$ for $\text{conv}(P_I)$ induces a lattice point free polyhedron*

$$L = \text{proj}_x(\{(x, y) \in P, a^T x + g^T y \leq \gamma\}).$$

This statement is easily verified noting that if $x \in \text{relint}(L) \cap \mathbb{Z}^n$, then there exists y such that $(x, y) \in P_I$ and $a^T x + g^T y < \gamma$. The existence of such a y contradicts the validity of $a^T x + g^T y \geq \gamma$ for $\text{conv}(P_I)$.

As a next step we deal with the converse direction. More precisely, given a lattice point free polyhedron L , we formally define how to generate mixed integer cuts from L . Let $P \subseteq \mathbb{R}^{n+d}$ be a polyhedron, where n denotes the number of integer variables and d the number of continuous variables. Let $L^x \subseteq \mathbb{R}^n$ be a split polyhedron. Its associated *mixed integer split polyhedron* is defined to be

$$L = \{(x, y) \in \mathbb{R}^{n+d} \mid x \in L^x\}.$$

If it is clear from the context we will sometimes drop the term “mixed integer”, and simply call also L a split polyhedron. Finally, note that $\text{relint}(L^x)$ is lattice point free in \mathbb{R}^n if and only if $\text{relint}(L)$ is lattice point free in \mathbb{R}^{n+d} . For a mixed integer split polyhedron L we next define as

$$R(L, P) = \text{conv}(P \setminus \text{relint}(L)), \quad (2)$$

the operation of adding cuts to P from the lattice point free polyhedron L .

The following remark summarizes the main properties of this operation. In particular, it shows that $R(L, P)$ is a valid relaxation of $\text{conv}(P_I)$.

Proposition 2. *(Basis properties of $R(L, P)$) [1]*

- $R(L, P) \cap (\mathbb{Z}^n \times \mathbb{R}^d) = P_I$.
- $R(L, P)$ is a polyhedron.

- $R(L, P) \neq P$ if and only if there exists a vertex v of P such that $v \in \text{relint}(L)$.
- The recession cone of $R(L, P)$ equals the recession cone of P , unless $R(L, P) = \emptyset$.
- A vertex u of $R(L, P)$ is an intersection point, i.e., there exists $0 \leq \lambda \leq 1$ and vertices v, w of P such that
 - (a) $u = \lambda v + (1 - \lambda)w$;
 - (b) u belongs to the boundary of L and
 - (c) $v \in \text{relint}(L)$ and $w \notin \text{relint}(L)$.

Observe that the outer description of $R(L, P)$ may have several new inequalities that are not part of the description of P . Precisely those new inequalities are the cutting planes that one can generate from the lattice point free polyhedron L . This raises the following basic question. Which lattice point free polyhedra L should be chosen so as to generate strong cuts? It is not clear at all how to answer this question in general. In fact, precise statements can be made only if one imposes some more structure on the feasible domain P_I . This topic will be discussed later.

2. Complexity and Closures of Split Polyhedra

Let us next introduce the question of how to measure complexity of cutting planes derived from lattice point free polyhedra. We will argue that split polyhedra can be used to derive so-called cutting plane proofs for a mixed integer linear set. We consider a polyhedron in \mathbb{R}^{n+d} of the form

$$P := \text{conv}(\{v^i\}_{i \in V}) + \text{cone}(\{r^j\}_{j \in E}), \tag{3}$$

where V and E are finite index sets, $\{v^i\}_{i \in V}$ denotes the vertices of P and $\{r^j\}_{j \in E}$ denotes the extreme rays of P . We assume P is rational, i.e., we assume $\{r^j\}_{j \in E} \subset \mathbb{Z}^{n+d}$ and $\{v^i\}_{i \in V} \subset \mathbb{Q}^{n+d}$.

Throughout this section we use the notation introduced in Section 1. In particular, $L \subseteq \mathbb{R}^{n+d}$ is the notation for a mixed integer split whose associated split polyhedron L^x has a Minkowski decomposition as in Eq. (1) with split dimension $\text{dim}_s(L^x) \leq n$, see also Eq. (2) and Proposition 2 for basic properties of the operator $R(L, P)$. Like in the previous section, we denote $P \cap (\mathbb{Z}^n \times \mathbb{R}^d)$ by P_I .

Definition 4 (Closure). *A finite family \mathcal{M} of mixed integer split polyhedra whose associated split polyhedra satisfy Eq. (1) gives an approximation of P_I of the form*

$$Cl(\mathcal{M}, P) := \bigcap_{L \in \mathcal{M}} R(L, P).$$

The set $Cl(\mathcal{M}, P)$ is called the closure with respect to the family \mathcal{M} .

Of course, improved approximations of P_I can be obtained by iteratively computing closures $P^1(\mathcal{M}, P), P^2(\mathcal{M}, P), \dots$, where $P^0(\mathcal{M}, P) := P$,

$$P^1(\mathcal{M}, P) := \text{Cl}(\mathcal{M}, P^0(\mathcal{M}, P)), \quad P^2(\mathcal{M}, P) := \text{Cl}(\mathcal{M}, P^1(\mathcal{M}, P))$$

etc. The questions emerge (i) which properties of the family \mathcal{M} ensure that the relaxation $\text{Cl}(\mathcal{M}, P)$ is again a polyhedron and (ii) how many rounds of closure computations are required in order to arrive at $\text{conv}(P_I)$. In order to shed some light on these questions, an appropriate measure of “size” or “complexity” of a split polyhedron is required.

Possible measures for the size of a mixed integer split polyhedron L are the max-facet-width of L or the lattice width of L , see Definition 3.

In addition, given that every mixed integer split polyhedron L can be written in the form $L = \mathcal{P} + \mathcal{S}$, where \mathcal{P} is a polytope and \mathcal{S} is a linear space, an alternative measure of the size of S could be its split dimension, i.e., the dimension of the polytope \mathcal{P} .

The simplest split polyhedra in terms of size should always be *splits*. A split is a split polyhedron of the form $\{x \in \mathbb{R}^{n+d} : \pi_0 \leq \pi^T x \leq \pi_0 + 1\}$, where $(\pi, \pi_0) \in \mathbb{Z}^{n+d+1}$ and $\pi_j = 0$ for $j \geq n + 1$. Note that a split has the smallest possible lattice width and the smallest possible max-facet-width among all split polyhedra. In general there does not exist a finite number k such that the operation of computing the closure of all splits k times yields the mixed integer hull, see [12] for such an example. This answers our second question negatively, at least for mixed integer splits.

We return to question (i). One important result in this context is

Theorem 3 (Split closure theorem). [12] *Let*

$$\mathcal{M}^1 := \{L \subseteq \mathbb{R}^{n+d} : L \text{ is a mixed integer split polyhedron satisfying } w_f(L) \leq 1\}$$

be the set of all splits. Then, for any rational polyhedron $P \subseteq \mathbb{R}^{n+d}$, the split closure

$$SC^1 := \bigcap_{L \in \mathcal{M}^1} R(L, P)$$

is a polyhedron.

A natural generalization of the split closure is to allow for split polyhedra that have max-facet-width larger than one. For any $w > 0$, define the set of mixed integer split polyhedra

$$\mathcal{M}^w := \{L \subseteq \mathbb{R}^{n+d} : L \text{ is a mixed integer split polyhedron satisfying } w_f(L) \leq w\}$$

with max-facet-width at most w . We define the w^{th} split closure to be the set

$$SC^w := \bigcap_{L \in \mathcal{M}^w} R(L, P).$$

The following result extends the split closure theorem to split polyhedra with arbitrary but fixed max-facet-width.

Theorem 4 (Split polyhedron closure theorem). *[1] For any family $\bar{\mathcal{M}} \subseteq \mathcal{M}^w$ of split polyhedra with bounded max-facet-width $w > 0$ and any rational polyhedron P , the set $\cap_{L \in \bar{\mathcal{M}}} R(L, P)$ is a polyhedron.*

The proof is an application of a more general result that gives a sufficient condition for the set $\cap_{L \in \bar{\mathcal{M}}} R(L, P)$ to be a polyhedron for *any* family $\bar{\mathcal{M}}$ of split polyhedra. We refer to [1] for further details.

3. Cutting Plane Proofs

In this section, we rely on the material presented in the previous section. We will, however, not choose a specific measure of the complexity of the split polyhedra. We simply use a generic function $\text{size}(L)$ for a split polyhedron L , and we assume that small values of $\text{size}(L)$ are desirable. We also assume that splits attain the smallest size.

Our point of view is that, if we happen to show that an inequality $a^T x + g^T y \geq \gamma$ that is valid for P_I is also valid for $R(L, P)$ for some split polyhedron L , then L provides a certificate (finite cutting plane proof) for the validity of the inequality.

More specifically, let us assume that we are given a generic measure of “size” or “complexity” of all split polyhedra. Moreover, let us assume that we are given a finite or even infinite set \mathcal{M} of split polyhedra. A natural question to ask is the following: What is the complexity of the split polyhedron $L \in \mathcal{M}$ that is required to prove validity of a specific inequality $a^T x + g^T y \geq \gamma$ that is known to be valid for $P_I := P \cap (\mathbb{Z}^n \times \mathbb{R}^d)$? A motivation for this question is the fact that there exist inequalities that are valid for a mixed integer hull of a problem and that cannot be derived by applying a finite number of closure operations using only splits [12]. On the other hand, for a mixed integer problem in dimension $n+d$ with n integer variables one may define \mathcal{M} to be the set of all lattice point free polyhedra of split dimension $\leq n$. Then it is an easy exercise to show that the closure operation $\text{Cl}(\mathcal{M}, P)$ applied recursively a finite number of times generates the mixed integer hull. This motivates to ask for the correct value of the size / complexity of the split polyhedron $L \in \mathcal{M}$ that is required to prove validity of a specific valid inequality for the mixed integer hull.

Definition 5 (Cutting plane proof). *Let $a^T x + g^T y \geq \gamma$ be valid for P_I . A cutting plane proof of $a^T x + g^T y \geq \gamma$ is a family \mathcal{M} of split polyhedra such that $a^T x + g^T y \geq \gamma$ is valid for $P^k(\mathcal{M}, P)$ for some $k < \infty$.*

A natural measure of the complexity of a cutting plane proof defined by the family \mathcal{M} is the largest size of a mixed integer split polyhedron $L \in \mathcal{M}$. This allows us to define the size of a valid inequality for P_I .

Definition 6 (Size of a valid inequality for P_I). Let $a^T x + g^T y \geq \gamma$ be valid for P_I . The size of the inequality $a^T x + g^T y \geq \gamma$ is defined to be the number

$$\text{size}(a, g, \gamma) := \min \left\{ \max \{ \text{size}(L) \mid L \in \mathcal{M} \} \mid \mathcal{M} \text{ is a cutting plane proof for } a^T x + g^T y \geq \gamma \right\}.$$

We now characterize which size is necessary to prove validity of an inequality $a^T x + g^T y \geq \gamma$ for P_I with a cutting plane proof, i.e., we characterize the number $\text{size}(a, g, \gamma)$. In what follows we assume that $a^T x + g^T y \geq \gamma$ is valid for P_I and that there exists at least one point $(x, y) \in P_I$ such that $a^T x + g^T y = \gamma$.

It turns out that the split polyhedra that are needed to prove validity of $a^T x + g^T y \geq \gamma$ depend on the facial structure of the following polyhedral set

$$L(a, g, \gamma) := \text{proj}_x \left((x, y) \in P \text{ and } a^T x + g^T y \geq \gamma \right).$$

Note that from the validity of the inequality $a^T x + g^T y \geq \gamma$ for P_I , it follows that the set $L(a, g, \gamma)$ is lattice point free (Proposition 1).

Let

$$\mathcal{F} := \{ F \text{ face of } L(a, g, \gamma) \mid \exists (x, y) \in P \text{ s.t. } x \in F \text{ and } a^T x + g^T y < \gamma \}$$

denote the set of all faces F of $L(a, g, \gamma)$ for which there exists $(x, y) \in P$ such that $x \in F$ and (x, y) violates the inequality $a^T x + g^T y \geq \gamma$. It can be shown that each $F \in \mathcal{F}$ is lattice point free and rational. Hence, for each $F \in \mathcal{F}$ there exists a split polyhedron L such that $F \subseteq L$.

Another important point is that if $F \in \mathcal{F}$, and if $G \notin \mathcal{F}$ for every proper face G of F , then one can verify that every split polyhedron $L \subseteq \mathbb{R}^n$ that satisfies $\text{int}(F) \subseteq \text{relint}(L)$ proves validity of $a^T x + g^T y \geq \gamma$ on F . Hence, by iteratively considering the finite number $|\mathcal{F}|$ of violated faces of $L(a, g, \gamma)$, we obtain a finite cutting plane proof for the validity of the inequality $a^T x + g^T y \geq \gamma$ for P_I . Conversely, if L is a split polyhedron of smaller size than $\max\{\text{size}(F) : F \in \mathcal{F}\}$, then finite convergence is not achievable. More specifically, there exists $F \in \mathcal{F}$ and $x' \in \text{int}(F)$ such that $x' \notin \text{relint}(L)$. Furthermore, since $x' \in \text{int}(F)$, it can be shown that there exists $y' \in \mathbb{R}^q$ such that $(x', y') \in P$ and $a^T x' + g^T y' < \gamma$. We now have $(x', y') \in R(L, P)$ and $a^T x' + g^T y' < \gamma$.

These arguments sketch the proof of the main result in this section.

Theorem 5 (A formula for the size of the inequality $a^T x + g^T y \geq \gamma$). [1] For an inequality $a^T x + g^T y \geq \gamma$ that is valid for P_I , we have that

$$\text{size}(a, g, \gamma) = \max \left\{ \min \{ \text{size}(L) \mid L \text{ a split polyhedron, } F \subseteq L \} \mid F \in \mathcal{F} \right\}.$$

Theorem 5 allows us to associate a complexity with every valid inequality for $P \cap (\mathbb{Z}^n \times \mathbb{R}^d)$. This complexity, can be computed face by face a-posteriori, only. On the other hand, given a concrete instance $P \cap (\mathbb{Z}^n \times \mathbb{R}^d)$, it would be

desirable to know an estimate on the size of the inequalities needed to describe the mixed integer hull of P a priori for the selection of split polyhedra of appropriate complexity. For performing this task Theorem 5 is, however, of no help.

4. Cutting Plane Generation from Split Polyhedra

In the previous section we have elucidated that cutting planes for mixed integer programs are in tight relation with lattice point free polyhedra. The purpose of this section is to discuss how to generate cutting planes from lattice point free polyhedra. This question is certainly interesting in its own right. It is yet essential to answer it in order to establish an algorithmic framework that is capable of computing cuts for concrete mixed integer models. One requirement in generating cuts from split polyhedra is that the complexity of the cut generation should be matched by the measure of the complexity of the corresponding lattice point free polyhedra that are used for the computations. Secondly, in order to compute with lattice point free polyhedra it is necessary to get control over the explicit description of such objects. Once such a description is at hand, computations can be performed based on an extended formulation, see [5] for details.

The problem here is that even in dimension two there are infinitely many maximally lattice point free polyhedra, even up to unimodular transformations. Here and throughout the paper affine mappings in \mathbb{R}^m which preserve \mathbb{Z}^m are referred to as (*affine*) *unimodular transformations*. There is evidence to believe that in order to design a finite cutting plane algorithm for a mixed integer linear program it is sufficient to compute with the subset of those maximally lattice point free polyhedra L that are integral, i.e., every minimal face of L contains an integral point.

Conjecture: Let \mathcal{M}_I be the set of all split polyhedra with integral vertices. There exists a positive integer k such that $\text{conv}(P_I) = P^k(\mathcal{M}_I, P)$.

This conjecture has been verified in [13] for two integer variables when P is an affine cone, see Section 5 for details regarding the model.

In dimension two, it can easily be verified that, up to unimodular transformation, only two integral maximally lattice point free polyhedra exist. These are given as

$$\text{conv}(0, e_1) + \text{lin}(e_2), \quad \text{conv}(0, 2e_1, 2e_2),$$

where e_1, \dots, e_d always denote the d unit vectors in \mathbb{R}^d . In higher dimensions a characterization of the subset of all maximally lattice point free polyhedra with the additional property that they are integral is more demanding. In di-

mension three, a recent paper of Treutlein [18] shows finiteness. An explicit characterization is derived in [4].

Theorem 6. [4] *Every maximally lattice point free polytope $P \subseteq \mathbb{R}^3$ with integral vertices is up to unimodular transformation one of the following polytopes.*

- one of the seven simplices

$$\begin{aligned} M_1 &= \text{conv}(\{0, 2e_1, 3e_2, 6e_3\}), \\ M_2 &= \text{conv}(\{0, 2e_1, 4e_2, 4e_3\}), \\ M_3 &= \text{conv}(\{0, 3e_1, 3e_2, 3e_3\}), \\ M_4 &= \text{conv}(\{0, e_1, 2e_1 + 4e_2, 3e_1 + 4e_3\}), \\ M_5 &= \text{conv}(\{0, e_1, 2e_1 + 5e_2, 3e_1 + 5e_3\}), \\ M_6 &= \text{conv}(\{0, 3e_1, e_1 + 3e_2, 2e_1 + 3e_3\}), \\ M_7 &= \text{conv}(\{0, 4e_1, e_1 + 2e_2, 2e_1 + 4e_3\}), \end{aligned}$$

- the pyramid $M_8 = \text{conv}(B \cup \{a\})$ with the base $B = \text{conv}(\{\pm 2e_1, \pm 2e_2\})$ and the apex $a = (1, 1, 2)$,
- the pyramid $M_9 = \text{conv}(B \cup \{a\})$ with the base $B = \text{conv}(\{-e_1, -e_2, 2e_1, 2e_2\})$ and the apex $a = (1, 1, 3)$,
- the prism $M_{10} = \text{conv}(B \cup (B + u))$ with the bases B and $B + u$, where $B = \text{conv}(\{e_1, e_2, -(e_1 + e_2)\})$ and $u = (1, 2, 3)$,
- the prism $M_{11} = \text{conv}(B \cup (B + u))$ with the bases B and $B + u$, where $B = \text{conv}(\{\pm e_1, 2e_2\})$ and $u = (1, 0, 2)$,
- the parallelepiped $M_{12} = \text{conv}(\{\sigma_1 u_1 + \sigma_2 u_2 + \sigma_3 u_3 : \sigma_1, \sigma_2, \sigma_3 \in \{0, 1\}\})$ where $u_1 = (-1, 1, 0)$, $u_2 = (1, 1, 0)$, and $u_3 = (1, 1, 2)$.

For dimension greater than three no explicit description of all maximally lattice point free polytopes with the additional property of being integral is known. In fact, it remains a challenge to prove even finiteness of these objects in general dimensions.

5. Integer Points in an Affine Cone

Next we introduce a basic mixed integer model that allows us to develop a deeper understanding of the connection between lattice point free sets on the one hand and cutting planes for the associated mixed integer hull. For a finite set of primitive vectors $\{g^j : j = 1, \dots, d\} \subseteq \mathbb{Z}^n$ and a given point $f \in \mathbb{Q}^n \setminus \mathbb{Z}^n$, we investigate the set

$$P_I = \left\{ (x, y) \in \mathbb{Z}^n \times \mathbb{R}_+^d : x = f + \sum_{j=1}^d y_j g^j \right\},$$

with associated polyhedron $P = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}_+^d : x = f + \sum_{j=1}^d y_j g^j\}$. Note that geometrically P describes an affine cone with apex f , hence the title of the section.

This model arises as a natural relaxation of a general mixed integer problem: then f denotes a vertex of the continuous relaxation of the mixed integer instance with incident edge directions g^j . By ignoring nonnegativity on the variables x and integrality on the variables associated with the edges g^j one arrives at a relaxation of the original problem in form of a set P_I . Let us also remark that there is a close connection to the so-called corner polyhedron introduced in [15] and studied further in [16], see also [9]. Roughly speaking, in the corner polyhedron introduced in [15] one keeps the integrality condition on all the integer variables associated with the rays g^j , whereas this condition is ignored in a relaxation of the form P_I . This basic mixed integer model has been introduced and studied in [2] for the case of $n = 2$. Several further papers emerged from this study, see [8], [10].

One property of $\text{conv}(P_I)$ is that any nontrivial valid inequality for $\text{conv}(P_I)$ can be written in so-called standard form, i.e.,

$$\sum_{j=1}^d \alpha_j y_j \geq 1,$$

where $\alpha_j \geq 0$ for all $j = 1, \dots, d$. We denote by $N_\alpha^0 := \{j \mid \alpha_j = 0\}$ the set of variables with coefficient zero.

The link to lattice point free polyhedra in this special situation is summarized below. The following notation is needed. If $\sum_{j=1}^d \alpha_j y_j \geq 1$ defines a facet-defining inequality for $\text{conv}(P_I)$, then we define

$$L_\alpha = \text{proj}_x \left\{ (x, y) \in P, \sum_{j=1}^d \alpha_j y_j \leq 1 \right\}.$$

Note that L_α for this special mixed integer model corresponds to $L(a, g, \gamma)$ that we introduced in Section 3 in the context of general mixed integer programs.

Theorem 7 (Facet Classification Theorem). [3] *Let $\sum_{j=1}^d \alpha_j y_j \geq 1$ be a facet-defining inequality for $\text{conv}(P_I)$, where $N_\alpha^0 \neq \emptyset$. Moreover, let $S_\alpha := \text{lin}(\{g^j\}_{j \in N_\alpha^0})$ and $\dim(S_\alpha) = s$.*

1. *There exists a rational lattice point free polytope $P_\alpha \subseteq \mathbb{R}^{n-s}$ such that $L_\alpha \subseteq P_\alpha + S_\alpha$.*
2. *If $s = n - 1$, there exists $(\pi, \pi_0) \in \mathbb{Z}^{n+1}$ such that $L_\alpha \subseteq \{x \in \mathbb{R}^n : \pi_0 \leq \pi^T x \leq \pi_0 + 1\}$.*

Theorem 7 allows us to classify facet-defining inequalities for $\text{conv}(P_I)$ according to the dimension of the linear subspace that is spanned by the rays

$\{g^j\}_{j \in N_\alpha^0}$. More precisely, we classify the facets of $\text{conv}(P_I)$ according to the split-dimension of their corresponding L_α -bodies, which is $n - \dim(S_\alpha)$ for $S_\alpha = \text{lin}(\{g^j\}_{j \in N_\alpha^0})$. In particular, the facet $\sum_{j=1}^d \alpha_j y_j \geq 1$ is a split cut whenever $\dim(S_\alpha) = n - 1$. Note that any facet-defining inequality $\sum_{j=1}^d \alpha_j y_j \geq 1$ for $\text{conv}(P_I)$, where $\dim(S_\alpha) \leq n - 2$ can never be a split cut. Next we argue that the set of all facets of $\text{conv}(P_I)$ whose associated split-dimension is equal to n is needed in order to provide a tight approximation of the mixed-integer hull.

For the case $n = 2$ this result has been shown by Basu et al. [7]. We present here its generalization to dimensions $n \geq 3$, i.e., we show that using only facets of $\text{conv}(P_I)$ which correspond to L_α -polyhedra with split-dimension $< n$ will lead to an arbitrarily bad approximation of the mixed-integer hull, in general.

We denote by $S^i(P_I)$ the *i-dimensional split closure*, i.e., the intersection of all valid inequalities for $\text{conv}(P_I)$ corresponding to L_α -bodies of split-dimension $\leq i$, for all $i = 1, \dots, n$. Note that $S^1(P_I)$ is the usual split closure, whereas $S^n(P_I) = \{y \in \mathbb{R}_+^d : \sum_{j=1}^d \alpha_j^i y_j \geq 1 \text{ for all } i \in \mathcal{T}\}$, where \mathcal{T} is the set of all facet-defining inequalities for $\text{conv}(P_I)$.

Theorem 8 (Inapproximability Theorem). [3] *For any $\alpha > 1$ there exists a choice of a fractional point $f \in \mathbb{Q}^n \setminus \mathbb{Z}^n$ and rays $g^1, \dots, g^d \in \mathbb{Q}^n$ such that $S^{n-1}(P_I) \not\subseteq \alpha S^n(P_I)$.*

Theorem 8 implies that the approximability of the mixed-integer hull by $S^{n-1}(P_I)$ may be arbitrarily bad. This statement applies to general sets P_I . However, by restricting the size of the input data, split polyhedra of split-dimension equal to n can be approximated using ordinary splits to within a constant factor. For a fractional point $f \in \mathbb{Q}^n \setminus \mathbb{Z}^n$ we define the *precision of f* as the smallest integer $q \in \mathbb{Z}_+$ such that f has a representation $f = (\frac{p_1}{q}, \dots, \frac{p_n}{q})$, where $p_j \in \mathbb{Z}$ for all $j = 1, \dots, n$. Recalling Definition 3, we are now prepared to state our result on the approximability of the mixed integer hull by means of the split closure.

Theorem 9 (Split closure approximation theorem). [3] *Let q be the precision of the fractional point f , and let \mathcal{F} be the family of all valid inequalities for $\text{conv}(P_I)$ arising from split polyhedra with split-dimension equal to n and max-facet-width $\leq w^*$. Define*

$$\bar{S}(P_I) := S^1(P_I) \cap \left\{ y \in \mathbb{R}_+^d : \sum_{j=1}^d \alpha_j^u y_j \geq 1 \ \forall u \in \mathcal{F} \right\}.$$

Then, $\bar{S}(P_I) \subseteq S^1(P_I) \subseteq \frac{nqw^}{2} \bar{S}(P_I)$.*

Computational tests by many researchers have revealed that split cuts perform extremely well in practice. Theorem 9 supports this hypothesis theoretically.

References

- [1] K. Andersen, Q. Louveaux, R. Weismantel, *An analysis of mixed integer linear sets based on lattice point free convex sets*, Mathematics of Operations Research, to appear (2010).
- [2] K. Andersen, Q. Louveaux, R. Weismantel, L. Wolsey, *Inequalities from two rows of a simplex tableau*, IPCO 2007, Lecture Notes in Computer Science 4513, Springer, (2007), 1–15.
- [3] K. Andersen, Ch. Wagner, R. Weismantel, *On an analysis of the strength of mixed integer cutting planes from multiple simplex tableau rows*, SIAM Journal on Optimization 20, (2009), 967–982.
- [4] G. Averkov, Ch. Wagner, R. Weismantel, *Maximally lattice free polyhedra in dimension three with integral vertices*, Manuscript, (2009).
- [5] E. Balas, *Disjunctive Programming*, Annals of Discrete Mathematics 5, 3–51, (1979).
- [6] E. Balas, S. Ceria, G. Cornuéjols, *A lift-and-project cutting plane algorithm for mixed 0/1 programs*, Mathematical Programming 58, 295–324, (1993).
- [7] A. Basu, P. Bonami, G. Cornuéjols, F. Margot, *On the relative strength of split, triangle and quadrilateral cuts*, Manuscript, Mathematical Programming, to appear (2010).
- [8] V. Borozan, G. Cornuéjols, *Minimal valid inequalities for integer constraints*, Mathematics of Operations Research 34, (2009), 538–546.
- [9] C. A. Burdet, E. L. Johnson, *A subadditive approach to the group problem of integer programming*, Mathematical Programming 2, (1974), 51–71.
- [10] G. Cornuéjols, F. Margot, *On the facets of mixed integer programs with two integer variables and two constraints*, Mathematical Programming 120, (2009), 429–456.
- [11] V. Chvátal, *Edmonds Polytopes and a Hierarchy of Combinatorial Problems*, Discrete Mathematics 4, (1973), 305–337.
- [12] W.J. Cook, R. Kannan, A. Schrijver, *Chvátal closures for mixed integer programming problems*, Mathematical Programming 47, (1990), 155–174.
- [13] S. Dey, Q. Louveaux, *Split rank of triangle and quadrilateral inequalities*, Manuscript, arXiv:0906.0887, (2009).
- [14] R. E. Gomory, *Outline of an Algorithm for Integer Solutions to Linear Programs*, Bulletin of the American Mathematical Society 64, (1958), 275–278.
- [15] R. E. Gomory, *Some polyhedra related to combinatorial problems*, Linear Algebra and Applications 2, (1969), 451–558.
- [16] R. E. Gomory, E. L. Johnson, *Some continuous functions related to corner polyhedra I*, Mathematical Programming 3, (1972), 23–85.
- [17] L. Lovász, *Geometry of numbers and integer programming*, Mathematical Programming: Recent developments and Applications , M. Iri and K. Tanabe (eds), Kluwer, (1989), 177–201.
- [18] J. Treutlein, *3-dimensional lattice polytopes without interior lattice points*, Manuscript arXiv:0809.1787, (2008).

A Unified Controllability/Observability Theory for Some Stochastic and Deterministic Partial Differential Equations

Xu Zhang*

Abstract

The purpose of this paper is to present a universal approach to the study of controllability/observability problems for infinite dimensional systems governed by some stochastic/deterministic partial differential equations. The crucial analytic tool is a class of fundamental weighted identities for stochastic/deterministic partial differential operators, via which one can derive the desired global Carleman estimates. This method can also give a unified treatment of the stabilization, global unique continuation, and inverse problems for some stochastic/deterministic partial differential equations.

Mathematics Subject Classification (2010). Primary 93B05; Secondary 35Q93, 93B07.

Keywords. Controllability, observability, parabolic equations, hyperbolic equations, weighted identity.

1. Introduction

We begin with the following controlled system governed by a linear Ordinary Differential Equation (ODE for short):

$$\begin{cases} \frac{dy(t)}{dt} = Ay(t) + Bu(t), & t > 0, \\ y(0) = y_0. \end{cases} \quad (1.1)$$

*This work is supported by the NSFC under grants 10831007, 60821091 and 60974035, and the project MTM2008-03541 of the Spanish Ministry of Science and Innovation.

School of Mathematics, Sichuan University, Chengdu 610064, China; and Key Laboratory of Systems Control, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China. E-mail: xuzhang@amss.ac.cn.

In (1.1), $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$ ($n, m \in \mathbb{N}$), $y(\cdot)$ is the *state variable*, $u(\cdot)$ is the *control variable*, \mathbb{R}^n and \mathbb{R}^m are the *state space and control space*, respectively. System (1.1) is said to be exactly controllable at a time $T > 0$ if for any initial state $y_0 \in \mathbb{R}^n$ and any final state $y_1 \in \mathbb{R}^n$, there is a control $u(\cdot) \in L^2(0, T; \mathbb{R}^m)$ such that the solution $y(\cdot)$ of (1.1) satisfies $y(T) = y_1$.

The above definition of controllability can be easily extended to abstract evolution equations. In the general setting, it may happen that the requirement $y(T) = y_1$ has to be relaxed in one way or another. This leads to the approximate controllability, null controllability, and partial controllability, etc. Roughly speaking, the controllability problem for an evolution process is driving the state of the system to a prescribed final target state (exactly or in some approximate way) at a finite time. Also, the above B can be unbounded for general controlled systems.

The controllability/observability theory for finite dimensional linear systems was introduced by R.E. Kalman ([19]). It is by now the basis of the whole control theory. Note that a finite dimensional system is usually an approximation of some infinite dimensional system. Therefore, stimulated by Kalman’s work, many mathematicians devoted to extend it to more general systems including infinite dimensional systems, and its nonlinear and stochastic counterparts. However, compared with Kalman’s classical theory, the extended theories are not very mature.

Let us review rapidly the main results of Kalman’s theory. First of all, it is shown that: *System (1.1) is exactly controllable at a time T if and only if $\text{rank}[B, AB, \dots, A^{n-1}B] = n$* . However, this criterion is not applicable for general infinite dimensional systems. Instead, in the general setting, one uses another method which reduces the controllability problem for a controlled system to an observability problem for its dual system. The dual system of (1.1) reads:

$$\begin{cases} \frac{dw}{dt} = -A^*w, & t \in (0, T), \\ w(T) = z_0. \end{cases} \tag{1.2}$$

It is shown that: *System (1.1) is exactly controllable at some time T if and only if the following observability inequality (or estimate) holds*

$$|z_0|^2 \leq C \int_0^T |B^*w(t)|^2 dt, \quad \forall z_0 \in \mathbb{R}^n. \tag{1.3}$$

Here and henceforth, C denotes a generic positive constant, which may be different from one place to another. We remark that similar results remain true in the infinite dimensional setting, where the theme of the controllability/observability theory is to establish suitable observability estimates through various approaches.

Systems governed by Partial Differential Equations (PDEs for short) are typically infinite dimensional. There exists many works on controllability/observability of PDEs. Contributions by D.L. Russell ([40]) and by J.L. Lions ([29]) are classical in this field. In particular, since it stimulated many

in-depth researches on related problems in PDEs, J.L. Lions's paper [29] triggered extensive works addressing the controllability/observability of infinite dimensional controlled system. After [29], important works in this field can be found in [1, 4, 8, 11, 13, 17, 21, 25, 26, 43, 46, 55, 56]. For other related works, we refer to [18, 28] and so on.

The controllability/observability of PDEs depends strongly on the nature of the underlying system, such as time reversibility or not, and propagation speed of solutions, etc. The wave equation and the heat equation are typical examples. Now it is clear that essential differences exist between the controllability/observability theories for these two equations. Naturally, one expects to know whether some relationship exist between the controllability/observability theories for these two equations of different nature. Especially, it would be quite interesting to establish, in some sense and to some extent, a unified controllability/observability theory for parabolic equations and hyperbolic equations. This problem was initially studied by D.L. Russell ([39]).

The main purpose of this paper is to present the author's and his collaborators' works with an effort towards a unified controllability/observability theory for stochastic/deterministic PDEs. The crucial analytic tool we employ is a class of elementary pointwise weighted identities for partial differential operators. Starting from these identities, we develop a unified approach, based on global Carleman estimate. This universal approach not only deduces the known controllability/observability results (that have been derived before via Carleman estimates) for the linear parabolic, hyperbolic, Schrödinger and plate equations, but also provides new/sharp results on controllability/observability, global unique continuation, stabilization and inverse problems for some stochastic/deterministic linear/nonlinear PDEs.

The rest of this paper is organized as follows. Section 2 analyzes the main differences between the existing controllability/observability theories for parabolic equations and hyperbolic equations. Sections 3 and 4 address, among others, the unified treatment of the controllability/observability problem for deterministic PDEs and stochastic PDEs, respectively.

2. Main Differences Between the Known Theories

In the sequel, unless otherwise indicated, G stands for a bounded domain (in \mathbb{R}^n) with a boundary $\Gamma \in C^2$, G_0 denotes an open non-empty subset of G , and T is a given positive number. Put $Q = (0, T) \times G$, $Q_{G_0} = (0, T) \times G_0$ and $\Sigma = (0, T) \times \Gamma$.

We begin with a controlled heat equation:

$$\begin{cases} y_t - \Delta y = \chi_{G_0}(x)u(t, x) & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y(0) = y_0 & \text{in } G \end{cases} \quad (2.1)$$

and a controlled wave equation:

$$\begin{cases} y_{tt} - \Delta y = \chi_{G_0}(x)u(t, x) & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y(0) = y_0, \quad y_t(0) = y_1 & \text{in } G. \end{cases} \tag{2.2}$$

In (2.1), y and u are the *state variable* and *control variable*, the *state space* and *control space* are chosen to be $L^2(G)$ and $L^2(Q_{G_0})$, respectively; while in (2.2), (y, y_t) and u are the *state variable* and *control variable*, $H_0^1(G) \times L^2(G)$ and $L^2(Q_{G_0})$ are respectively the *state space* and *control space*. System (2.1) is said to be null controllable (*resp.* approximately controllable) in $L^2(G)$ if for any given $y_0 \in L^2(G)$ (*resp.* for any given $\varepsilon > 0, y_0, y_1 \in L^2(G)$), one can find a control $u \in L^2(Q_{G_0})$ such that the weak solution $y(\cdot) \in C([0, T]; L^2(G)) \cap C((0, T]; H_0^1(G))$ of (2.1) satisfies $y(T) = 0$ (*resp.* $|y(T) - y_1|_{L^2(G)} \leq \varepsilon$). In the case of null controllability, the corresponding control u is called a null-control (with initial state y_0). Note that, due to the smoothing effect of solutions to the heat equation, the exact controllability for (2.1) is impossible, i.e., the above ε cannot be zero. On the other hand, since one can rewrite system (2.2) as an evolution equation in a form like (1.1), it is easy to define the exact controllability of this system. The dual systems of (2.1) and (2.2) read respectively

$$\begin{cases} \psi_t + \Delta \psi = 0 & \text{in } Q, \\ \psi = 0 & \text{on } \Sigma, \\ \psi(T) = \psi_0 & \text{in } G \end{cases} \tag{2.3}$$

and

$$\begin{cases} \psi_{tt} - \Delta \psi = 0 & \text{in } Q, \\ \psi = 0 & \text{on } \Sigma, \\ \psi(T) = \psi_0, \quad \psi_t(T) = \psi_1 & \text{in } G. \end{cases} \tag{2.4}$$

The controllability/observability theories for parabolic equations and hyperbolic equations turns out to be quite different. First of all, we recall the related result for the heat equation.

Theorem 2.1. ([25]) *Let G be a bounded domain of class C^∞ . Then: i) System (2.1) is null controllable and approximately controllable in $L^2(G)$ at time T ; ii) Solutions of equation (2.3) satisfy*

$$|\psi(0)|_{L^2(G)} \leq C|\psi|_{L^2(Q_{G_0})}, \quad \forall \psi_0 \in L^2(G). \tag{2.5}$$

Since solutions to the heat equation have an infinite propagation speed, the “waiting” time T can be chosen as small as one likes, and the control domain G_0 does not need to satisfy any geometric condition but being open and non-empty. On the other hand, due to the time irreversibility and the strong dissipativity of (2.3), one cannot replace $|\psi(0)|_{L^2(G)}$ in inequality (2.5) by $|\psi_0|_{L^2(G)}$.

Denote by $\{\mu_i\}_{i=1}^\infty$ the eigenvalues of the homogenous Dirichlet Laplacian on G , and $\{\varphi_i\}_{i=1}^\infty$ the corresponding eigenvectors satisfying $|\varphi_i|_{L^2(G)} = 1$. The proof of Theorem 2.1 is based on the following observability estimate on sums of eigenfunctions for the Laplacian ([25]):

Theorem 2.2. *Under the assumption of Theorem 2.1, for any $r > 0$, it holds*

$$\sum_{\mu_i \leq r} |a_i|^2 \leq C e^{C\sqrt{r}} \int_{G_0} \left| \sum_{\mu_i \leq r} a_i \varphi_i(x) \right|^2 dx, \quad \forall \{a_i\}_{\mu_i \leq r} \text{ with } a_i \in \mathbb{C}. \quad (2.6)$$

Note that Theorem 2.2 has some other applications in control problems of PDEs ([32, 34, 44, 49, 55, 56]). Besides, to prove Theorem 2.1, one needs to utilize a time iteration method ([25]), which uses essentially the Fourier decomposition of solutions to (2.3) and especially, the strong dissipativity of this equation. Hence, this method cannot be applied to conservative systems (say, system (2.2)) or the system that the underlined equation is time-dependent.

As for the controllability/observability for the wave equation, we need to introduce the following notations. Fix any $x_0 \in \mathbb{R}^n$, put

$$\Gamma_0 \triangleq \{x \in \Gamma \mid (x - x_0) \cdot \nu(x) > 0\}, \quad (2.7)$$

where $\nu(x)$ is the unit outward normal vector of G at $x \in \Gamma$. For any set $S \in \mathbb{R}^n$ and $\varepsilon > 0$, put $\mathcal{O}_\varepsilon(S) = \{y \in \mathbb{R}^n \mid |y - x| < \varepsilon \text{ for some } x \in S\}$.

The exact controllability of system (2.2) is equivalent to the following *observability estimate* for system (2.4):

$$|(\psi_0, \psi_1)|_{L^2(G) \times H^{-1}(G)} \leq C |\psi|_{L^2(Q_{G_0})}, \quad \forall (\psi_0, \psi_1) \in L^2(G) \times H^{-1}(G). \quad (2.8)$$

Note that the left hand side of (2.8) can be replaced by $|(\psi(0), \psi_t(0))|_{L^2(G) \times H^{-1}(G)}^2$ (because (2.4) is conservative). The following classical result can be found in [29].

Theorem 2.3. *Assume $G_0 = O_\varepsilon(\Gamma_0) \cap G$ and $T_0 = 2 \sup_{x \in G \setminus G_0} |x - x_0|$. Then, inequality (2.8) holds for any time $T > T_0$.*

The proof of Theorem 2.3 is based on a classical Rellich-type multiplier method. Indeed, it is a consequence of the following identity (e.g. [47]):

Proposition 2.4. *Let $h \triangleq (h^1, \dots, h^n) : \mathbb{R} \times \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a vector field of class C^1 . Then for any $z \in C^2(\mathbb{R} \times \mathbb{R}^n)$, it holds that*

$$\begin{aligned} & \nabla \cdot \left\{ 2(h \cdot \nabla z)(\nabla z) + h \left[z_t^2 - \sum_{i=1}^n z_{x_i}^2 \right] \right\} \\ &= -2(z_{tt} - \Delta z)h \cdot \nabla z + (2z_t h \cdot \nabla z)_t - 2z_t h_t \cdot \nabla z \\ &+ (\nabla \cdot h) \left[z_t^2 - \sum_{i=1}^n z_{x_i}^2 \right] + 2 \sum_{i,j=1}^n \left(\frac{\partial h^j}{\partial x_i} z_{x_i} z_{x_j} \right). \end{aligned}$$

The observability time T in Theorem 2.3 should be large enough. This is due to the finite propagation speed of solutions to the wave equation (except when

the control is acting in the whole domain G). On the other hand, it is shown in [4] that exact controllability of (2.2) is impossible without geometric conditions on G_0 . Note also that, the multiplier method rarely provides the optimal control/observation domain and minimal controll/observation time except for some very special geometries. These restrictions are weakened by the microlocal analysis ([4]). In [4, 5, 6], the authors proved that, roughly speaking, inequality (2.8) holds if and only if every ray of Geometric Optics that propagates in G and is reflected on its boundary Γ enters G_0 at time less than T .

The above discussion indicates that the results and methods for the controllability/observability of the heat equation differ from those of the wave equation. As we mentioned before, this leads to the problem of establishing a unified theory for the controllability/observability of parabolic equations and hyperbolic equations. The first result in this direction was given in [39], which showed that *the exact controllability of the wave equation implies the null controllability of the heat equation with the same controller but in a short time*. Further results were obtained in [32, 49], in which organic connections were established for the controllability theories between parabolic equations and hyperbolic equations. More precisely, it has been shown that: *i) By taking the singular limit of some exactly controllable hyperbolic equations, one gives the null controllability of some parabolic equations ([32]); and ii) Controllability results of the heat equation can be derived from the exact controllability of some hyperbolic equations ([49])*. Other interesting related works can be found in [34, 36, 43]. In the sequel, we shall focus mainly on a unified treatment of the controllability/observability for both deterministic PDEs and stochastic PDEs, from the methodology point of view.

3. The Deterministic Case

The key to solve controllability/observability problems for PDEs is the obtention of suitable observability inequalities for the underlying homogeneous systems. Nevertheless, as we see in Section 2, the techniques that have been developed to obtain such estimates depend heavily on the nature of the equations, especially when one expects to obtain sharp results for time-invariant equations. As for the time-variant case, in principle one needs to employ Carleman estimates, see [17] for the parabolic equation and [47] for the hyperbolic equation. The Carleman estimate is simply a weighted energy method. However, at least formally, the Carleman estimate used to derive the observability inequality for parabolic equations is quite different from that for hyperbolic ones. The main purpose of this section is to present a universal approach for the controllability/observability of some deterministic PDEs. Our approach is based on global Carleman estimates via a fundamental pointwise weighted identity for partial differential operators of second order (It was established in [13, 15]. See [27] for an earlier result). This approach is stimulated by [24, 20], both of which are addressed for ill-posed problems.

3.1. A stimulating example. The basic idea of Carleman estimates is available in proving the stability of ODEs ([27]). Indeed, consider an ODE in \mathbb{R}^n :

$$\begin{cases} x_t(t) = a(t)x(t), & t \in [0, T], \\ x(0) = x_0, \end{cases} \tag{3.1}$$

where $a \in L^\infty(0, T)$. A well-known simple result reads: *Solutions of (3.1) satisfy*

$$\max_{t \in [0, T]} |x(t)| \leq C|x_0|, \quad \forall x_0 \in \mathbb{R}^n. \tag{3.2}$$

A Carleman-type Proof of (3.2). For any $\lambda \in \mathbb{R}$, by (3.1), one obtains

$$\frac{d}{dt} \left(e^{-\lambda t} |x(t)|^2 \right) = -\lambda e^{-\lambda t} |x(t)|^2 + 2e^{-\lambda t} x_t(t) \cdot x(t) = (2a(t) - \lambda) e^{-\lambda t} |x(t)|^2. \tag{3.3}$$

Choosing λ large enough so that $2a(t) - \lambda \leq 0$ for a.e. $t \in (0, T)$, we find that

$$|x(t)| \leq e^{\lambda T/2} |x_0|, \quad t \in [0, T],$$

which proves (3.2). □

Remark 3.1. *By (3.3), we see the following pointwise identity:*

$$2e^{-\lambda t} x_t(t) \cdot x(t) = \frac{d}{dt} \left(e^{-\lambda t} |x(t)|^2 \right) + \lambda e^{-\lambda t} |x(t)|^2. \tag{3.4}$$

Note that $x_t(t)$ is the principal operator of the first equation in (3.1). The main idea of (3.4) is to establish a pointwise identity (and/or estimate) on the principal operator $x_t(t)$ in terms of the sum of a “divergence” term $\frac{d}{dt} (e^{-\lambda t} |x(t)|^2)$ and an “energy” term $\lambda e^{-\lambda t} |x(t)|^2$. As we see in the above proof, one chooses λ to be big enough to absorb the undesired terms. This is the key of all Carleman-type estimates. In the sequel, we use exactly the same method, i.e., the method of Carleman estimate via pointwise estimate, to derive observability inequalities for both parabolic equations and hyperbolic equations.

3.2. Pointwise weighted identity. We now show a fundamental pointwise weighted identity for general partial differential operator of second order. Fix real functions $\alpha, \beta \in C^1(\mathbb{R}^{1+m})$ and $b^{jk} \in C^1(\mathbb{R}^{1+m})$ satisfying $b^{jk} = b^{kj}$ ($j, k = 1, 2, \dots, m$). Define a formal differential operator of second order: $\mathcal{P}z \triangleq (\alpha + i\beta)z_t + \sum_{j,k=1}^m (b^{jk} z_{x_j})_{x_k}$, $i = \sqrt{-1}$. The following identity was established in [13, 15]:

Theorem 3.2. *Let $z \in C^2(\mathbb{R}^{1+m}; \mathbb{C})$ and $\ell \in C^2(\mathbb{R}^{1+m}; \mathbb{R})$. Put $\theta = e^\ell$ and $v = \theta z$. Let $a, b, \lambda \in \mathbb{R}$ be parameters. Then*

$$\begin{aligned}
 & \theta(\mathcal{P}z\overline{I_1} + \overline{\mathcal{P}z}I_1) + M_t + \sum_{k=1}^m \partial_{x_k} V^k \\
 &= 2|I_1|^2 + \sum_{j,k,j',k'=1}^m \left[2(b^{j'k} \ell_{x_{j'}})_{x_k} b^{jk'} - (b^{jk} b^{j'k'} \ell_{x_{j'}})_{x_{k'}} + \frac{1}{2}(\alpha b^{jk})_t \right. \\
 & \quad \left. - a b^{jk} b^{j'k'} \ell_{x_{j'} x_{k'}} \right] (v_{x_k} \overline{v}_{x_j} + \overline{v}_{x_k} v_{x_j}) + \left[- \sum_{j,k=1}^m b^{jk} \ell_{x_j} + b\lambda \right] (I_1 \overline{v} + \overline{I_1} v) \\
 & \quad + i \sum_{j,k=1}^m \left\{ [(\beta b^{jk} \ell_{x_j})_t + b^{jk} (\beta \ell_t)_{x_j}] (\overline{v}_{x_k} v - v_{x_k} \overline{v}) \right. \\
 & \quad \left. + [(\beta b^{jk} \ell_{x_j})_{x_k} + a \beta b^{jk} \ell_{x_j x_k}] (\overline{v} v_t - v \overline{v}_t) \right\} - \sum_{j,k=1}^m b^{jk} \alpha_{x_k} (v_{x_j} \overline{v}_t + \overline{v}_{x_j} v_t) \\
 & \quad - a \sum_{j,k,j',k'=1}^m b^{jk} (b^{j'k'} \ell_{x_{j'} x_{k'}})_{x_k} (\overline{v}_{x_j} v + v_{x_j} \overline{v}) + B|v|^2,
 \end{aligned} \tag{3.5}$$

where

$$\left\{ \begin{aligned}
 I_1 &\triangleq i\beta v_t - \alpha \ell_t v + \sum_{j,k=1}^m (b^{jk} v_{x_j})_{x_k} + Av, \\
 A &\triangleq \sum_{j,k=1}^m b^{jk} \ell_{x_j} \ell_{x_k} - (1+a) \sum_{j,k=1}^m b^{jk} \ell_{x_j x_k} - b\lambda, \\
 B &\triangleq (\alpha^2 \ell_t + \beta^2 \ell_t - \alpha A)_t \\
 &\quad + 2 \sum_{j,k=1}^m \left[(b^{jk} \ell_{x_j} A)_{x_k} - (\alpha b^{jk} \ell_{x_j} \ell_t)_{x_k} + a(A - \alpha \ell_t) b^{jk} \ell_{x_j x_k} \right], \\
 M &\triangleq [(\alpha^2 + \beta^2) \ell_t - \alpha A] |v|^2 + \alpha \sum_{j,k=1}^m b^{jk} v_{x_j} \overline{v}_{x_k} \\
 &\quad + i\beta \sum_{j,k=1}^m b^{jk} \ell_{x_j} (\overline{v}_{x_k} v - v_{x_k} \overline{v}), \\
 V^k &\triangleq \sum_{j,j',k'=1}^m \left\{ -i\beta \left[b^{jk} \ell_{x_j} (v \overline{v}_t - \overline{v} v_t) + b^{jk} \ell_t (v_{x_j} \overline{v} - \overline{v}_{x_j} v) \right] \right. \\
 &\quad \left. - \alpha b^{jk} (v_{x_j} \overline{v}_t + \overline{v}_{x_j} v_t) \right. \\
 &\quad \left. + (2b^{jk'} b^{j'k} - b^{jk} b^{j'k'}) \ell_{x_j} (v_{x_{j'}} \overline{v}_{x_{k'}} + \overline{v}_{x_{j'}} v_{x_{k'}}) \right. \\
 &\quad \left. - a b^{j'k'} \ell_{x_{j'} x_{k'}} b^{jk} (v_{x_j} \overline{v} + \overline{v}_{x_j} v) + 2b^{jk} (A \ell_{x_j} - \alpha \ell_{x_j} \ell_t) |v|^2 \right\}.
 \end{aligned} \right.$$

As we shall see later, Theorem 3.2 can be applied to study the controllability/observability as well as the stabilization of parabolic equations and hyperbolic equations. Also, as pointed by [13], starting from Theorem 3.2, one can deduce the controllability/observability for the Schrödinger equation and plate equation appeared in [23] and [48], respectively. Note also that, Theorem 3.2 can be applied to study the controllability of the linear/nonlinear complex Ginzburg-Landau equation (see [13, 15, 38]).

3.3. Controllability/Observability of Linear PDEs. In this subsection, we show that, starting from Theorem 3.2, one can establish sharp observability/controllability results for both parabolic systems and hyperbolic systems.

We need to introduce the following assumptions.

Condition 3.3. Matrix-valued function $(p^{ij})_{1 \leq i, j \leq n} \in C^1(\overline{Q}; \mathbb{R}^{n \times n})$ is uniformly positive definite.

Condition 3.4. Matrix-valued function $(h^{ij})_{1 \leq i, j \leq n} \in C^1(\overline{G}; \mathbb{R}^{n \times n})$ is uniformly positive definite.

Also, for any $N \in \mathbb{N}$, we introduce the following

Condition 3.5. Matrix-valued functions $a \in L^\infty(0, T; L^p(G; \mathbb{R}^{N \times N}))$ for some $p \in [n, \infty]$, and $a_1^1, \dots, a_1^n, a_2 \in L^\infty(Q; \mathbb{R}^{N \times N})$.

Let us consider first the following parabolic system:

$$\begin{cases} \varphi_t - \sum_{i,j=1}^n (p^{ij} \varphi_{x_i})_{x_j} = a\varphi + \sum_{k=1}^n a_1^k \varphi_{x_k}, & \text{in } Q, \\ \varphi = 0, & \text{on } \Sigma, \\ \varphi(0) = \varphi^0, & \text{in } G, \end{cases} \tag{3.6}$$

where φ takes values in \mathbb{R}^N . By choosing $\alpha = 1$ and $\beta = 0$ in Theorem 3.2, one obtains a weighted identity for the parabolic operator. Along with [27], this identity leads to the existing controllability/observability result for parabolic equations ([9, 17]). One can go a little further to show the following result ([10]):

Theorem 3.6. Let Conditions 3.3 and 3.5 hold. Then, solutions of (3.6) satisfy

$$\begin{aligned} & |\varphi(T)|_{(L^2(G))^N} \\ & \leq \exp \left\{ C \left[1 + \frac{1}{T} + T|a|_{L^\infty(0,T;L^p(G;\mathbb{R}^{N \times N}))} + |a|_{L^\infty(0,T;L^p(G;\mathbb{R}^{N \times N}))}^{\frac{1}{\frac{3}{2} - \frac{n}{p}}} \right. \right. \\ & \quad \left. \left. + (1+T) \left(\sum_{k=1}^N |a_i^k|_{L^\infty(Q;\mathbb{R}^{N \times N})} \right)^2 \right] \right\} |\varphi|_{(L^2(Q_{G_0}))^N}, \quad \forall \varphi^0 \in (L^2(G))^N. \end{aligned} \tag{3.7}$$

Note that (3.7) provides the observability inequality for the parabolic system (3.6) with an explicit estimate on the observability constant, depending on the observation time T , the potential a and a_1^k . Earlier result in this respect can be found in [9] and the references cited therein. Inequality (3.7) will play a key role in the study of the null controllability problem for semilinear parabolic equations, as we shall see later.

Remark 3.7. *It is shown in [10] that when $n \geq 2, N \geq 2$ and $(p^{ij})_{1 \leq i, j \leq n} = I$, the exponent $\frac{2}{3}$ in $|a|_{L^\infty(0, T; L^p(G; \mathbb{R}^{N \times N}))}^{\frac{2}{3}}$ (for the case that $p = \infty$ in the inequality (3.7)) is sharp. In [10], it is also proved that the quadratic dependence on $\sum_{k=1}^N |a_1^k|_{L^\infty(Q; \mathbb{R}^{N \times N})}$ is sharp under the same assumptions. However, it is not clear whether the exponent $\frac{3}{2} - \frac{n}{p}$ in $|a|_{L^\infty(0, T; L^p(G; \mathbb{R}^{N \times N}))}^{\frac{1}{\frac{3}{2} - \frac{n}{p}}}$ is optimal when $p < \infty$.*

Next, we consider the following hyperbolic system:

$$\begin{cases} v_{tt} - \sum_{i,j=1}^n (h^{ij} v_{x_i})_{x_j} = av + \sum_{k=1}^n a_1^k v_{x_k} + a_2 v_t, & \text{in } Q, \\ v = 0, & \text{on } \Sigma, \\ v(0) = v^0, \quad v_t(0) = v^1, & \text{in } G, \end{cases} \tag{3.8}$$

where v takes values in \mathbb{R}^N .

Compared with the parabolic case, one needs more assumptions on the coefficient matrix $(h^{ij})_{1 \leq i, j \leq n}$ as follows ([10, 16]):

Condition 3.8. *There is a positive function $d(\cdot) \in C^2(\overline{G})$ satisfying*

i) For some constant $\mu_0 \geq 4$, it holds

$$\sum_{i,j=1}^n \left\{ \sum_{i',j'=1}^n \left[2h^{ij'} (h^{i'j} d_{x_{i'}})_{x_{j'}} - h_{x_{j'}}^{ij} h^{i'j'} d_{x_{i'}} \right] \right\} \xi^i \xi^j \geq \mu_0 \sum_{i,j=1}^n h^{ij} \xi^i \xi^j, \\ \forall (x, \xi^1, \dots, \xi^n) \in \overline{G} \times \mathbb{R}^n;$$

ii) There is no critical point of $d(\cdot)$ in \overline{G} , i.e., $\min_{x \in \overline{G}} |\nabla d(x)| > 0$;

iii) $\frac{1}{4} \sum_{i,j=1}^n h^{ij}(x) d_{x_i}(x) d_{x_j}(x) \geq \max_{x \in \overline{G}} d(x), \quad \forall x \in \overline{G}$.

We put

$$T^* = 2 \max_{x \in \overline{G}} \sqrt{d(x)}, \quad \Gamma^* \triangleq \left\{ x \in \Gamma \mid \sum_{i,j=1}^n h^{ij}(x) d_{x_i}(x) \nu_j(x) > 0 \right\}. \tag{3.9}$$

By choosing $b^{jk}(t, x) \equiv h^{jk}(x)$ and $\alpha = \beta = 0$ in Theorem 3.2 (and noting that only the symmetry condition is assumed for b^{jk} in this theorem), one obtains the fundamental identity derived in [16] to establish the controllability/observability of the general hyperbolic equations. One can go a little further to show the following result ([10]).

Theorem 3.9. *Let Conditions 3.4, 3.5 and 3.8 hold, $T > T^*$ and $G_0 = G \cap \mathcal{O}_\varepsilon(\Gamma^*)$ for some $\varepsilon > 0$. Then one has the following conclusions:*

1) *For any $(v^0, v^1) \in (H_0^1(G))^N \times (L^2(G))^N$, the corresponding weak solution $v \in C([0, T]; (H_0^1(G))^N) \cap C^1([0, T]; (L^2(G))^N)$ of system (3.8) satisfies*

$$\begin{aligned}
 & |v^0|_{H_0^1(G)^N} + |v^1|_{(L^2(G))^N} \\
 & \leq \exp \left[C \left(1 + |a|_{L^\infty(0, T; L^p(G; \mathbb{R}^{N \times N}))}^{\frac{1}{\frac{3}{2} - \frac{n}{p}}} \right. \right. \\
 & \quad \left. \left. + \left(\sum_{k=1}^N |a_i^k|_{L^\infty(Q; \mathbb{R}^{N \times N})} + |a_2|_{L^\infty(Q; \mathbb{R}^{N \times N})} \right)^2 \right) \right] \left| \frac{\partial v}{\partial \nu} \right|_{(L^2((0, T) \times \Gamma^*))^N}. \tag{3.10}
 \end{aligned}$$

2) *If $a_1^k \equiv 0$ ($k = 1, \dots, n$) and $a_2 \equiv 0$, then for any $(v^0, v^1) \in (L^2(G))^N \times (H^{-1}(G))^N$, the weak solution $v \in C([0, T]; (L^2(G))^N) \cap C^1([0, T]; (H^{-1}(G))^N)$ of system (3.8) satisfies*

$$\begin{aligned}
 & |v^0|_{(L^2(G))^N} + |v^1|_{H^{-1}(G)^N} \\
 & \leq \exp \left[C \left(1 + |a|_{L^\infty(0, T; L^p(G; \mathbb{R}^{N \times N}))}^{\frac{1}{\frac{3}{2} - \frac{n}{p}}} \right) \right] |v|_{(L^2(Q_{G_0}))^N}. \tag{3.11}
 \end{aligned}$$

As we shall see in the next subsection, inequality (3.11) plays a crucial role in the study of the exact controllability problem for semilinear hyperbolic equations.

Remark 3.10. *As in the parabolic case, it is shown in [10] that the exponent $\frac{2}{3}$ in the estimate $|a|_{L^\infty(0, T; L^p(G; \mathbb{R}^{N \times N}))}^{\frac{2}{3}}$ in (3.11) (for the special case $p = \infty$) is sharp for $n \geq 2$ and $N \geq 2$. Also, the exponent 2 in the term $\left(\sum_{k=1}^N |a_i^k|_{L^\infty(Q; \mathbb{R}^{N \times N})} + |a_2|_{L^\infty(Q; \mathbb{R}^{N \times N})} \right)^2$ in (3.10) is sharp. However, it is unknown whether the estimate is optimal for the case that $p < \infty$.*

By the standard duality argument, Theorems 3.6 and 3.9 can be applied to deduce the controllability results for parabolic systems and hyperbolic systems, respectively. We omit the details.

3.4. Controllability of Semi-linear PDEs. The study of exact/null controllability problems for semi-linear PDEs began in the 1960s. Early

works in this respect were mainly devoted to the local controllability problem. By the local controllability of a system, we mean that the controllability property holds under some smallness assumptions on the initial data and/or the final target, or the Lipschitz constant of the nonlinearity.

In this subsection we shall present some global controllability results for both semilinear parabolic equations and hyperbolic equations. These results can be deduced from Theorems 3.6 and 3.9, respectively.

Consider first the following controlled semi-linear parabolic equation:

$$\begin{cases} y_t - \sum_{i,j=1}^n (p^{ij} y_{x_i})_{x_j} + f(y, \nabla y) = \chi_{G_0} u, & \text{in } Q, \\ y = 0, & \text{on } \Sigma, \\ y(0) = y_0, & \text{in } G. \end{cases} \tag{3.12}$$

For system (3.12), the state variable and control variable, state space and control space, controllability, are chosen/defined in a similar way as for system (2.1). Concerning the nonlinearity $f(\cdot, \cdot)$, we introduce the following assumption ([9]).

Condition 3.11. *Function $f(\cdot, \cdot) \in C(\mathbb{R}^{1+n})$ is locally Lipschitz-continuous. It satisfies $f(0, 0) = 0$ and*

$$\begin{cases} \lim_{|(s,p)| \rightarrow \infty} \frac{\int_0^1 f_s(\tau s, \tau p) d\tau}{\ln^{\frac{3}{2}}(1 + |s| + |p|)} = 0, \\ \lim_{|(s,p)| \rightarrow \infty} \frac{|\int_0^1 f_{p_1}(\tau s, \tau p) d\tau, \dots, \int_0^1 f_{p_n}(\tau s, \tau p) d\tau|}{\ln^{\frac{1}{2}}(1 + |s| + |p|)} = 0, \end{cases} \tag{3.13}$$

where $p = (p_1, \dots, p_n)$.

As shown in [9] (See [2] and the references therein for earlier results), linearizing the equation, estimating the cost of the control in terms of the size of the potential entering in the system (thanks to Theorem 3.6), and using the classical fixed point argument, one can show the following result.

Theorem 3.12. *Assume that Conditions 3.3 and 3.11 hold. Then system (3.12) is null controllable.*

In particular, Theorem 3.12 provides the possibility of controlling some blowing-up equations. More precisely, assume that $f(s, p) \equiv f(s)$ in system (3.12) has the form

$$f(s) = -s \ln^r(1 + |s|), \quad r \geq 0. \tag{3.14}$$

When $r > 1$, solutions of (3.12), in the absence of control, i.e. with $u \equiv 0$, blow-up in finite time. According to Theorem 3.12 the process can be controlled, and,

in particular, the blow-up can be avoided when $1 < r \leq 3/2$. By the contrary, it is proved in [2, 12] that for some nonlinearities f satisfying

$$\lim_{|s| \rightarrow \infty} \frac{|f(s)|}{s \ln^r(1 + |s|)} = 0, \quad (3.15)$$

where $r > 2$, the corresponding system is not controllable. The reason is that the controls cannot help the system to avoid blow-up.

Remark 3.13. *It is still an unsolved problem whether the controllability holds for system (3.12) in which the nonlinear function $f(\cdot)$ satisfies (3.15) with $3/2 \leq r \leq 2$. Note that, the growth condition in (3.13) comes from the observability inequality (3.7). Indeed, the logarithmic function in (3.13) is precisely the inverse of the exponential one in (3.7). According to Remark 3.7, the estimate (3.7) cannot be improved, and therefore, the usual linearization approach cannot lead to any improvement of the growth condition (3.13).*

Next, we consider the following controlled semi-linear hyperbolic equation:

$$\begin{cases} y_{tt} - \sum_{i,j=1}^n (h^{ij} y_{x_i})_{x_j} = h(y) + \chi_{G_0} u & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y(0) = y_0, \quad y_t(0) = y_1 & \text{in } G. \end{cases} \quad (3.16)$$

For system (3.16), the state variable and control variable, state space and control space, controllability, are chosen/defined in a similar way as that for system (2.2). Concerning the nonlinearity $h(\cdot)$, we need the following assumption ([10]).

Condition 3.14. *Function $h(\cdot) \in C(\mathbb{R})$ is locally Lipschitz-continuous, and for some $r \in [0, \frac{3}{2})$, it satisfies that*

$$\lim_{|s| \rightarrow \infty} \frac{\int_0^1 h_s(\tau s) d\tau}{\ln^r(1 + |s|)} = 0. \quad (3.17)$$

As mentioned in [10], proceeding as in the proof of [16, Theorem 2.2], i.e., by the linearization approach (thanks to the second conclusion in Theorem 3.9), noting that the embedding $H_0^1(G) \hookrightarrow L^2(G)$ is compact, and using the fixed point technique, one can show the following result.

Theorem 3.15. *Assume that Conditions 3.4, 3.8 and 3.14 are satisfied, and T and G_0 are given as in Theorem 3.9. Then system (3.12) is exactly controllable.*

Due to the blow-up and the finite propagation speed of solutions to hyperbolic equations, one cannot expect exact controllability of system (3.12) for nonlinearities of the form (3.17) with $r > 2$. One could expect the system to be controllable for $r \leq 2$. However, in view of Remark 3.10, the usual fixed point method cannot be applied for $r \geq 3/2$. Therefore, when $n \geq 2$, the controllability problem for system (3.16) is open for $3/2 \leq r \leq 2$.

Remark 3.16. Note that the above “3/2 logarithmic growth” phenomenon (arising in the global exact controllability for nonlinear PDEs) does not occur in the pure PDE problem, and therefore the study of nonlinear controllability is of independent interest. More precisely, this means that for the controllability problem of nonlinear systems, there exist some extra difficulties.

3.5. Controllability of Quasilinear PDEs. In this subsection, we consider the controllability of quasilinear parabolic/hyperbolic equations.

We begin with the following controlled quasilinear hyperbolic equation:

$$\begin{cases} y_{tt} - \sum_{i,j=1}^n (h^{ij} y_{x_i})_{x_j} = F(t, x, y, \nabla_{t,x} y, \nabla_{t,x}^2 y) + qy + \phi_{G_0} u, & \text{in } Q, \\ y = 0, & \text{on } \Sigma, \\ y(0) = y_0, \quad y_t(0) = y_1, & \text{in } G. \end{cases} \tag{3.18}$$

Here, $(h^{ij})_{1 \leq i,j \leq n} \in H^{s+1}(G; \mathbb{R}^{n \times n})$ and $q \in H^s(Q)$ with $s > \frac{n}{2} + 1$, and similar to [54], the nonlinear term $F(\cdot)$ has the form

$$F(t, x, y, \nabla_{t,x} y, \nabla_{t,x}^2 y) = \sum_{i=1}^n \sum_{\alpha=0}^n f_{i\alpha}(t, x, \nabla_{t,x} y) y_{x_i x_\alpha} + O(|y|^2 + |\nabla_{t,x} y|^2),$$

where $f_{i\alpha}(t, x, 0) = 0$ and $x_0 = t$, ϕ_{G_0} is a nonnegative smooth function defined on \bar{G} and satisfying $\min_{x \in \bar{G}_0} \phi(x) > 0$. In system (3.18), as before, (y, y_t) is the *state variable* and u is the *control variable*. However, as we shall see later, the state space and control space have to be chosen in a different way from those used in the linear/semilinear setting.

The controllability of quasilinear hyperbolic equations is well understood in one space dimension ([26]). With regard to the multidimensional case, we introduce the following assumption.

Condition 3.17. The linear part in (3.18), i.e., hyperbolic equation

$$\begin{cases} y_{tt} - \sum_{i,j=1}^n (h^{ij} y_{x_i})_{x_j} = qy + \chi_{G_0} u, & \text{in } Q, \\ y = 0, & \text{in } \Sigma, \\ y(0) = y_0, \quad y_t(0) = y_1, & \text{in } G \end{cases} \tag{3.19}$$

is exactly controllable in $H_0^1(G) \times L^2(G)$ at some time T .

Theorem 3.9 provides a sufficient condition to guarantee Condition 3.17 is satisfied. The following result is a slight generalization of that shown in [52].

Theorem 3.18. *Assume Condition 3.17 holds. Then, there is a sufficiently small $\varepsilon_0 > 0$ such that for any $(y_0, y_1), (z_0, z_1) \in (H^{s+1}(G) \cap H_0^1(G)) \times H^s(G)$ satisfying $|(y_0, y_1)|_{H^{s+1}(G) \times H^s(G)} < \varepsilon_0$, $|(z_0, z_1)|_{H^{s+1}(G) \times H^s(G)} < \varepsilon_0$ and the compatibility condition, one can find a control $u \in \bigcap_{k=0}^{s-2} C^k([0, T]; H^{s-k}(G))$ such that the corresponding solution of system (3.18) verifies $y(T) = z_0$ and $y_t(T) = z_1$ in G .*

The key in the proof of Theorem 3.18 is to reduce the local exact controllability of quasilinear equations to the exact controllability of the linear equation by means of a new unbounded perturbation technique (developed in [52]), which is a universal approach to solve the local controllability problem for a large class of quasilinear time-reversible evolution equations.

Note however that the above approach does not apply to the controllability problem for quasilinear time-irreversible evolution equations, such as the following controlled quasilinear parabolic equation:

$$\begin{cases} y_t - \sum_{i,j=1}^n (a^{ij}(y)y_{x_i})_{x_j} = \chi_{G_0} u & \text{in } Q, \\ y = 0 & \text{on } \Sigma, \\ y(0) = y_0 & \text{in } G. \end{cases} \tag{3.20}$$

In (3.20), y is the *state variable* and u is the *control variable*, the nonlinear matrix-valued function $(a^{ij})_{1 \leq i, j \leq n} \in C^2(\mathbb{R}; \mathbb{R}^{n \times n})$ is locally positive definite. One can find very limited papers on the controllability of quasilinear parabolic-type equations ([31] and the references therein). One of the main difficulty to solve this problem is to show the “good enough” regularity for solutions of system (3.20) with a desired control.

We introduce the dual system of the linearized equation of (3.20).

$$\begin{cases} p_t - \sum_{i,j=1}^n (p^{ij}p_{x_i})_{x_j} = 0 & \text{in } Q, \\ p = 0 & \text{on } \Sigma, \\ p(0) = p_0 & \text{in } G, \end{cases} \tag{3.21}$$

where $(p^{ij})_{1 \leq i, j \leq n}$ is assumed to satisfy Condition 3.3. Put $B = 1 + \sum_{i,j=1}^n |p^{ij}|_{C^1(\overline{Q})}^2$. Starting from Theorem 3.2, one can show the following observability result ([31]).

Theorem 3.19. *There exist suitable real functions α and φ , and a constant $C_0 = C_0(\rho, n, G, T) > 0$, such that for any $\lambda \geq C_0 e^{C_0 B}$, solutions of (3.21) satisfy*

$$|p(T)|_{L^2(G)} \leq C e^{e^{C_0 B}} |e^{\lambda \alpha} |\varphi|^{3/2} p|_{L^2(Q_{G_0})}, \quad \forall p_0 \in L^2(G). \quad (3.22)$$

In Theorem 3.19, the observability constant in (3.22) is obtained explicitly in the form of $C e^{e^{C_0 B}}$ in terms of the C^1 -norms of the coefficients in the principal operator appeared in the first equation of (3.21). This is the key in the argument of fixed point technique to show the following local controllability of system (3.20) ([31]).

Theorem 3.20. *There is a constant $\gamma > 0$ such that, for any initial value $y_0 \in C^{2+\frac{1}{2}}(\overline{G})$ satisfying $|y_0|_{C^{2+\frac{1}{2}}(\overline{G})} \leq \gamma$ and the first order compatibility condition, one can find a control $u \in C^{\frac{1}{2}, \frac{1}{4}}(\overline{Q})$ with $\text{supp } u \subseteq [0, T] \times G_0$ such that the solution y of system (3.20) satisfies $y(T) = 0$ in G .*

From Theorem 3.20, it is easy to see that the *state space* and *control space* for system (3.20) are chosen to be $C^{2+\frac{1}{2}}(\overline{G})$ and $C^{\frac{1}{2}, \frac{1}{4}}(\overline{Q})$, respectively. The key observation in [31] is that, thanks to an idea in [2], for smooth initial data, the regularity of the null-control function for the linearized system can be improved, and therefore, the fixed point method is applicable.

3.6. Stabilization of hyperbolic equations and further comments. In this subsection, we give more applications of Theorem 3.2 to the stabilization of hyperbolic equations and comment other applications of this theorem and some related open problems.

One of the main motivation to introduce the controllability/obseervability theory is to design the feedback regulator ([19]). Stimulated by [29], there exist a lot of works addressing the stabilization problem of PDEs from the control point of view. To begin with, we fix a nonnegative function $a \in L^\infty(\Gamma)$ such that $\{x \in \Gamma \mid a(x) > 0\} \neq \emptyset$, and consider the following hyperbolic equation with a boundary damping:

$$\begin{cases} u_{tt} - \sum_{j,k=1}^n (h^{jk} u_{x_j})_{x_k} = 0 & \text{in } (0, \infty) \times G, \\ \sum_{j,k=1}^n h^{jk} u_{x_j} \nu_k + a(x) u_t = 0 & \text{on } (0, \infty) \times \Gamma, \\ u(0) = u^0, \quad u_t(0) = u^1 & \text{in } G. \end{cases} \quad (3.23)$$

Put $H \triangleq \{(f, g) \in H^1(G) \times L^2(G) \mid \int_G f dx = 0\}$, which is a Hilbert space with the canonic norm. Define an unbounded operator $\mathcal{A} : H \rightarrow H$ by

$$\left\{ \begin{aligned} \mathcal{A} &\triangleq \begin{pmatrix} 0 & I \\ \sum_{j,k=1}^n \partial_{x_k} (h^{jk} \partial_{x_j}) & 0 \end{pmatrix}, \\ D(\mathcal{A}) &\triangleq \left\{ u = (u^0, u^1) \in H \mid \mathcal{A}u \in H, \left(\sum_{j,k=1}^n h^{jk} u_{x_j}^0 \nu_k + au^1 \right) \Big|_{\Gamma} = 0 \right\}. \end{aligned} \right.$$

It is easy to show that \mathcal{A} generates an C_0 -semigroup $\{e^{t\mathcal{A}}\}_{t \in \mathbb{R}}$ on H . Hence, system (3.23) is well-posed in H . Clearly, H is the *finite energy space* of system (3.23). One can show that the energy of any solution of (3.23) tends to zero as $t \rightarrow \infty$ (There is no any geometric conditions on Γ).

Starting from Theorem 3.2, one can show the following result, which is a slight improvement of the main result in [14]:

Theorem 3.21. *Assume Conditions 3.4 holds. Then solutions $u \in C([0, \infty); D(\mathcal{A})) \cap C^1([0, \infty); H)$ of system (3.23) satisfy*

$$\|(u, u_t)\|_H \leq \frac{C}{\ln(2+t)} \|(u^0, u^1)\|_{D(\mathcal{A})}, \quad \forall (u^0, u^1) \in D(\mathcal{A}), \quad \forall t > 0. \quad (3.24)$$

Next, we consider a semilinear hyperbolic equation with a local damping:

$$\left\{ \begin{aligned} u_{tt} - \sum_{j,k=1}^n (h^{jk} u_{x_j})_{x_k} + f(u) + b(x)g(u_t, \nabla u) &= 0 && \text{in } (0, \infty) \times G, \\ u = 0 &&& \text{on } (0, \infty) \times \Gamma, \\ u(0) = u^0, \quad u_t(0) = u^1 &&& \text{in } G. \end{aligned} \right. \quad (3.25)$$

In (3.25), h^{jk} is supposed to satisfy Conditions 3.4 and 3.8; $f : \mathbb{R} \rightarrow \mathbb{R}$ is a differentiable function satisfying $f(0) = 0$, $sf(s) \geq 0$ and $|f'(s)| \leq C(1 + |s|^q)$ for any $s \in \mathbb{R}$, where $q \geq 0$ and $(n - 2)q \leq 2$; b is a nonnegative function satisfying $\min_{x \in G_0} b(x) > 0$, where G_0 is given in Theorem 3.9; and $g : \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ is a globally Lipschitz function satisfying $g(0, w) = 0$, $|g(r, w) - g(r_1, w_1)| \leq C(|r - r_1| + |w - w_1|)$ and $g(r, w)r \geq c_0 r^2$ for some $c_0 > 0$, any $w, w_1 \in \mathbb{R}^n$ and any $r, r_1 \in \mathbb{R}$.

Define the energy of any solution u to (3.25) by setting

$$E(t) = \frac{1}{2} \int_G \left[|u_t|^2 + \sum_{j,k=1}^n h^{jk} u_{x_j} u_{x_k} \right] dx + \int_G \int_0^u f(s) ds dx.$$

Starting from Theorem 3.2, one can show the following stabilization result for system (3.25) ([42]).

Theorem 3.22. *Let $(u^0, u^1) \in H_0^1(G) \times L^2(G)$. Then there exist positive constants M and r , possibly depending on $E(0)$, such that the energy $E(t)$ of the solution of (3.25) satisfies $E(t) \leq Me^{-rt}E(0)$ for any $t \geq 0$.*

Several comments are in order.

Remark 3.23. *In [25], the authors need C^∞ -regularity for the data to establish Theorem 2.2. Recently, based on Theorem 3.2, this result was extended in [33] as follows: Denote by $\{\lambda_i\}_{i=1}^\infty$ the eigenvalues of any general elliptic operator of second order (with C^1 -principal part coefficients) on Ω (of class C^2) with Dirichlet or Robin boundary condition, and $\{e_i\}_{i=1}^\infty$ the corresponding eigenvectors satisfying $\|e_i\|_{L^2(\Omega)} = 1$. Then, for any $r > 0$, it holds*

$$\sum_{\lambda_i \leq r} |a_i|^2 \leq Ce^{C\sqrt{r}} \int_{G_0} \left| \sum_{\lambda_i \leq r} a_i e_i(x) \right|^2 dx, \quad \forall \{a_i\}_{\lambda_i \leq r} \text{ with } a_i \in \mathbb{C}.$$

Remark 3.24. *As indicated in [13, 22, 23], Theorem 3.2 can be employed to study the global unique continuation and inverse problems for some PDEs. Note also that this Carleman estimate based approach can be applied to solve some optimal control problems ([45]).*

Remark 3.25. *In practice, constrained controllability is more realizable. It is shown in [37] that the study of this problem is unexpectedly difficult even for the 1 – d wave equation and heat equation. We refer to [30] for an interesting example showing that this problem is nontrivial even if the control is effective everywhere in the domain in which the system is evolved.*

Remark 3.26. *Note that the above mentioned approach applies mainly to the controllability, observability and stabilization of second order non-degenerate PDEs. It is quite interesting to extend it to the coupled and/or higher order systems, or degenerate systems but in general, this is nontrivial even for linear problems ([7, 53]).*

Remark 3.27. *Similar to other nonlinear problems, nonlinear controllability problems are usually quite difficult. It seems that there is no satisfactory controllability results published for nonlinear hyperbolic-parabolic coupled equations. Also, there exists no controllability results for fully nonlinear PDEs. In the general case, of course, one could expect only local results. Therefore, the following three problems deserve deep studies: 1) The characterization of the controllability subspace; 2) Controllability problem with (sharp) lower regularity for the*

data; 3) The problem that cannot be linearized. Of course, all of these problems are usually challenging.

4. The Stochastic Case

In this section, we extend some of the results/approaches in Section 3 to the stochastic case. As we shall see later, the stochastic counterpart is far from satisfactory, compared to the deterministic setting.

In what follows, we fix a complete filtered probability space $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t \geq 0}, P)$ on which a one dimensional standard Brownian motion $\{B(t)\}_{t \geq 0}$ is defined. Let H be a Fréchet space. Denote by $L^2_{\mathcal{F}}(0, T; H)$ the Fréchet space consisting of all H -valued $\{\mathcal{F}_t\}_{t \geq 0}$ -adapted processes $X(\cdot)$ such that $\mathbb{E}(|X(\cdot)|^2_{L^2(0, T; H)}) < \infty$, with the canonical quasi-norms; by $L^\infty_{\mathcal{F}}(0, T; H)$ the Fréchet space consisting of all H -valued $\{\mathcal{F}_t\}_{t \geq 0}$ -adapted bounded processes, with the canonical quasi-norms; and by $L^2_{\mathcal{F}}(\Omega; C([0, T]; H))$ the Fréchet space consisting of all H -valued $\{\mathcal{F}_t\}_{t \geq 0}$ -adapted continuous processes $X(\cdot)$ such that $\mathbb{E}(|X(\cdot)|^2_{C([0, T]; H)}) < \infty$, with the canonical quasi-norms (similarly, one can define $L^2_{\mathcal{F}}(\Omega; C^k([0, T]; H))$ for $k \in \mathbb{N}$).

4.1. Stochastic Parabolic Equations. We begin with the following stochastic parabolic equation:

$$\begin{cases} dz - \sum_{i,j=1}^n (p^{ij} z_{x_i})_{x_j} dt = [\langle a, \nabla z \rangle + bz] dt + cz dB(t) & \text{in } Q, \\ z = 0 & \text{on } \Sigma, \\ z(0) = z_0 & \text{in } G \end{cases} \tag{4.1}$$

with suitable coefficients a, b and c , where $p^{ij} \in C^2(\bar{Q})$ is assumed to satisfy Condition 3.3 (Note that, technically we need here more regularity for p^{ij} than the deterministic case). We are concerned with an observability estimate for system (4.1), i.e., to find a constant $C = C(a, b, c, T) > 0$ such that solutions of (4.1) satisfy

$$|z(T)|_{L^2(\Omega, \mathcal{F}_T, P; L^2(G))} \leq C |z|_{L^2_{\mathcal{F}}(0, T; L^2(G_0))}, \quad \forall z_0 \in L^2(\Omega, \mathcal{F}_0, P; L^2(G)). \tag{4.2}$$

Similar to Theorem 3.2, we have the following weighted identity ([41]).

Theorem 4.1. *Let $m \in \mathbb{N}$, $b^{ij} = b^{ji} \in L^2_{\mathcal{F}}(\Omega; C^1([0, T]; W^{2,\infty}(\mathbb{R}^m)))$ ($i, j = 1, 2, \dots, m$), $\ell \in C^{1,3}((0, T) \times \mathbb{R}^m)$ and $\Psi \in C^{1,2}((0, T) \times \mathbb{R}^m)$. Assume u is an $H^2(\mathbb{R}^m)$ -valued continuous semi-martingale. Set $\theta = e^\ell$ and $v = \theta u$. Then for*

a.e. $x \in \mathbb{R}^m$ and P -a.s. $\omega \in \Omega$,

$$\begin{aligned} & 2 \int_0^T \theta \left[- \sum_{i,j=1}^m (b^{ij} v_{x_i})_{x_j} + Av \right] \left[du - \sum_{i,j=1}^m (b^{ij} u_{x_i})_{x_j} dt \right] + 2 \int_0^T \sum_{i,j=1}^m (b^{ij} v_{x_i} dv)_{x_j} \\ & + 2 \int_0^T \sum_{i,j=1}^m \left[\sum_{i',j'=1}^m \left(2b^{ij} b^{i'j'} \ell_{x_{i'}} v_{x_i} v_{x_{j'}} - b^{ij} b^{i'j'} \ell_{x_i} v_{x_{i'}} v_{x_{j'}} \right) \right. \\ & \left. + \Psi b^{ij} v_{x_i} v - b^{ij} \left(A \ell_{x_i} + \frac{\Psi x_i}{2} \right) v^2 \right]_{x_j} dt \\ & = 2 \int_0^T \sum_{i,j=1}^m \left\{ \sum_{i',j'=1}^m \left[2b^{ij'} (b^{i'j} \ell_{x_{i'}})_{x_{j'}} - (b^{ij} b^{i'j'} \ell_{x_{i'}})_{x_{j'}} \right] - \frac{b_t^{ij}}{2} + \Psi b^{ij} \right\} v_{x_i} v_{x_j} dt \\ & + \int_0^T B v^2 dt + 2 \int_0^T \left[- \sum_{i,j=1}^m (b^{ij} v_{x_i})_{x_j} + Av \right] \left[- \sum_{i,j=1}^m (b^{ij} v_{x_i})_{x_j} + (A - \ell_t) v \right] dt \\ & + \left(\sum_{i,j=1}^m b^{ij} v_{x_i} v_{x_j} + Av^2 \right) \Big|_0^T \\ & - \int_0^T \theta^2 \sum_{i,j=1}^m b^{ij} [(du_{x_i} + \ell_{x_i} du)(du_{x_j} + \ell_{x_j} du)] - \int_0^T \theta^2 A (du)^2, \end{aligned}$$

where

$$\begin{cases} A \triangleq - \sum_{i,j=1}^m (b^{ij} \ell_{x_i} \ell_{x_j} - b_{x_j}^{ij} \ell_{x_i} - b^{ij} \ell_{x_i x_j}) - \Psi, \\ B \triangleq 2 \left[A \Psi - \sum_{i,j=1}^m (A b^{ij} \ell_{x_i})_{x_j} \right] - A_t - \sum_{i,j=1}^m (b^{ij} \Psi_{x_j})_{x_i}. \end{cases}$$

Remark 4.2. Note that, in Theorem 4.1, we assume only the symmetry for matrix $(b^{ij})_{1 \leq i,j \leq n}$ (without assuming the positive definiteness). Hence, this theorem can be applied to study not only the observability/controllability of stochastic parabolic equations, but also similar problems for deterministic parabolic and hyperbolic equations, as indicated in Section 3. In this way, we give a unified treatment of controllability/observability problems for some stochastic and deterministic PDEs of second order.

Starting from Theorem 4.1, one can show the following observability result in [41] (See [3] and the references therein for some earlier results).

Theorem 4.3. Assume that

$$a \in L_{\mathcal{F}}^\infty(0, T; L^\infty(G; \mathbb{R}^n)), \quad b \in L_{\mathcal{F}}^\infty(0, T; L^{n^*}(G)), \quad c \in L_{\mathcal{F}}^\infty(0, T; W^{1,\infty}(G)),$$

where $n^* \geq 2$ if $n = 1$; $n^* > 2$ if $n = 2$; $n^* \geq n$ if $n \geq 3$. Then there is a constant $\mathcal{C} = \mathcal{C}(a, b, c, T) > 0$ such that all solutions z of system (4.1) satisfy

(4.2). Moreover, the observability constant \mathcal{C} may be bounded as

$$\mathcal{C}(a, b, c, T) = Ce^{C[T^{-4}(1+\tau^2)+T\tau^2]},$$

with $\tau \triangleq |a|_{L^\infty_{\mathcal{F}}(0,T;L^\infty(G;\mathbb{R}^n))} + |b|_{L^\infty_{\mathcal{F}}(0,T;L^{n^*}(G))} + |c|_{L^\infty_{\mathcal{F}}(0,T;W^{1,\infty}(G))}$.

As a consequence of Theorem 4.3, one can deduce a controllability result for backward stochastic parabolic equations. Unlike the deterministic case, the study of controllability problems for forward stochastic differential equations is much more difficult than that for the backward ones. We refer to [35] for some important observation in this respect. It deserves to mention that, as far as I know, there exists no satisfactory controllability result published for forward stochastic parabolic equations. Note however that, as a consequence of Theorem 2.2 and its generalization (see Remark 3.23), one can deduce a null controllability result for forward stochastic parabolic equations with time-invariant coefficients ([33]).

Theorem 4.1 has another application in global unique continuation of stochastic PDEs. To see this, we consider the following stochastic parabolic equation:

$$\mathcal{F}z \equiv dz - \sum_{i,j=1}^n (f^{ij} z_{x_i})_{x_j} dt = [\langle a_1, \nabla z \rangle + b_1 z] dt + c_1 z dB(t) \quad \text{in } Q, \quad (4.3)$$

where $f^{ij} \in C^{1,2}([0, T] \times G)$ satisfy $f^{ij} = f^{ji}$ ($i, j = 1, 2, \dots, n$) and for any open subset G_1 of G , there is a constant $s_0 = s_0(G_1) > 0$ so that $\sum_{i,j=1}^n f^{ij} \xi^i \xi^j \geq s_0 |\xi|^2$ for all $(t, x, \xi) \equiv (t, x, \xi^1, \dots, \xi^n) \in (0, T) \times G_1 \times \mathbb{R}^n$; $a_1 \in L^\infty_{\mathcal{F}}(0, T; L^\infty_{loc}(G; \mathbb{R}^n))$, $b_1 \in L^\infty_{\mathcal{F}}(0, T; L^\infty_{loc}(G))$, and $c_1 \in L^\infty_{\mathcal{F}}(0, T; W^{1,\infty}_{loc}(G))$.

Starting from Theorem 4.1, one can show the following result ([50]).

Theorem 4.4. Any solution $z \in L^2_{\mathcal{F}}(\Omega; C([0, T]; L^2_{loc}(G))) \cap L^2_{\mathcal{F}}(0, T; H^1_{loc}(G))$ of (4.3) vanishes identically in $Q \times \Omega$, a.s. dP provided that $z = 0$ in $Q_{G_0} \times \Omega$, a.s. dP .

Note that the solution of a stochastic equation is generally *non-analytic in time* even if all coefficients of the equation are constants. Therefore, one cannot expect a Holmgren-type uniqueness theorem for stochastic equations except for some very special cases. On the other hand, the usual approach to employ Carleman-type estimate for the unique continuation needs to localize the problem. The difficulty of our present stochastic problem consists in the fact that *one cannot simply localize the problem as usual* because the usual localization technique may change the adaptedness of solutions, which is a key feature in the stochastic setting. In equation (4.3), for the space variable x , we may proceed as in the classical argument. However, for the time variable t , due

to the adaptedness requirement, we will have to treat it separately and globally. We need to introduce *partial global Carleman estimate* (indeed, global in time) even for local unique continuation for stochastic parabolic equation. Note that this idea comes from the study of controllability problem even though unique continuation itself is purely an PDE problem.

4.2. Stochastic Hyperbolic Equations. We consider now the following stochastic wave equation:

$$\begin{cases} dz_t - \Delta z dt = (a_1 z_t + \langle a_2, \nabla z \rangle + a_3 z + f) dt + (a_4 z + g) dB(t) & \text{in } Q, \\ z = 0 & \text{on } \Sigma, \\ z(0) = z_0, \quad z_t(0) = z_1 & \text{in } G, \end{cases} \tag{4.4}$$

where $a_1 \in L^\infty_{\mathcal{F}}(0, T; L^\infty(G))$, $a_2 \in L^\infty_{\mathcal{F}}(0, T; L^\infty(G; \mathbb{R}^n))$, $a_3 \in L^\infty_{\mathcal{F}}(0, T; L^n(G))$, $a_4 \in L^\infty_{\mathcal{F}}(0, T; L^\infty(G))$, $f \in L^2_{\mathcal{F}}(0, T; L^2(G))$, $g \in L^2_{\mathcal{F}}(0, T; L^2(G))$ and $(z_0, z_1) \in L^2(\Omega, \mathcal{F}_0, P; H^1_0(G) \times L^2(G))$. We shall derive an observability estimate for (4.4), i.e., find a constant $\mathcal{C}(a_1, a_2, a_3, a_4) > 0$ such that solutions of system (4.4) satisfy

$$\begin{aligned} & |(z(T), z_t(T))|_{L^2(\Omega, \mathcal{F}_T, P; H^1_0(G) \times L^2(G))} \\ & \leq \mathcal{C}(a_1, a_2, a_3, a_4) \left[\left| \frac{\partial z}{\partial \nu} \right|_{L^2_{\mathcal{F}}(0, T; L^2(\Gamma_0))} + |f|_{L^2_{\mathcal{F}}(0, T; L^2(G))} + |g|_{L^2_{\mathcal{F}}(0, T; L^2(G))} \right], \\ & \quad \forall (z_0, z_1) \in L^2(\Omega, \mathcal{F}_0, P; H^1_0(G) \times L^2(G)), \end{aligned} \tag{4.5}$$

where Γ_0 is given by (2.7) for some $x_0 \in \mathbb{R}^d \setminus \overline{G}$.

It is clear that, $0 < R_0 \triangleq \min_{x \in G} |x - x_0| < R_1 \triangleq \max_{x \in G} |x - x_0|$. We choose a sufficiently small constant $c \in (0, 1)$ so that $\frac{(4+5c)R_0^2}{9c} > R_1^2$. In what follows, we take T sufficiently large such that $\frac{4(4+5c)R_0^2}{9c} > c^2 T^2 > 4R_1^2$. Our observability estimate for system (4.4) is stated as follows ([51]).

Theorem 4.5. *Solutions of system (4.4) satisfy (4.5) with*

$$\begin{aligned} & \mathcal{C}(a_1, a_2, a_3, a_4) \\ & = C \exp \left\{ C \left[|(a_1, a_4)|_{L^\infty_{\mathcal{F}}(0, T; (L^\infty(G))^2)}^2 + |a_2|_{L^\infty_{\mathcal{F}}(0, T; L^\infty(G; \mathbb{R}^n))}^2 + |a_3|_{L^\infty_{\mathcal{F}}(0, T; L^n(G))}^2 \right] \right\}. \end{aligned}$$

Surprisingly, Theorem 4.5 was improved in [33] by replacing the left hand side of (4.5) by $|(z_0, z_1)|_{L^2(\Omega, \mathcal{F}_0, P; H^1_0(G) \times L^2(G))}$, exactly in a way of the deterministic setting. This is highly nontrivial by considering the very fact that the stochastic wave equation is time-irreversible.

The proof of Theorem 4.5 (and its improvement in [33]) is based on the following identity for a stochastic hyperbolic-like operator, which is in the spirit of Theorems 3.2 and 4.1.

Theorem 4.6. Let $b^{ij} \in C^1((0, T) \times \mathbb{R}^n)$ satisfy $b^{ij} = b^{ji}$ ($i, j = 1, 2, \dots, n$), $\ell, \Psi \in C^2((0, T) \times \mathbb{R}^n)$. Assume u is an $H^2_{loc}(\mathbb{R}^n)$ -valued $\{\mathcal{F}_t\}_{t \geq 0}$ -adapted process such that u_t is an $L^2_{loc}(\mathbb{R}^n)$ -valued semimartingale. Set $\theta = e^\ell$ and $v = \theta u$. Then, for a.e. $x \in \mathbb{R}^n$ and P -a.s. $\omega \in \Omega$,

$$\begin{aligned} & \theta \left(-2\ell_t v_t + 2 \sum_{i,j=1}^n b^{ij} \ell_{x_i} v_{x_j} + \Psi v \right) \left[du_t - \sum_{i,j=1}^n (b^{ij} u_{x_i})_{x_j} dt \right] \\ & + \sum_{i,j=1}^n \left[\sum_{i',j'=1}^n \left(2b^{ij} b^{i'j'} \ell_{x_i} v_{x_i} v_{x_j} - b^{ij} b^{i'j'} \ell_{x_i} v_{x_i'} v_{x_j} \right) - 2b^{ij} \ell_t v_{x_i} v_t + b^{ij} \ell_{x_i} v_t^2 \right. \\ & \quad \left. + \Psi b^{ij} v_{x_i} v - \left(A \ell_{x_i} + \frac{\Psi_{x_i}}{2} \right) b^{ij} v^2 \right]_{x_j} dt \\ & + d \left[\sum_{i,j=1}^n b^{ij} \ell_t v_{x_i} v_{x_j} - 2 \sum_{i,j=1}^n b^{ij} \ell_{x_i} v_{x_j} v_t + \ell_t v_t^2 - \Psi v_t v + \left(A \ell_t + \frac{\Psi_t}{2} \right) v^2 \right] \\ & = \left\{ \left[\ell_{tt} + \sum_{i,j=1}^n (b^{ij} \ell_{x_i})_{x_j} - \Psi \right] v_t^2 - 2 \sum_{i,j=1}^n \left[(b^{ij} \ell_{x_j})_t + b^{ij} \ell_{tx_j} \right] v_{x_i} v_t \right. \\ & \quad \left. + \sum_{i,j=1}^n \left[(b^{ij} \ell_t)_t + \sum_{i',j'=1}^n \left(2b^{ij'} (b^{i'j} \ell_{x_i'})_{x_j'} - (b^{ij} b^{i'j'} \ell_{x_i'})_{x_j'} \right) + \Psi b^{ij} \right] v_{x_i} v_{x_j} \right. \\ & \quad \left. + B v^2 + \left(-2\ell_t v_t + 2 \sum_{i,j=1}^n b^{ij} \ell_{x_i} v_{x_j} + \Psi v \right)^2 \right\} dt + \theta^2 \ell_t (du_t)^2, \end{aligned}$$

where $(du_t)^2$ denotes the quadratic variation process of u_t ,

$$\begin{cases} A \triangleq (\ell_t^2 - \ell_{tt}) - \sum_{i,j=1}^n (b^{ij} \ell_{x_i} \ell_{x_j} - b_{x_j}^{ij} \ell_{x_i} - b^{ij} \ell_{x_i x_j}) - \Psi, \\ B \triangleq A \Psi + (A \ell_t)_t - \sum_{i,j=1}^n (A b^{ij} \ell_{x_i})_{x_j} + \frac{1}{2} \left[\Psi_{tt} - \sum_{i,j=1}^n (b^{ij} \Psi_{x_i})_{x_j} \right]. \end{cases}$$

4.3. Further comments. Compared to the deterministic case, the controllability/observability of stochastic PDEs is in its “enfant” stage. Therefore, the main concern of the controllability/observability theory in the near future should be that for stochastic PDEs. Some most relevant open problems are listed below.

- **Controllability of forward stochastic PDEs.** Very little is known although there are some significant progress in the recent work [33]. Also,

it would be quite interesting to extend the result in [4] to the stochastic setting but this seems to be highly nontrivial.

- **Controllability of nonlinear stochastic PDEs.** Almost nothing is known in this direction although there are some papers addressing the problem in abstract setting by imposing some assumption which is usually very difficult to check for the nontrivial case.
- **Stabilization and inverse problems for stochastic PDEs.** Almost nothing is known in this respect.

Acknowledgement

This paper is a summary of part of the work I have done in close collaboration with my colleagues, coworkers and students. I am grateful to all of them. In particular, I would like to express my sincerely gratitude to my mentors, Xunjing Li, Jiongmin Yong and Enrique Zuazua, for their continuous encouragement and for so many fruitful discussions and suggestions that have been extremely influential on the formulation and solution of most of problems addressed here. Also, I would like to mention some of my colleagues with who I had the opportunity to develop part of the theory and learn so many things and, in particular, Xiaoyu Fu, Xu Liu, Qi Lü, Kim Dang Phung, Shanjian Tang, Gengsheng Wang, Jiongmin Yong and Enrique Zuazua. Finally, I thank Viorel Barbu, Piermarco Cannarsa, Xiaoyu Fu, Yamilet Quintana, Louis Tébou, Marius Tucsnak and Gengsheng Wang for their useful comments on the first version of this paper that allowed to improve its presentation and to avoid some inaccuracies.

References

- [1] S.A. Avdonin and S.A. Ivanov, *Families of exponentials. The method of moments in controllability problems for distributed parameter systems*, Cambridge University Press, Cambridge, UK, 1995.
- [2] V. Barbu, *Controllability of parabolic and Navier-Stokes equations*, Sci. Math. Jpn., **56** (2002), 143–211.
- [3] V. Barbu, A. Răscanu and G. Tessitore, *Carleman estimate and controllability of linear stochastic heat equations*, Appl. Math. Optim., **47** (2003), 97–120.
- [4] C. Bardos, G. Lebeau and J. Rauch, *Sharp sufficient conditions for the observation, control and stabilization of waves from the boundary*, SIAM J. Control Optim., **30** (1992), 1024–1065.
- [5] N. Burq, *Contrôle de l'équation des ondes dans des ouverts peu réguliers*, Asymptot. Anal., **14** (1997), 157–191.
- [6] N. Burq and P. Gérard, *Condition nécessaire et suffisante pour la contrôlabilité exacte des ondes*, C. R. Acad. Sci. Paris Sér. I Math., **325** (1997), 749–752.

- [7] P. Cannarsa, P. Martinez and J. Vancostenoble, *Carleman estimates for a class of degenerate parabolic operators*, SIAM J. Control Optim. **47** (2008), 1–19.
- [8] J.M. Coron, *Control and nonlinearity*, Mathematical Surveys and Monographs, vol. 136, American Mathematical Society, Providence, RI, 2007.
- [9] A. Doubova, E. Fernández-Cara, M. Gouzález-Burgos and E. Zuazua, *On the controllability of parabolic systems with a nonlinear term involving the state and the gradient*, SIAM J. Control Optim., **41** (2002), 798–819.
- [10] T. Duyckaerts, X. Zhang and E. Zuazua, *On the optimality of the observability inequalities for parabolic and hyperbolic systems with potentials*, Ann. Inst. H. Poincaré Anal. Non Linéaire, **25** (2008), 1–41.
- [11] E. Fernández-Cara, S. Guerrero, O.Yu. Imanuvilov and J.-P. Puel, *Local exact controllability of the Navier-Stokes system*, J. Math. Pures Appl., **83** (2004), 1501–1542.
- [12] E. Fernández-Cara and E. Zuazua, *Null and approximate controllability for weakly blowing-up semilinear heat equations*, Ann. Inst. H. Poincaré Anal. Non Linéaire, **17** (2000), 583–616.
- [13] X. Fu, *A weighted identity for partial differential operators of second order and its applications*, C. R. Math. Acad. Sci. Paris, **342** (2006), 579–584.
- [14] X. Fu, *Logarithmic decay of hyperbolic equations with arbitrary small boundary damping*, Comm. Partial Differential Equations, **34** (2009), 957–975.
- [15] X. Fu, X. Liu and X. Zhang, *Well-posedness and local controllability for quasi-linear complex Ginzburg-Landau equations*, Preprint.
- [16] X. Fu, J. Yong and X. Zhang, *Exact controllability for multidimensional semilinear hyperbolic equations*, SIAM J. Control Optim., **46** (2007), 1578–1614.
- [17] A.V. Fursikov and O.Yu. Imanuvilov, *Controllability of evolution equations*, Lecture Notes Series, vol. 34, Research Institute of Mathematics, Seoul National University, Seoul, Korea, 1996.
- [18] V. Isakov, *Inverse problems for partial differential equations. Second edition*, Applied Mathematical Sciences, vol. 127, Springer-Verlag, New York, 2006.
- [19] R.E. Kalman, *On the general theory of control systems*, Proc. 1st IFAC Congress, Moscow, 1960, vol. 1, Butterworth, London, 1961, 481–492.
- [20] M. Kazemi and M.V. Klibanov, *Stability estimates for ill-posed Cauchy problem involving hyperbolic equations and inequalities*, Appl. Anal., **50** (1993), 93–102.
- [21] V. Komornik and P. Loretí, *Fourier series in control theory*, Springer Monographs in Mathematics, Springer-Verlag, New York, 2005.
- [22] I. Lasiecka, R. Triggiani and X. Zhang, *Nonconservative wave equations with purely Neumann B.C.: Global uniqueness and observability in one shot*, Contemp. Math., **268** (2000), 227–326.
- [23] I. Lasiecka, R. Triggiani and X. Zhang, *Global uniqueness, observability and stabilization of nonconservative Schrödinger equations via pointwise Carleman estimates: Part I. H^1 -estimates*, J. Inv. Ill-posed Problems, **11** (2004), 43–123.
- [24] M. M. Lavrentév, V. G. Romanov and S. P. Shishat-skii, *Ill-posed problems of mathematical physics and analysis*, Translations of Mathematical Monographs, vol. 64, American Mathematical Society, Providence, RI, 1986.

- [25] G. Lebeau and L. Robbiano, *Contrôle exact de l'équation de la chaleur*, Comm. Partial Differential Equations, **20** (1995), 335–336.
- [26] T.T. Li, *Controllability and observability for quasilinear hyperbolic systems*, AIMS Series on Applied Mathematics, vol. 3, American Institute of Mathematical Sciences (AIMS), Springfield, MO, 2010.
- [27] W. Li and X. Zhang, *Controllability of parabolic and hyperbolic equations: Towards a unified theory*, Control Theory of Partial Differential Equations, Lect. Notes Pure Appl. Math., vol. 242, Chapman & Hall/CRC, Boca Raton, FL, 2005, 157–174.
- [28] X. Li and J. Yong, *Optimal control theory for infinite-dimensional systems*, Systems & Control: Foundations & Applications, Birkhäuser Boston, Inc., Boston, MA, 1995.
- [29] J.-L. Lions, *Contrôlabilité exacte, perturbations et stabilisation de systèmes distribués, Tome 1*, Recherches en Mathématiques Appliquées, vol. 8, Masson, Paris, 1988.
- [30] X. Liu and H. Gao, *Controllability of a class of Newtonian filtration equations with control and state constraints*, SIAM J. Control Optim., **46** (2007), 2256–2279.
- [31] X. Liu and X. Zhang, *On the local controllability of a class of multidimensional quasilinear parabolic equations*, C. R. Math. Acad. Sci. Paris, **347** (2009), 1379–1384.
- [32] A. López, X. Zhang and E. Zuazua, *Null controllability of the heat equation as singular limit of the exact controllability of dissipative wave equations*, J. Math. Pures Appl., **79** (2000), 741–808.
- [33] Q. Lü, *Control and observation of stochastic PDEs*, PhD Thesis, Sichuan University, 2010.
- [34] L. Miller, *The control transmutation method and the cost of fast controls*, SIAM J. Control Optim., **45** (2006), 762–772.
- [35] S. Peng, *Backward stochastic differential equation and exact controllability of stochastic control systems*, Progr. Natur. Sci. (English Ed.), **4** (1994), 274–284.
- [36] K.D. Phung, *Observability and control of Schrödinger equations*, SIAM J. Control Optim., **40** (2001), 211–230.
- [37] K.D. Phung, G. Wang and X. Zhang, *On the existence of time optimal control of some linear evolution equations*, Discrete Contin. Dyn. Syst. B, **8** (2007), 925–941.
- [38] L. Rosier and B.Y. Zhang, *Null Controllability of the complex Ginzburg-Landau equation*, Ann. Inst. H. Poincaré Anal. Non Linéaire, **26** (2009), 649–673.
- [39] D.L. Russell, *A unified boundary controllability theory for hyperbolic and parabolic partial differential equations*, Stud. Appl. Math., **52** (1973), 189–221.
- [40] D.L. Russell, *Controllability and stabilizability theory for linear partial differential equations: recent progress and open problems*, SIAM Rev., **20** (1978), 639–739.
- [41] S. Tang and X. Zhang, *Null controllability for forward and backward stochastic parabolic equations*, SIAM J. Control Optim., **48** (2009), 2191–2216.

-
- [42] L. Tébou, *A Carleman estimates based approach for the stabilization of some locally damped semilinear hyperbolic equations*, ESAIM Control Optim. Calc. Var., **14** (2008), 561–574.
- [43] M. Tucsnak and G. Weiss, *Observation and control for operator semigroups*, Birkhäuser Advanced Texts: Basler Lehrbücher, Birkhäuser Verlag, Basel, 2009.
- [44] G. Wang, *L^∞ -null controllability for the heat equation and its consequences for the time optimal control problem*, SIAM J. Control Optim., **48** (2008), 1701–1720.
- [45] G. Wang and L. Wang, *The Carleman inequality and its application to periodic optimal control governed by semilinear parabolic differential equations*, J. Optim. Theory Appl., **118** (2003), 249–461.
- [46] M. Yamamoto, *Carleman estimates for parabolic equations and applications*, Inverse Problems, **25** (2009), 123013.
- [47] X. Zhang, *Explicit observability estimate for the wave equation with potential and its application*, R. Soc. Lond. Proc. Ser. A Math. Phys. Eng. Sci., **456** (2000), 1101–1115.
- [48] X. Zhang, *Exact controllability of the semilinear plate equations*, Asymptot. Anal., **27** (2001), 95–125.
- [49] X. Zhang, *A remark on null exact controllability of the heat equation*, SIAM J. Control Optim., **40** (2001), 39–53.
- [50] X. Zhang, *Unique continuation for stochastic parabolic equations*, Differential Integral Equations, **21** (2008), 81–93.
- [51] X. Zhang, *Carleman and observability estimates for stochastic wave equations*, SIAM J. Math. Anal., **40** (2008), 851–868.
- [52] X. Zhang, *Remarks on the controllability of some quasilinear equations*, Preprint (see <http://arxiv.org/abs/0904.2427v1>).
- [53] X. Zhang and E. Zuazua, *Polynomial decay and control of a 1 – d hyperbolic-parabolic coupled system*, J. Differential Equations, **204** (2004), 380–438.
- [54] Y. Zhou and Z. Lei, *Local exact boundary controllability for nonlinear wave equations*, SIAM J. Control Optim., **46** (2007), 1022–1051.
- [55] E. Zuazua, *Propagation, observation, and control of waves approximated by finite difference methods*, SIAM Rev., **47** (2005), 197–243.
- [56] E. Zuazua, *Controllability and observability of partial differential equations: some results and open problems*, Handbook of Differential Equations: Evolutionary Differential Equations, vol. 3, Elsevier Science, 2006, 527–621.

Section 18

**Mathematics in Science
and Technology**

This page is intentionally left blank

Deterministic and Stochastic Aspects of Single-crossover Recombination

Ellen Baake*

Abstract

This contribution is concerned with mathematical models for the dynamics of the genetic composition of populations evolving under recombination. Recombination is the genetic mechanism by which two parent individuals create the mixed type of their offspring during sexual reproduction. The corresponding models are large, nonlinear dynamical systems (for the deterministic treatment that applies in the infinite-population limit), or interacting particle systems (for the stochastic treatment required for finite populations). We review recent progress on these difficult problems. In particular, we present a closed solution of the deterministic continuous-time system, for the important special case of single crossovers; we extract an underlying linearity; we analyse how this carries over to the corresponding stochastic setting; and we provide a solution of the analogous deterministic discrete-time dynamics, in terms of its generalised eigenvalues and a simple recursion for the corresponding coefficients.

Mathematics Subject Classification (2010). Primary 92D10, 34L30; Secondary 37N25, 06A07, 60J25.

Keywords. Population genetics, recombination dynamics, Möbius linearisation and diagonalisation, correlation functions, Moran model.

1. Introduction

Biological evolution is a complex phenomenon driven by various processes, such as mutation and recombination of genetic material, reproduction of individuals, and selection of favourable types. The area of *population genetics* is concerned with how these processes shape and change the genetic structure of populations. *Mathematical population genetics* was founded in the 1920's by Ronald Fisher, Sewall Wright, and John Haldane, and thus is among the oldest areas

*Faculty of Technology, Bielefeld University, 33594 Bielefeld, Germany.
E-mail: ebaake@techfak.uni-bielefeld.de.

of mathematical biology. The reason for its continuing (and actually increasing) attractiveness for both mathematicians and biologists is at least twofold: Firstly, there is a true need for mathematical models and methods, since the outcome of evolution is impossible to predict (and, thus, today's genetic data are impossible to analyse) without their help. Second, the processes of genetics lend themselves most naturally to a mathematical formulation and give rise to a wealth of fascinating new problems, concepts, and methods.

This contribution will focus on the phenomenon of *recombination*, in which two parent individuals are involved in creating the mixed type of their offspring during sexual reproduction. The essence of this process is illustrated in Figure 1 and may be idealised and summarised as follows.

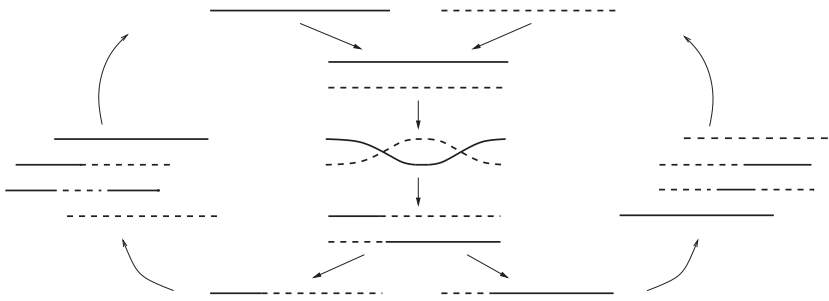


Figure 1. Life cycle of a population under sexual reproduction and recombination. Each line symbolises a sequence of sites that defines a gamete (like the two at the top that start the cycle as ‘egg’ and ‘sperm’). The pool of gametes at the left and the right comes from a large population of recombining individuals. These sequences meet randomly to start the next round of the cycle.

Genetic information is encoded in terms of sequences of finite length. Eggs and sperm (i.e., female and male germ cells or *gametes*) each carry a single copy of such a sequence. They go through the following life cycle: At fertilisation, two gametes meet randomly and unite, thus starting the life of a new individual, which is equipped with both the maternal and the paternal sequence. At maturity, this individual will generate its own germ cells. This process includes recombination, that is, the maternal and paternal sequences perform one or more *crossovers* and are cut and relinked accordingly, so that two ‘mixed’ sequences emerge. These are the new gametes and start the next round of fertilisation (by random mating within a large population).

Models of this process aim at describing the dynamics of the genetic composition of a population that goes through this life cycle repeatedly. These models come in various flavours: in discrete or continuous time; with various assumptions about the crossover pattern; and, most importantly, in a deterministic or a stochastic formulation, depending on whether the population is assumed to be so large that stochastic fluctuations may be neglected. In any case, however, the resulting process appears difficult to treat, due to the large number

of possible states and the nonlinearity generated by the random mixture of gametes. Nevertheless, a number of solution procedures have been discovered for the deterministic discrete-time setting [8, 10, 14], and the underlying mathematical structures were investigated within the framework of genetic algebras, see [18, 19, 22]. Quite generally, the solution relies on a certain nonlinear transformation (known as Haldane linearisation) from (gamete or type) frequencies to suitable correlation functions, which decouple from each other and decay geometrically. But if sequences of more than three sites are involved, this transformation must be constructed via recursions that involve the parameters of the recombination process, and is not available explicitly except in the trivial case of independent sites. For a review of the area, see [9, Ch. V.4].

In this contribution, we concentrate on a special case that is both biologically and mathematically relevant, namely, the situation in which at most one crossover happens at any given time. That is, only recombination events may occur that partition the sites of a sequence into two parts that correspond to the sites before and after a given crossover point. We analyse the resulting models in continuous time (both deterministic and stochastic), as well as in discrete time. For the deterministic continuous-time system (Section 2), a simple explicit solution can be given. This simplicity is due to some underlying linearity; actually, the system may even be diagonalised (via a *nonlinear* transformation). In Section 3, we consider the corresponding stochastic process (still in continuous time), namely, the Moran model with recombination. This also takes into account the resampling effect that comes about via random reproduction in a finite population. In particular, we investigate the relationship between the expectation of the Moran model and the solution of the deterministic continuous-time model. We finally tackle deterministic single-crossover dynamics in *discrete* time (Section 4). This setting implies additional dependencies, which become particularly transparent when the so-called *ancestral recombination process* is considered. A solution may still be given, but its coefficients must be determined recursively.

Altogether, it will turn out that the corresponding models, and their analysis, have various mathematical facets that are intertwined with each other, such as differential equations, probability theory, and combinatorics.

2. Deterministic Dynamics, Continuous Time

2.1. The model. We describe populations at the level of their gametes and thus identify gametes with individuals. Their genetic information is encoded in terms of a linear arrangement of sites, indexed by the set $S := \{0, 1, \dots, n\}$. For each site $i \in S$, there is a set X_i of ‘letters’ that may possibly occur at that site. To allow for a convenient notation, we restrict ourselves to the simple but important case of *finite* sets X_i ; for the full generality of arbitrary locally compact spaces X_i , the reader is referred to [3] and [5].

A *type* is thus defined as a sequence

$$x = (x_0, x_1, \dots, x_n) \in X_0 \times X_1 \times \dots \times X_n =: X,$$

where X is called the *type space*. By construction, x_i is the i -th coordinate of x , and we define $x_I := (x_i)_{i \in I}$ as the collection of coordinates with indices in I , where I is a subset of S . A *population* is identified with a non-negative measure ω on X . Namely, $\omega(\{x\})$ denotes the frequency of individuals of type $x \in X$ and $\omega(A) := \sum_{x \in A} \omega(\{x\})$ for $A \subseteq X$; we abbreviate $\omega(\{x\})$ as $\omega(x)$. The set of all nonnegative measures on X is denoted by $\mathcal{M}_{\geq 0}(X)$. If we define δ_x as the point measure on x (i.e., $\delta_x(y) = \delta_{x,y}$ for $x, y \in X$), we can also write $\omega = \sum_{x \in X} \omega(x)\delta_x$. We may, alternatively, interpret δ_x as the basis vector of $\mathbb{R}_{\geq 0}^{|X|}$ that corresponds to x (where a suitable ordering of types is implied, and $|X|$ is the number of elements in X); ω is thus identified with a vector in $\mathbb{R}_{\geq 0}^{|X|}$.

At this stage, frequencies need not be normalised; $\omega(x)$ may simply be thought of as the size of the subpopulation of type x , measured in units so large that it may be considered a continuous quantity. The corresponding normalised version $p := \omega/\|\omega\|$ (where $\|\omega\| := \sum_{x \in X} \omega(x) = \omega(X)$ is the total population size) is then a probability distribution on X , and may be identified with a probability vector.

Recombination acts on the links between the sites; the links are collected into the set $L := \{\frac{1}{2}, \frac{3}{2}, \dots, \frac{2n-1}{2}\}$. We shall use Latin indices for the sites and Greek indices for the links, and the implicit rule will always be that $\alpha = \frac{2i+1}{2}$ is the link between sites i and $i + 1$; see Figure 2.

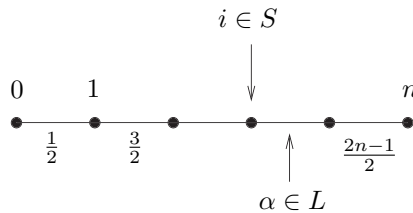


Figure 2. Sites and links.

Let recombination happen in every individual, and at every link $\alpha \in L$, at rate $\varrho_\alpha > 0$. More precisely, for every $\alpha \in L$, every individual exchanges, at rate $\varrho_\alpha/2$, the sites after link α with those of a randomly chosen partner. Explicitly, if the ‘active’ and the partner individual are of types x and y , then the new pair has types $(x_0, x_1, \dots, x_{\lfloor \alpha \rfloor}, y_{\lceil \alpha \rceil}, \dots, y_n)$ and $(y_0, y_1, \dots, y_{\lfloor \alpha \rfloor}, x_{\lceil \alpha \rceil}, \dots, x_n)$, where $\lfloor \alpha \rfloor$ ($\lceil \alpha \rceil$) is the largest integer below α (the smallest above α); see Figure 3. Since every individual can occur as either the ‘active’ individual or as its randomly chosen partner, we have a total rate of ϱ_α for crossovers at link α . For later use, we also define $\varrho := \sum_{\alpha \in L} \varrho_\alpha$.

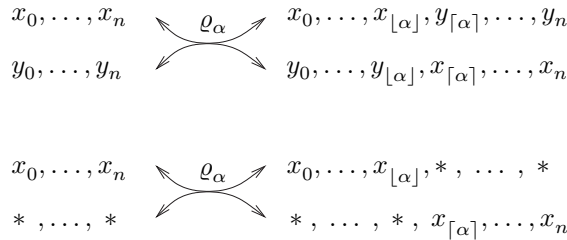


Figure 3. Upper panel: Recombination between individuals of type x and y . Lower panel: The corresponding ‘marginalised’ version that summarises all events by which individuals of type x are gained or lost (a ‘*’ at site i stands for an arbitrary element of X_i). Note that, in either case, the process can go both ways, as indicated by the arrows.

In order to formulate the corresponding model, let us introduce the projection operators $\pi_i, i \in S$, via

$$\pi_i : X_0 \times X_1 \times \dots \times X_n \longrightarrow X_i \tag{1}$$

$$(x_0, x_1, \dots, x_n) \mapsto x_i,$$

i.e., π_i is the canonical projection to the i -th coordinate. Likewise, for any index set $I \subseteq S$, one defines a projector

$$\pi_I : X \longrightarrow \prod_{i \in I} X_i =: X_I$$

$$(x_0, x_1, \dots, x_n) \mapsto (x_i)_{i \in I} =: x_I.$$

We shall frequently use the abbreviations $\pi_{<\alpha} := \pi_{\{1, \dots, [\alpha]\}}$ and $\pi_{>\alpha} := \pi_{\{[\alpha], \dots, n\}}$, as well as $x_{<\alpha} := \pi_{<\alpha}(x), x_{>\alpha} := \pi_{>\alpha}(x)$. The projectors $\pi_{<\alpha}$ and $\pi_{>\alpha}$ may be thought of as *cut and forget* operators because they take the leading or trailing segment of a sequence x , and forget about the rest.

Whereas the π_I act on the types, we also need the induced mapping at the level of the population, namely,

$$\pi_I . : \mathcal{M}_{\geq 0} \longrightarrow \mathcal{M}_{\geq 0} \tag{2}$$

$$\omega \mapsto \omega \circ \pi_I^{-1} =: \pi_I . \omega,$$

where π_I^{-1} denotes the preimage under π_I . The operation $.$ (where the dot is on the line) is the ‘pullback’ of π_I w.r.t. ω ; so, $\pi_I . \omega$ is the marginal distribution of ω with respect to the sites in I . In particular, $(\pi_{<\alpha} . \omega)(x_{<\alpha})$ is the marginal frequency of sequences prescribed at the sites before α , and vice versa for the sites after α .

Now, single-crossover recombination (at the level of the population) means the relinking of a randomly chosen leading segment with a randomly chosen trailing segment. We therefore introduce (elementary) recombination operators (or *recombinators*, for short), $R_\alpha : \mathcal{M}_{\geq 0} \rightarrow \mathcal{M}_{\geq 0}$ for $\alpha \in L$, defined by

$$R_\alpha(\omega) := \frac{1}{\|\omega\|} ((\pi_{<\alpha} . \omega) \otimes (\pi_{>\alpha} . \omega)). \tag{3}$$

Here, the tensor product reflects the independent combination (i.e., the product measure) of the two marginals $\pi_{<\alpha}.\omega$ and $\pi_{>\alpha}.\omega$. R_α is therefore a *cut and relink* operator. $R_\alpha(\omega)$ may be understood as the population that emerges if *all* individuals of the population ω disintegrate into their leading and trailing segments and these are relinked randomly. Note that $\|R_\alpha(\omega)\| = \|\omega\|$.

The recombination dynamics may thus be compactly written as

$$\dot{\omega}_t = \sum_{\alpha \in L} \varrho_\alpha (R_\alpha(\omega_t) - \omega_t) = \sum_{\alpha \in L} \varrho_\alpha (R_\alpha - \mathbb{1})(\omega_t) =: \Phi(\omega_t), \tag{4}$$

where $\mathbb{1}$ is the identity operator. Note that (4) is a large, nonlinear system of ordinary differential equations (ODEs).

2.2. Solution of the ODE system. The solution of (4) relies on some elementary properties of our recombinators. Most importantly, they are idempotents and commute with each other, i.e.,

$$R_\alpha^2 = R_\alpha, \quad \alpha \in L, \tag{5}$$

$$R_\alpha R_\beta = R_\beta R_\alpha, \quad \alpha, \beta \in L. \tag{6}$$

These properties are intuitively plausible: if the links before α are already independent of those after α due to a previous recombination event, then further recombination at that link does not change the situation; and if a product measure is formed with respect to two links α and β , the result does not depend on the order in which the links are affected. For the proof, we refer to [5, Prop. 2]; let us only mention here that it relies on the elementary fact that, for $\omega \in \mathcal{M}_{\geq 0}$,

$$\begin{aligned} \pi_{<\alpha}.(R_\beta(\omega)) &= \pi_{<\alpha}.\omega, & \text{for } \beta \geq \alpha, \text{ and} \\ \pi_{>\alpha}.(R_\beta(\omega)) &= \pi_{>\alpha}.\omega, & \text{for } \beta \leq \alpha; \end{aligned}$$

that is, recombination at or after α does not affect the marginal frequencies at sites before α , and vice versa.

We now define *composite* recombinators as

$$R_G := \prod_{\alpha \in G} R_\alpha \quad \text{for } G \subseteq L.$$

Here, the product is to be read as composition; it is, indeed, a product if the recombinators are written in terms of their multilinear matrix representations, which is available in the case of finite types considered here (see [2]). By property (6), the order in the composition plays no role. Furthermore, (5) and (6) obviously entail $R_G R_H = R_{G \cup H}$ for $G, H \subseteq L$.

With this in hand, we can now state an explicit solution of our problem, namely,

Theorem 1. *The solution of the single-crossover dynamics (4) with initial value ω_0 can be given in closed form as*

$$\omega_t = \sum_{G \subseteq L} a_G(t) R_G(\omega_0) =: \varphi_t(\omega_0) \tag{7}$$

with coefficient functions

$$a_G(t) = \prod_{\alpha \in L \setminus G} e^{-\varrho_\alpha t} \prod_{\beta \in G} (1 - e^{-\varrho_\beta t});$$

i.e., φ_t is the semigroup belonging to the recombination equation (4). □

For the proof, the reader is referred to [5, Thm. 2] or [3, Thm. 3] (the former article contains the original, the latter a shorter and more elegant version of the proof). Let us note that the coefficient functions can be interpreted probabilistically. Given an individual sequence in the population, $a_G(t)$ is the probability that the set of links that have seen at least one crossover event until time t is precisely the set G (obviously, $\sum_{G \subseteq L} a_G(t) = 1$). Note that the product structure of the $a_G(t)$ implies independence of links, a decisive feature of the single-crossover dynamics in continuous time, as we shall see later on. Note also that, as $t \rightarrow \infty$, ω_t converges to the stationary state

$$\omega_\infty = \frac{1}{\|\omega_0\|^{n-1}} \bigotimes_{i=1}^n (\pi_i \cdot \omega_0), \tag{8}$$

in which all sites are independent.

2.3. Underlying linearity. The simplicity of the solution in Theorem 1 comes as a certain surprise. After all, explicit solutions to large, nonlinear ODE systems are rare – they are usually available for linear systems at best. For this reason, the recombination equation and its solution have already been taken up in the framework of functional analysis, where they have led to an extension of potential theory [21]. We will now show that there is an underlying linear structure that is hidden behind the solution. It can be stated as follows, compare [5, Sec. 3.2] for details.

Theorem 2. *Let $\{c_{G'}^{(L')}(t) \mid \emptyset \subseteq G' \subseteq L' \subseteq L\}$ be a family of non-negative functions with $c_G^{(L)}(t) = c_{G_1}^{(L_1)}(t) c_{G_2}^{(L_2)}(t)$, valid for any partition $L = L_1 \dot{\cup} L_2$ of the set L and all $t \geq 0$, where $G_i := G \cap L_i$. Assume further that these functions satisfy $\sum_{H \subseteq L'} c_H^{(L')}(t) = 1$ for any $L' \subseteq L$ and $t \geq 0$. If $v \in \mathcal{M}_{\geq 0}(X)$*

and $H \subseteq L$, one has the identity

$$R_H \left(\sum_{G \subseteq L} c_G^{(L)}(t) R_G(v) \right) = \sum_{G \subseteq L} c_G^{(L)}(t) R_{G \cup H}(v),$$

which is then satisfied for all $t \geq 0$. □

Here, the upper index specifies the respective set of links. Clearly, the coefficient functions $a_G(t)$ of Theorem 1 satisfy the conditions of Theorem 2. The result then means that the recombinators act linearly along the solutions (7) of the recombination equation (4). Theorem 2 thus has the consequence that, on $\mathcal{M}_{\geq 0}(X)$, the forward flow of (4) commutes with all recombinators, that is, $R_G \circ \varphi_t = \varphi_t \circ R_G$ for all $t \geq 0$ and all $G \subseteq L$.

But let us go one step further here. The conventional approach to solve the recombination dynamics consists in transforming the type frequencies to certain functions (known as principal components) that diagonalise the dynamics, see [8, 10, 18] and references therein for more. We will now show that, in continuous time, they have a particularly simple structure: they are given by certain correlation functions, known as *linkage disequilibria* (LDE) in biology, which play an important role in applications. They have a counterpart at the level of operators (on $\mathcal{M}_{\geq 0}(X)$). Namely, let us define *LDE operators* via

$$T_G := \sum_{H \supseteq G} (-1)^{|H \setminus G|} R_H, \quad G \subseteq L, \tag{9}$$

where the underdot indicates the summation variable. Note that T_G maps $\mathcal{M}_{\geq 0}(X)$ into $\mathcal{M}(X)$, the set of *signed* measures on X . Eq. (9) leads to the inverse $R_G = \sum_{H \supseteq G} T_H$ by the combinatorial Möbius inversion formula, see [1, Thm. 4.18]. We then have

Theorem 3. *If ω_t is the solution (7), the transformed quantities $T_G(\omega_t)$ satisfy*

$$\frac{d}{dt} T_G(\omega_t) = - \left(\sum_{\alpha \in L \setminus G} \varrho_\alpha \right) T_G(\omega_t), \quad G \subseteq L. \tag{10}$$

Proof. See [5, Sec. 3.3]. □

Obviously, Eq. (10) is a system of *decoupled, linear, homogeneous* differential equations with the usual exponential solution. Note that this simple form emerged through the *nonlinear* transform (9) as applied to the solution of the *coupled, nonlinear* differential equation (4).

Suitable components of the signed measure $T_G(\omega_t)$ may then be identified to work with in practice (see [5, 6] for details); they correspond to correlation functions of all orders and decouple and decay exponentially. These functions turn out to be particularly well-adapted to the problem since they rely on ordered partitions, in contrast to conventional LDE’s used elsewhere in population genetics, which rely on general partitions (see [9, Ch. V.4] for review).

3. Stochastic Dynamics, Continuous Time

3.1. The model. The effect of finite population size in population genetics is, in continuous time, well captured by the *Moran model*. It describes a population of fixed size N and takes into account the stochastic fluctuations due to random reproduction, which manifest themselves via a *resampling effect* (known as genetic drift in biology). More precisely, the finite-population counterpart of our deterministic model is the Moran model with single-crossover recombination. To simplify matters (and in order to clearly dissect the individual effects of recombination and resampling), we shall use the decoupled (or parallel) version of the model, which assumes that resampling and recombination occur independently of each other, as illustrated in Figure 4. More precisely, in our finite population of fixed size N , every individual experiences, independently of the others,

- resampling at rate $b/2$. The individual reproduces, the offspring inherits the parent's type and replaces a randomly chosen individual (possibly its own parent).
- recombination at (overall) rate ρ_α at link $\alpha \in L$. Every individual picks a random partner (maybe itself) at rate $\rho_\alpha/2$, and the pair exchanges the sites after link α . That is, if the recombining individuals have types x and y , they are replaced by the two offspring individuals $(x_{<\alpha}, y_{>\alpha})$ and $(y_{<\alpha}, x_{>\alpha})$, as in the deterministic case, and Figure 3. As before, the per-capita rate of recombination at link α is then ρ_α , because both orderings of the individuals lead to the same type count in the population.

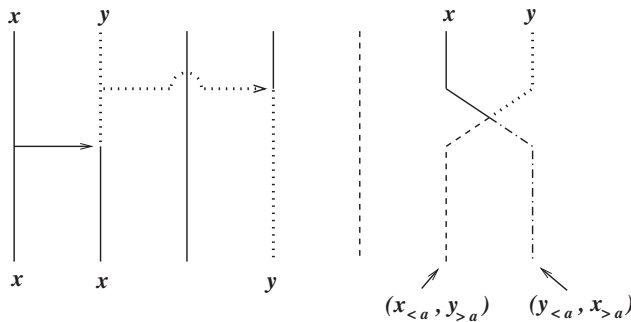


Figure 4. Graphical representation of the Moran model (with parallel resampling and recombination). Every individual is represented by a vertical line; time runs down the page. Resampling is indicated by arrows, with the parent individual at the tail and the offspring at the tip. Recombination is depicted by a crossing between two individuals. Note that the spatial information suggested by the graphical representation does not play a role in the model; one is only interested in the frequencies of the various types.

Note that the randomly chosen second individual (for resampling or recombination) may be the active individual itself; then, effectively, nothing happens. One might, for biological reasons, prefer to exclude these events by sampling from the remaining population only; but this means nothing but a change of time scale of order $1/N$.

To formalise this verbal description of the process, let the state of the population at time t be given by the collection (the random vector)

$$Z_t = (Z_t(x))_{x \in X} \in E := \left\{ z \in \{0, 1, \dots, N\}^{|X|} \mid \sum_x z(x) = N \right\},$$

where $Z_t(x)$ is the number of individuals of type x at time t ; clearly, $\sum_{x \in X} Z_t(x) = N$. We also use Z_t in the sense of a (random counting) measure, in analogy with ω_t (but keep in mind that Z_t is integer-valued and counts single individuals, whereas ω_t denotes continuous frequencies in an infinite population). The letter z will be used to denote realisations of Z_t — but note that the symbols x, y , and z are not on equal footing (x and y will continue to be types). The stochastic process $\{Z_t\}_{t \geq 0}$ is the continuous-time Markov chain on E defined as follows. If the current state is $Z_t = z$, two types of transitions may occur:

$$\begin{aligned} \text{resampling: } \quad z &\rightarrow z + s(x, y), \quad s(x, y) := \delta_x - \delta_y, \\ &\text{at rate } \frac{1}{2N} b z(x) z(y) \text{ for } (x, y) \in X \times X \end{aligned} \tag{11}$$

$$\begin{aligned} \text{recombination: } \quad z &\rightarrow z + r(x, y, \alpha), \\ r(x, y, \alpha) &:= \delta_{(x_{<\alpha}, y_{>\alpha})} + \delta_{(y_{<\alpha}, x_{>\alpha})} - \delta_x - \delta_y, \\ &\text{at rate } \frac{1}{2N} \varrho_\alpha z(x) z(y) \text{ for } (x, y) \in X \times X, \alpha \in L \end{aligned} \tag{12}$$

(where δ_x is the point measure on x , as before). Note that, in (11) and (12), transitions that leave E are automatically excluded by the fact that the corresponding rates vanish. On the other hand, ‘empty transitions’ ($s(x, y) = 0$ or $r(x, y, \alpha) = 0$) are explicitly included (they occur if $x = y$ in resampling or recombination, and if $x_{<\alpha} = y_{<\alpha}$ or $x_{>\alpha} = y_{>\alpha}$ in recombination).

3.2. Connecting stochastic and deterministic models. Let us now explore the connection between the stochastic process $\{Z_t\}_{t \geq 0}$ on E , its normalised version $\{\widehat{Z}_t\}_{t \geq 0} = \{Z_t\}_{t \geq 0}/N$ on E/N , and the solution $\omega_t = \varphi_t(\omega_0)$ (Eq. (7)) of the differential equation. It is easy to see (and no surprise) that

$$\frac{d}{dt} \mathbb{E}(Z_t) = \mathbb{E}(\Phi(Z_t)), \tag{13}$$

with Φ of (4). But this does not, *per se*, lead to a ‘closed’ differential equation for $\mathbb{E}(Z_t)$, because it is not clear whether $\mathbb{E}(\Phi(Z_t))$ can be written as a function

of $\mathbb{E}(Z_t)$ alone—after all, Φ is nonlinear. In the absence of resampling, however, we have

Theorem 4. *Let $\{Z_t\}_{t \geq 0}$ be the recombination process without resampling (i.e., $b = 0$), and let Z_0 be fixed. Then, $\mathbb{E}(Z_t)$ satisfies the differential equation*

$$\frac{d}{dt}\mathbb{E}(Z_t) = \Phi(\mathbb{E}(Z_t))$$

with initial value Z_0 , and Φ from (4); therefore,

$$\mathbb{E}(Z_t) = \varphi_t(Z_0), \quad \text{for all } t \geq 0,$$

with φ_t from (7). Likewise, for all $t \geq 0$,

$$\mathbb{E}(T_G Z_t) = T_G(\varphi_t(Z_0)).$$

Proof. See [6, Thm. 1 and Cor. 1]. □

The result again points to some underlying linearity, which, in the context of the *stochastic* model, should be connected to some kind of independence. Indeed, the key to the proof of Theorem 4 is a lemma concerning the independence of marginal processes. For $I \subseteq S$, we introduce the ‘stretch’ of I as

$$J(I) := \{i \in S \mid \min(I) \leq i \leq \max(I)\},$$

and look at the projection of the recombination process on non-overlapping stretches. This is the content of

Lemma 5. *Let $\{Z_t\}_{t \geq 0}$ be the recombination process without resampling (i.e., $b = 0$). Let $A, B \subseteq S$ with $J(A) \cap J(B) = \emptyset$. Then, $\{\pi_A \cdot Z_t\}_{t \geq 0}$ and $\{\pi_B \cdot Z_t\}_{t \geq 0}$ are conditionally (on Z_0) independent Markov chains on E_A and E_B .*

Proof. See [6, Lemma 1]. □

Let us now re-include resampling, at rate $b/2 > 0$, and consider the stochastic process $\{Z_t^{(N)}\}_{t \geq 0}$ defined by both (11) and (12), where we add the upper index here to indicate the dependence on N . Now, Lemma 5 and Theorem 4 are no longer valid. The processes $\{\pi_{<\alpha} \cdot Z_t^{(N)}\}_{t \geq 0}$ and $\{\pi_{>\alpha} \cdot Z_t^{(N)}\}_{t \geq 0}$ are still individually Markov, but their resampling events are coupled (replacement of $y_{<\alpha}$ by $x_{<\alpha}$ is always tied to replacement of $y_{>\alpha}$ by $x_{>\alpha}$). Hence the marginal processes fail to be independent, so that no equivalent of Lemma 5 holds.

Let us, therefore, change focus and consider the normalised version $\{\widehat{Z}_t^{(N)}\}_{t \geq 0} = \{Z_t^{(N)}\}_{t \geq 0}/N$. In line with general folklore in population genetics, in the limit $N \rightarrow \infty$, the relative frequencies $\{\widehat{Z}_t^{(N)}\}_{t \geq 0}$ cease to fluctuate

and are then given by the solution of the corresponding deterministic equation. More precisely, we have

Proposition 6. *Consider the family of processes $\{\widehat{Z}_t^{(N)}\}_{t \geq 0} = \frac{1}{N}\{Z_t^{(N)}\}_{t \geq 0}$, $N = 1, 2, \dots$, where $\{Z_t^{(N)}\}_{t \geq 0}$ is defined by (11) and (12). Assume that the initial states are chosen so that $\lim_{N \rightarrow \infty} \widehat{Z}_0^{(N)} = p_0$. Then, for every given $t \geq 0$, one has*

$$\lim_{N \rightarrow \infty} \sup_{s \leq t} |\widehat{Z}_s^{(N)} - p_s| = 0 \tag{14}$$

with probability 1, where $p_s := \varphi_s(p_0)$ is the solution of the deterministic recombination equation (4). □

The proof is an elementary application of Thm. 11.2.1 of [12]; see Prop. 1 of [6] for the explicit workout.

Note that the convergence in (14) applies for any given t , but need not carry over to $t \rightarrow \infty$. Indeed, if resampling is present, the population size required to get close to the deterministic solution is expected to grow over all bounds with increasing t . This is because, for every finite N , the Moran model with resampling and recombination is an *absorbing* Markov chain, which leads to fixation (i.e., to a homogeneous population of uniform type) in finite time with probability one (for the special case of just two types without recombination, the expected time is known to be of order N if the initial frequencies are both $1/2$ [13, p. 93]). In sharp contrast, the deterministic system never loses any type, and the stationary state, the complete product measure (8), is, in a sense, even the most variable state accessible to the system. For increasing N , finite populations stay close to the deterministic limit for an increasing length of time.

4. Discrete Time

Let us return to the deterministic setting and consider the *discrete-time* version of our single-crossover dynamics (4), that is,

$$\omega_{t+1} = \omega_t + \sum_{\alpha \in L} \tilde{\varrho}_\alpha (R_\alpha - \mathbb{1})(\omega_t) =: \tilde{\Phi}(\omega_t). \tag{15}$$

Here, the coefficients $\tilde{\varrho}_\alpha > 0$, $\alpha \in L$, are the *probabilities* for a crossover at link α in every generation (as opposed to the *rates* ϱ_α of the continuous-time setting). Consequently, we must have $0 < \sum_{\alpha \in L} \tilde{\varrho}_\alpha \leq 1$.

Based on the result for the continuous-time model, the solution is expected to be of the form

$$\omega_t = \tilde{\Phi}^t(\omega_0) = \sum_{G \subseteq L} \tilde{a}_G(t) R_G(\omega_0), \tag{16}$$

with non-negative $\tilde{a}_G(t)$, $G \subseteq L$, $\sum_{G \subseteq L} \tilde{a}_G(t) = 1$, describing the (still unknown) coefficient functions arising from the dynamics. This representation of

the solution was first stated by Geiringer [14]. The coefficient functions will have the same probabilistic interpretation as the corresponding $a_G(t)$ in the continuous-time model, so that $\tilde{a}_G(t)$ is the probability that the links that have been involved in recombination until time t are exactly those of the set G .

But there is a crucial difference. Recall that, in continuous time, single crossovers imply *independence* of links, which is expressed in the product structure of the $a_G(t)$ (see Theorem 1). This independence is lost in discrete time, where a crossover event at one link forbids any other cut at other links in the same time step. It is therefore not surprising that a closed solution is not available in this case. It will, however, turn out that a solution can be stated in terms of the (generalised) eigenvalues of the system (which are known explicitly), together with coefficients to be determined via a simple recursion. But it is rewarding to take a closer look at the dynamics first.

Let us introduce the following abbreviations:

$$L_{\leq \alpha} := \{i \in L \mid i \leq \alpha\}, \quad L_{\geq \alpha} := \{i \in L \mid i \geq \alpha\},$$

and, for each $G \subseteq L$,

$$G_{< \alpha} := \{i \in G \mid i < \alpha\}, \quad G_{> \alpha} := \{i \in G \mid i > \alpha\}.$$

Furthermore, we set $\eta := 1 - \sum_{\alpha \in L} \tilde{\varrho}_\alpha$. The dynamics (15) is then reflected in the following dynamics of the coefficient functions:

Theorem 7. *For all $G \subseteq L$ and $t \in \mathbb{N}_0$, the coefficient functions $\tilde{a}_G(t)$ evolve according to*

$$\tilde{a}_G(t+1) = \eta \tilde{a}_G(t) + \sum_{\alpha \in G} \tilde{\varrho}_\alpha \left(\sum_{H \subseteq L_{\geq \alpha}} \tilde{a}_{G_{< \alpha} \cup H}(t) \right) \left(\sum_{K \subseteq L_{\leq \alpha}} \tilde{a}_{K \cup G_{> \alpha}}(t) \right), \tag{17}$$

with initial condition $\tilde{a}_G(0) = \delta_{G, \emptyset}$. □

A verbal description of this dynamics was already given by Geiringer [14]; a formal proof may be found in [24, Thm. 3].

The above iteration is easily understood intuitively: A type x resulting from recombination at link α is composed of two segments $x_{< \alpha}$ and $x_{> \alpha}$. These segments themselves may have been pieced together in previous recombination events already, and the iteration explains the possible cuts these segments may carry along. The first term in the product stands for the type delivering the leading segment (which may bring along arbitrary cuts in the trailing segment), the second for the type delivering the trailing one (here any leading segment is allowed). The term $\eta \tilde{a}_G(t)$ covers the case of no recombination.

Let us now have a closer look at the structure of the dependence between links in discrete time. To this end, note first that the set $G = \{\alpha_1, \dots, \alpha_{|G|}\} \subseteq L$

with $\alpha_1 < \alpha_2 < \dots < \alpha_{|G|}$ partitions $L \setminus G$ into $\mathcal{L}_G := \{I_0^G, I_1^G, \dots, I_{|G|}^G\}$, where

$$I_0^G = \{\alpha \in L : \frac{1}{2} \leq \alpha < \alpha_1\}, \quad I_{|G|}^G = \{\alpha \in L : \alpha_{|G|} < \alpha \leq \frac{2n-1}{2}\}, \quad (18)$$

and $I_\ell^G = \{\alpha \in L : \alpha_\ell < \alpha < \alpha_{\ell+1}\}$ for $1 \leq \ell \leq |G| - 1$.

Cutting all links in G decomposes the original system (of sites and links) into subsystems which are independent of each other from then on. In particular, the links in I_j become independent of those in I_k , for $k \neq j$. The probability that none of these subsystems experiences any further recombination is

$$\lambda_G = \prod_{i=0}^{|G|} \left(1 - \sum_{\alpha_i \in I_i^G} \tilde{\rho}_{\alpha_i}\right). \quad (19)$$

In particular, $\lambda_\emptyset = \eta = 1 - \sum_{\alpha \in L} \tilde{\rho}_\alpha \geq 0$. The λ_G are, at the same time, the generalised eigenvalues that appear when the system is diagonalised and have been previously identified by Bennett [8], Lyubich [18] and Dawson [10].

A most instructive way to detail the effect of dependence is the *ancestral recombination process*: start from an individual in the present population, let time run backwards and consider how this individual’s type has been pieced together from different fragments in the past. In the four-sites example of Figure 5, the probability that exactly 1/2 and 3/2 have been cut reads

$$\begin{aligned} \tilde{a}_{\{\frac{1}{2}, \frac{3}{2}\}}(t) &= \tilde{\rho}_{\frac{1}{2}} \tilde{\rho}_{\frac{3}{2}} \sum_{k=0}^{t-2} \lambda_\emptyset^k \sum_{i=0}^{t-2-k} \lambda_{\frac{1}{2}}^i \lambda_{\{\frac{1}{2}, \frac{3}{2}\}}^{t-2-k-i} \\ &+ \tilde{\rho}_{\frac{1}{2}} \tilde{\rho}_{\frac{3}{2}} (1 - \tilde{\rho}_{\frac{3}{2}}) \sum_{k=0}^{t-2} \lambda_\emptyset^k \sum_{i=0}^{t-2-k} \lambda_{\frac{3}{2}}^i \lambda_{\{\frac{1}{2}, \frac{3}{2}\}}^{t-2-k-i}. \end{aligned} \quad (20)$$

Here, the first (second) term corresponds to the possibility that link 1/2 (3/2) is the first to be cut. Obviously, the two possibilities are not symmetric: If 3/2 is the first to break, an additional factor of $(1 - \tilde{\rho}_{5/2})$ is required to guarantee that, at the time of the second recombination event (at 1/2), the trailing segment (sites 2 and 3) remains intact while the leading segment (sites 0 and 1) is cut.

Despite these complications, the discrete-time dynamics can again be solved, even directly at the level of the $\tilde{a}_G(t)$, albeit slightly less explicitly than in continuous time. Indeed, it may be shown (and will be detailed in a forthcoming paper) that the coefficient functions have the form

$$\tilde{a}_G^{(L)}(t) = \sum_{H \subseteq G} \gamma_G^{(L)}(H) (\lambda_H^{(L)})^t,$$

where the upper index has again been added to indicate the dependence on the system. The coefficients $\gamma_G^{(L)}(H)$ ($H \subseteq G$) are defined recursively as follows.

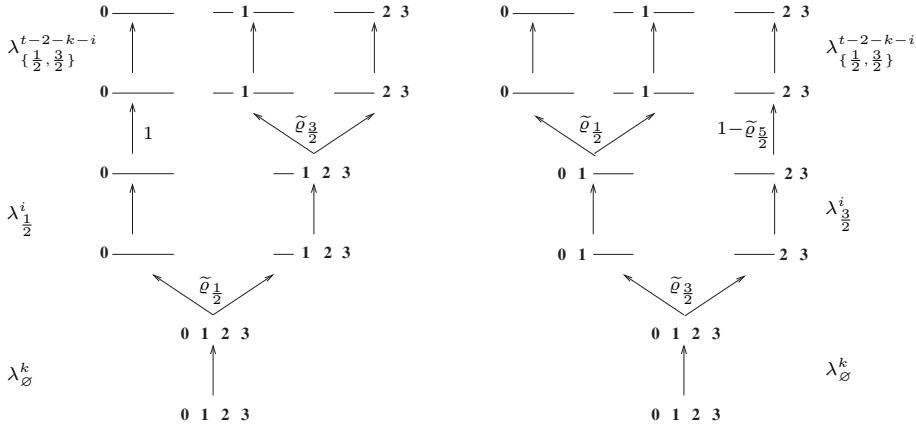


Figure 5. The ancestral recombination process: possible histories of the sequence 0123 (at the bottom). The two panels illustrate the two terms of $\tilde{a}_{\{1/2,3/2\}}(t)$ in Eq. (20) (left: link 1/2 is cut first; right: link 3/2 is cut first.) Arrows point in the backward direction of time. Blank lines indicate arbitrary leading or trailing segments with which parts of the sequence have joined during recombination (they correspond to the asterisks (*) in Figure 3). The probability that nothing happens for a while (straight arrows only) is given by (powers of) the generalised eigenvalues (19).

For $G \neq \emptyset$,

$$\gamma_G^{(L)}(H) = \begin{cases} \frac{1}{\lambda_H^{(L)} - \lambda_{\emptyset}^{(L)}} \sum_{\alpha \in H} \tilde{\varrho}_{\alpha} \gamma_{G_{<\alpha}}^{(L < \alpha)}(H_{<\alpha}) \gamma_{G_{>\alpha}}^{(L > \alpha)}(H_{>\alpha}), & H \neq \emptyset \\ - \sum_{\emptyset \neq J \subseteq G} \gamma_G^{(L)}(J), & H = \emptyset. \end{cases} \quad (21)$$

Together with the initial value $\gamma_{\emptyset}^{(L)}(\emptyset) = 1$, this may be solved recursively.

A diagonalisation of the system (analogous to that in Theorem 3) may also be achieved via a related, albeit technically more involved recursion [24].

5. Concluding Remarks and Outlook

The results presented here can naturally only represent a narrow segment from a large area with lively recent and current activities. Let us close this contribution by mentioning some important further directions in the context of recombination.

Our restriction to single crossovers provided a starting point with a uniquely transparent structure (mainly due to the independence of links in continuous time). However, arbitrary recombination patterns (which partition the set of links into two *arbitrary* parts) can also be dealt with, as has been done for the deterministic case in [18, 10]. The underlying mathematical structure will be

further investigated in a forthcoming paper, for both the deterministic and the stochastic models.

Above, genetic material was exchanged reciprocally at recombination events, so that the length of each sequence remains constant. But sequences may also shift relative to each other before cutting and relinking (so-called *unequal crossover*), which entails changes in length, see [4] and references therein for more.

The most important aspect of modern population genetics is the backward-in-time point of view. This is natural because evolution is mainly a historical science and today's researchers try to infer the past from samples of individuals taken from present-day populations. We have hinted at this with our version of an ancestral recombination process, but would like to emphasise that this is only a toy version. The full version of this process also takes into account resampling (as in Sec. 3, with $b > 0$) and aims at the law of *genealogies of samples* from *finite* populations. This point of view was introduced by Hudson [16]. The fundamental concept here is the *ancestral recombination graph*: a branching-coalescing graph, where branching (backwards in time) comes about by recombination (as in Figure 5), but lines may also coalesce where two individuals go back to a common ancestor (this corresponds to a resampling event forward in time). For recent introductions into this topic, see [11, Ch. 3], [15, Ch. 5], or [23, Ch. 7]; these texts also contain overviews of how recombination may be inferred from genomic datasets.

Last not least, recombination and resampling are but two of the various processes that act on genes in populations. Further inclusion of mutation and/or selection leads to a wealth of challenging problems, whose investigation has stimulated the exploration of new mathematical structures, concepts, and methods; let us only mention [7], [20], and [17] as recent examples. This development is expected to continue and intensify in the years to come – not least because it concerns the processes that have shaped present-day genomes.

References

- [1] M. Aigner, *Combinatorial Theory*, Springer, Berlin, 1979. Reprint 1997.
- [2] E. Baake, Mutation and recombination with tight linkage, *J. Math. Biol.* **42** (2001), 455–488.
- [3] M. Baake, Recombination semigroups on measure spaces, *Monatsh. Math.* **146** (2005), 267–278 and **150** (2007), 83–84 (Addendum). arXiv:math.CA/0506099.
- [4] M. Baake, Repeat distributions from unequal crossovers, *Banach Center Publ.* **80** (2008), 53–70. arXiv:0803.1270.
- [5] M. Baake and E. Baake, An exactly solved model for mutation, recombination and selection, *Canad. J. Math.* **55** (2003), 3–41 and **60** (2008), 264–265 (Erratum). arXiv:math.CA/0210422.
- [6] E. Baake and I. Herms, Single-crossover dynamics: Finite versus infinite populations, *Bull. Math. Biol.* **70** (2008), 603–624. arXiv:q-bio/0612024.

-
- [7] N. H. Barton, A. M. Etheridge, and A. K. Sturm, Coalescence in a random background, *Ann. Appl. Prob.* **14** (2004), 754–785. arXiv:math/0406174.
- [8] J. Bennett, On the theory of random mating, *Ann. Hum. Genet.* **18** (1954), 311–317.
- [9] R. Bürger, *The Mathematical Theory of Selection, Recombination, and Mutation*, Wiley, Chichester, 2000.
- [10] K. Dawson, The evolution of a population under recombination: How to linearize the dynamics, *Linear Algebra Appl.* **348** (2002), 115–137.
- [11] R. Durrett, *Probability Models for DNA Sequence Evolution*, 2nd ed., Springer, New York, 2008.
- [12] S. N. Ethier and T. G. Kurtz, *Markov Processes – Characterization and Convergence*, Wiley, New York, 1986. Reprint 2005.
- [13] W. Ewens, *Mathematical Population Genetics*, Springer, Berlin, 2nd ed., 2004.
- [14] H. Geiringer, On the probability theory of linkage in Mendelian heredity, *Ann. Math. Statist.* **15** (1944), 25–57.
- [15] J. Hein, M. H. Schierup, and C. Wiuf, *Gene Genealogies, Variation and Evolution*, Oxford University Press, Oxford, 2005.
- [16] R. R. Hudson, Properties of a neutral allele model with intragenic recombination, *Theor. Pop. Biol.* **23** (1983), 183–201.
- [17] P.A. Jenkins and Y.S. Song, An asymptotic sampling formula for the coalescent with recombination, *Ann. Appl. Prob.*, in press.
- [18] Y. Lyubich, *Mathematical Structures in Population Genetics*, Springer, New York, 1992.
- [19] D. McHale and G. Ringwood, Haldane linearization of baric algebras, *J. London Math. Soc.* **28** (1983), 17–26.
- [20] P. Pfaffelhuber, B. Haubold, and A. Wakolbinger, Approximate genealogies under genetic hitchhiking, *Genetics* **174** (2006), 1995–2008.
- [21] E. Popa, Some remarks on a nonlinear semigroup acting on positive measures, in: O. Carja and I. I. Vrabie, eds., *Applied Analysis and Differential Equations*, pp. 308–319, World Scientific, Singapore, 2007.
- [22] G. Ringwood, Hypergeometric algebras and Mendelian genetics, *Nieuw Archief v. Wiskunde* **3** (1985), 69–83.
- [23] J. Wakeley, *Coalescent Theory*. Roberts, Greenwood Village, CO, 2009.
- [24] U. von Wangenheim, E. Baake, and M. Baake, Single-crossover dynamics in discrete time, *J. Math. Biol.* **60** (2010), 727–760. arXiv:0906.1678.

BSDE and Risk Measures

Freddy Delbaen*

Abstract

The study of dynamic coherent risk measures and risk adjusted values is intimately related to the study of Backward Stochastic Differential Equations. We will present some of these relations and will also present some links with quasi-linear PDE.

Mathematics Subject Classification (2010). 91G80

Keywords. BSDE, Risk Measures, Time Consistency, Quasi-linear PDE

We will use the following notation:

$(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. We will work with bounded random variables and hence we use the space L^∞ . \mathbf{P} represents the set of all probability measures $\mathbb{Q} \ll \mathbb{P}$. \mathbf{P} is seen as a subset of L^1 . We will identify measures with their Radon-Nikodym derivatives and in the multivariate framework with their density or likelihood functions.

A utility function will be a function defined on L^∞ satisfying:

1. $u(0) = 0$, $\xi \geq 0$ implies $u(\xi) \geq 0$
2. u is concave
3. $u(\xi + a) = u(\xi) + a$ (monetary)
4. The set $\mathcal{A} = \{\xi \mid u(\xi) \geq 0\}$ is weak* closed: if $\xi_n \rightarrow \xi$ in probability and $\sup_n \|\xi_n\| < \infty$, then $u(\xi) \geq \limsup u(\xi_n)$.

Coherent means that $u(\lambda\xi) = \lambda u(\xi)$ for $\lambda \geq 0$.

The structure of such utility functions is well known.

Theorem 1. *There is a convex, lsc function $c : \mathbf{P} \rightarrow [0, +\infty]$ such that*

1. $\inf_{\mathbb{Q} \in \mathbf{P}} c(\mathbb{Q}) = 0$

*Eidgenössische Technische Hochschule, Department of Mathematics, 8092 Zürich, Switzerland. E-mail: delbaen@math.ethz.ch.

$$2. u(\xi) = \inf_{\mathbb{Q} \in \mathbb{P}} (\mathbb{E}_{\mathbb{Q}}[\xi] + c(\mathbb{Q}))$$

$$3. c(\mathbb{Q}) = \sup\{-\mathbb{E}_{\mathbb{Q}}[\xi] \mid \xi \in \mathcal{A}\}$$

If u is coherent c is the “indicator” of a convex closed set $\mathcal{S} \subset L^1$ and $u(\xi) = \inf_{\mathbb{Q} \in \mathcal{S}} \mathbb{E}_{\mathbb{Q}}[\xi]$.

We add the hypothesis $c(\mathbb{P}) = 0$ or for a coherent utility, $\mathbb{P} \in \mathcal{S}$.

In the multiperiod framework we get a family of utility functions, related to a filtration $(\mathcal{F}_t)_t$ and we use stopping times. Of course we ask that the filtration satisfies the usual assumptions. For the functions $u_\sigma : L^\infty(\mathcal{F}_T) \rightarrow L^\infty(\mathcal{F}_\sigma)$ we require:

$$1. u_\sigma(\xi) \in L^\infty(\Omega, \mathcal{F}_\sigma, \mathbb{P})$$

$$2. u_\sigma(0) = 0 \text{ and } u_\sigma(\xi) \geq 0 \text{ for } \xi \geq 0$$

$$3. 0 \leq \lambda \leq 1, \lambda \in L^\infty(\mathcal{F}_\sigma), \xi_1, \xi_2 \in L^\infty(\mathcal{F}_T) \text{ then}$$

$$u_\sigma(\lambda\xi_1 + (1-\lambda)\xi_2) \geq \lambda u_\sigma(\xi_1) + (1-\lambda)u_\sigma(\xi_2)$$

$$4. \text{ if } \eta \in L^\infty(\mathcal{F}_\sigma) \text{ then } u_\sigma(\xi + \eta) = u_\sigma(\xi) + \eta$$

$$5. \text{ if } \xi_n \downarrow \xi \text{ then } u_\sigma(\xi_n) \downarrow u_\sigma(\xi)$$

The results of the one period case can be generalised to the multiperiod case:

$$c_{[\sigma, \tau]}(\mathbb{Q}) = \text{ess.sup} \left\{ \mathbb{E}_{\mathbb{Q}}[-\xi \mid \mathcal{F}_\sigma] \mid \begin{array}{l} \xi \in L^\infty(\mathcal{F}_\tau) \\ u_\sigma(\xi) \geq 0 \end{array} \right\}$$

and for $\xi \in L^\infty(\mathcal{F}_\tau)$ we have

$$u_\sigma(\xi) = \text{ess.inf}_{\mathbb{Q} \sim \mathbb{P}} (\mathbb{E}_{\mathbb{Q}}[\xi \mid \mathcal{F}_\sigma] + c_{[\sigma, \tau]}(\mathbb{Q}))$$

As pointed out by Koopmans (1960 — 1961) we need a concept called time consistency. This is defined as follows: for $\xi, \eta \in L^\infty(\mathcal{F}_T)$, two stopping times $\sigma \leq \tau$,

$$u_\tau(\xi) \geq u_\tau(\eta) \text{ implies } u_\sigma(\xi) \geq u_\sigma(\eta)$$

This implies conditions on the family

$$\mathcal{A}_{\sigma, \tau} = \{\xi \in L^\infty(\mathcal{F}_\tau) \mid u_\sigma(\xi) \geq 0\}.$$

Theorem 2. *u is time consistent if and only if for all σ : $\mathcal{A}_{0,T} = \mathcal{A}_{0,\sigma} + \mathcal{A}_{\sigma,T}$. Given u_0 there is at most one time consistent family u_σ that extends u_0 and*

$$\mathcal{A}_{\sigma,T} = \{\xi \in L^\infty(\mathcal{F}_T) \mid \forall A \in \mathcal{F}_\sigma : \xi \mathbf{1}_A \in \mathcal{A}_{0,T}\}.$$

For $\sigma \leq \tau$ this property gives the cocycle property:

$$c_{[\sigma,T]}(\mathbb{Q}) = c_{[\sigma,\tau]}(\mathbb{Q}) + \mathbb{E}_{\mathbb{Q}}[c_{[\tau,T]}(\mathbb{Q}) \mid \mathcal{F}_\sigma]$$

The dynamic programming principle: $u_0(\xi) = u_0(u_\sigma(\xi))$, is equivalent to time consistency, at least on closed time intervals (and it is wrong on open end time intervals).

A family of random variables

$$\{u_\sigma(\xi), c_{\sigma,T}(\mathbb{Q}) \mid \sigma \text{ a stopping time}\}$$

is NOT the same as a stochastic process! We need to answer the question about regularity of trajectories. For this we need an extra assumption on u .

Definition 1. *u is relevant if $\xi \leq 0$ and $\mathbb{P}[\xi < 0] > 0$ imply $u_0(\xi) < 0$, this is e.g. implied by $c_0(\mathbb{P}) = 0$.*

Theorem 3. *If u is relevant there are càdlàg versions of $(c_{t,T}(\mathbb{Q}))_{0 \leq t \leq T}$ and $(c_{t,T}(\mathbb{Q}))_{0 \leq t \leq T}$ is a \mathbb{Q} -potential of class D.*

Example 1. We take a d -dimensional Brownian Motion W . If $\mathbb{Q} \sim \mathbb{P}$ then we write $\frac{d\mathbb{Q}}{d\mathbb{P}} \Big|_{\mathcal{F}_t} = \mathcal{E}(q \cdot W)_t$, $\mathcal{E}(q \cdot W)_t = \exp\left(\int_0^t q_u dW_u - \frac{1}{2} \int_0^t q_u^2 du\right)$. We now take a function $f : \mathbb{R} \rightarrow [0, +\infty]$, convex, lsc, $f(0) = 0$. Its Fenchel-Legendre transform is denoted by g . The function c is then defined as $c_{[t,T]}(\mathbb{Q}) = \mathbb{E}_{\mathbb{Q}}\left[\int_t^T f(q_u) du \mid \mathcal{F}_t\right]$ and this gives

$$u_t(\xi) = \text{ess.inf}_{\mathbb{Q} \sim \mathbb{P}} \mathbb{E}_{\mathbb{Q}} \left[\xi + \int_t^T f(q_u) du \mid \mathcal{F}_t \right].$$

Example 2. For $f(q) = \frac{1}{2}q^2$ we have $g(x) = \frac{1}{2}x^2$ and using the Girsanov-Maruyama theorem we get $c_0(\mathbb{Q}) = \mathbb{E} \left[\frac{d\mathbb{Q}}{d\mathbb{P}} \log \left(\frac{d\mathbb{Q}}{d\mathbb{P}} \right) \right]$, the so-called entropy. This leads to

$$u_t(\xi) = -\log \mathbb{E}[\exp(-\xi) \mid \mathcal{F}_t]$$

and $\exp(-u_t(\xi))$ is a martingale. Up to scaling it is the only law determined time consistent utility function.

The preceding example is very close to the general structure of time consistent utility functions. The following theorem characterises the utility functions in the Brownian Motion setting.

Theorem 4. *Suppose $c_0(\mathbb{P}) = 0$. There is a function*

$$\begin{aligned} f: \mathbb{R}^d \times [0, T] \times \Omega &\rightarrow \overline{\mathbb{R}}_+; f \in \mathcal{R}^d \otimes \mathcal{P} \\ \text{for all } (t, \omega), f &\text{ is convex, lsc in } q \in \mathbb{R}^d \\ f(0, \cdot, \cdot) &= 0 \\ \text{for all } q \in \mathbb{R}^d, f &\text{ is predictable in } (t, \omega) \text{ and} \\ c_0(\mathbb{Q}) &= \mathbb{E}_{\mathbb{Q}} \left[\int_0^T f(q_t(\cdot), t, \cdot) dt \right] \end{aligned}$$

Let us denote by g_t the Fenchel Legendre transform of f_t . Take $\xi \in L^\infty$.

Theorem 5. *In case there is $\mathbb{Q} \sim \mathbb{P}$ with*

$$u_0(\xi) = \min_{\mathbb{Q} \sim \mathbb{P}} \mathbb{E} \left[\xi + \int_0^T f_u(q_u) du \right],$$

we can say more:

$$u_t(\xi) = u_0(\xi) + \int_0^t g_u(Z_u) du - \int_0^t Z_u dW_u.$$

Hence $u_t(\xi)$ is a bounded solution of the BSDE:

$$dY_t = g_t(Z_t) dt - Z_t dW_t, \quad Y_T = \xi$$

In some cases $u_t(\xi)$ is the unique bounded solution of the BSDE. The reader can see that there is a close relation with the concept of g -expectations. However the concept of utility functions is more general than the concept of g -expectations. In case there is $K: \Omega \rightarrow \mathbb{R}_+$ with

$$0 \leq g_t(z, \omega) \leq K(\omega)(1 + |z|^2),$$

g is convex in z , $g(0, t, \omega) = 0$, we can prove:

Theorem 6. *For each $\xi \in L^\infty$, $u_t(\xi)$ is a bounded solution of*

$$dY_t = g_t(Z_t) dt - Z_t dW_t; \quad Y_T = \xi$$

However, uniqueness is not guaranteed. Even when the model is Markovian.

The proof uses the Bishop-Phelps theorem, the preceding result on minimizer + (probably) well known result on submartingales. Let us give an overview of the steps.

Lemma 1. *If $\phi: \mathbb{R}^d \rightarrow \mathbb{R}_+$ is a convex function such that $\phi(z) \leq K(1 + |z|^2)$ for all $z \in \mathbb{R}^d$. Then every $q \in \partial_z \phi$ (q an element of the subgradient of ϕ in the point z) satisfies $|q| \leq 5K(1 + |z|)$.*

Theorem 7. *Suppose that for every $\omega \in \Omega$ there is a constant $K(\omega) < \infty$ depending on ω such that for all t : $g_t(z) \leq K(1 + |z|^2)$. Suppose that \mathbb{Q} is a minimising measure for ξ , then necessarily $\mathbb{Q} \sim \mathbb{P}$.*

Theorem 8. *Suppose X^n is a uniformly bounded (in L^∞) sequence of (continuous) submartingales. Suppose $X^n = A^n + M^n$ is the Doob-Meyer decomposition. Suppose*

$$\| \sup_t |X_t^n - X_t| \|_\infty \rightarrow 0.$$

Then $X = A + M$ where $M^n \rightarrow M$ in BMO.

Remark 1. If M_n is a uniformly bounded sequence of martingales that tends to 0 a.s., then this does not imply that $M_n \rightarrow 0$ in BMO. So we (almost) need $\|\cdot\|_\infty$ convergence. This is structural in the sense that the existence of such sequences is equivalent to the statement: The mapping $H^1 \rightarrow L^1$ is not weakly compact. Hence $L^\infty \rightarrow BMO$ is not weakly compact.

There is also a close connection to non-linear PDE. To make this clear we need to restrict to utility functions that depend on diffusion processes.

We take a d -dimensional Markov process

$$\begin{aligned} dX_t &= \sigma(t, X_t) dW_t + b(t, X_t) dt \\ X_0 &= x_0. \end{aligned}$$

We assume that there is no explosion for the process X but the matrix σ can be degenerate in which case we need some conditions to guarantee that the flow defined by X exists and covers \mathbb{R}^d . This is technical and we do not yet have the best conditions. For $\xi \in L^\infty$ we treat the BSDE:

$$Y_T = \xi; \quad Y \text{ bounded}$$

$$dY_t = g(t, X_t, Z_t) dt - Z_t dW_t.$$

To apply general results, we need to impose conditions on g . This function should satisfy:

For every compact set $D \subset \mathbb{R}^d$:

$$\sup_{x \in D, 0 \leq t \leq T} g(t, x, z) \leq K_D(1 + |z|^2)$$

The Fenchel-Legendre transform of g is then written as:

$$f(u, x, q) = \sup_{z \in \mathbb{R}^d} (q \cdot z - g(u, x, z)).$$

Theorem 9. For $\xi \in L^\infty$, the utility function

$$u_t(\xi) = \text{ess.inf}_{\mathbb{Q} \sim \mathbb{P}} \mathbb{E}_{\mathbb{Q}} \left[\xi + \int_t^T f(u, X_u, q_u) du \mid \mathcal{F}_t \right]$$

satisfies the BSDE.

If $\xi = \phi(X_T)$, then applying the Markov property yields $\phi(t, X_t) = u_t(\xi)$ and because u satisfies the BSDE we immediately get (in viscosity sense)

$$\begin{aligned} \partial_t \phi + \frac{1}{2} \text{trace}(\sigma(t, x) \sigma^*(t, x) D_{x,x}^2 \phi) \\ + \nabla_x \phi b(t, x) = g(t, x, -\nabla_x \phi \sigma(x)) \end{aligned}$$

with ϕ bounded, terminal condition $\phi(T, x)$. We do not discuss the regularity of the solution u and the convergence at T is of course related to the continuity of the endpoint ϕ . The non-uniqueness of the solution is illustrated by the following example.

Example 3. we need the following setup

- $dL_t = -L_t^2 dW_t$ with $L_0 = 1$, this means that L is $\frac{1}{BES^3}$.
- $\xi = \frac{1}{1+L_T} = \frac{R_T}{1+R_T}$, $R = \frac{1}{L}$.
- $g_t(z) = \frac{L_t^2}{1+L_t} |z| = \frac{1}{R_t(1+R_t)} |z|$. This means that u is even coherent.
- There are at least two solutions:
 1. $\frac{1}{1+L_t}$ where $\frac{1}{1+L_0} = 1/2$ and
 2. $u_t(\xi)$
- Both are of the form $\phi(t, R_t)$
- They are different since

$$\lim_{T \rightarrow \infty} u_0(\xi) = 1 > \frac{1}{2}$$

(this is non-trivial)

What is the PDE? Itô's formula gives that $\phi(t, x); x > 0$ satisfies (at least in viscosity sense)

$$\partial_t \phi + \frac{1}{2} \partial_{xx} \phi + \frac{1}{x} \partial_x \phi = \frac{1}{x(x+1)} |\partial_x \phi|,$$

and

$$\phi(T, x) = \frac{x}{1+x}$$

In our case

$$\partial_x \phi \geq 0$$

for both solutions (uses stochastic theory). The PDE now simplifies to

$$\partial_t \phi + \frac{1}{2} \partial_{xx} \phi + \frac{1}{x+1} \partial_x \phi = 0 \text{ on } [0, T] \times (0, \infty).$$

The above equation is the equation that gives martingales for a diffusion of the form

$$dV_t = \frac{1}{V_t + 1} dt + dW_t, V_0 = x > 0.$$

This diffusion does not stay positive since its solution is $V_t = R_t - 1$. So to solve the PDE we either need the value of $\phi(T, x)$ for $0 \geq x \geq -1$ or we need a boundary condition for $u(t, 0)$ when $0 \leq t \leq T$.

Novel Concepts for Nonsmooth Optimization and their Impact on Science and Technology

Kazufumi Ito and Karl Kunisch*

Abstract

A multitude of important problems can be cast as nonsmooth variational problems in function spaces, and hence in an infinite-dimensional, setting. Traditionally numerical approaches to such problems are based on first order methods. Only more recently Newton-type methods are systematically investigated and their numerical efficiency is explored. The notion of Newton differentiability combined with path following is of central importance. It will be demonstrated how these techniques are applicable to problems in mathematical imaging, and variational inequalities. Special attention is paid to optimal control with partial differential equations as constraints.

Mathematics Subject Classification (2010). 35Q93, 46N10, 49K20, 65K10.

Keywords. Non-smooth optimization, semi-smooth Newton methods, optimal control, complementarity problems, ill-posed problems.

1. Introduction

Let X , H be real Hilbert spaces and K a closed convex subset of X . Identify H with H^* and let $\langle \cdot, \cdot \rangle$ denote the duality product on $X^* \times X$. We consider the minimization problem

$$\min \quad f(x) + \varphi(\Lambda x) \quad \text{over } x \in K, \quad (P)$$

*The work of this author was supported by the Austrian Science Fund (FWF) under grant SFB F32 (SFB “Mathematical Optimization and Applications in Biomedical Sciences”).

K. Ito, Department of Mathematics, North Carolina State University, Raleigh, North Carolina, 27695-8205, USA. E-mail: kito@unity.ncsu.edu.

K. Kunisch, Institute of Mathematics and Scientific Computing, University of Graz, Austria. E-mail: karl.kunisch@uni-graz.at.

where $f : X \rightarrow \mathbb{R}$ is a continuously differentiable, convex function, $\Lambda \in \mathcal{L}(X, H)$ and $\varphi : H \rightarrow (-\infty, \infty]$ is a proper, lower semi-continuous, convex function. Typically X and H will be real-valued function spaces over a bounded domain $\Omega \subset \mathbb{R}^n$ with smooth boundary $\partial\Omega$. This is a problem that is well-studied within convex analysis framework. This aspect, as well as first order numerical iterative solution methods are reviewed in part from a non-classical perspective in Section 2. Since φ is not assumed to be regular, classical Newton methods are not directly applicable. In Section 3 the concept of Newton-differentiability and semi-smooth Newton methods are introduced. In the subsequent sections the applicability of these tools is demonstrated for a wide variety of topics, including optimal boundary control in Section 4, optimal control with sparsity constraints in Section 5, time optimal control in Section 6, and data fitting problems in Section 7. The final Section 8 is devoted to a general class of non-linear, non-differentiable complementarity problems. Most of these applications involve differential equations.

2. First Order Augmented Lagrangian Method

In this section we summarize convex analysis techniques for solving (P). For basic convex analysis concepts see [ET, ETu]. Throughout we assume that

$$f, \varphi \text{ are bounded below by zero on } K \quad (\text{A1})$$

$$\langle f'(x_1) - f'(x_2), x_1 - x_2 \rangle \geq \sigma |x_1 - x_2|_X^2 \text{ for all } x_1, x_2 \in K \text{ and } \sigma > 0, \quad (\text{A2})$$

$$\varphi(\Lambda x_0) < \infty \text{ for some } x_0 \in K. \quad (\text{A3})$$

Note that

$$\begin{aligned} f(x) - f(x_0) - \langle f'(x_0), x - x_0 \rangle \\ = \int_0^1 \langle f'(x_0 + t(x - x_0)) - f'(x_0), x - x_0 \rangle dt \geq \frac{\sigma}{2} |x - x_0|^2. \end{aligned}$$

Since φ is proper there exists an element $y_0 \in D(\partial\varphi)$ and

$$\varphi(\Lambda x) - \varphi(y_0) \geq (y_0^*, \Lambda x - y_0)_H \text{ for } y_0^* \in \partial\varphi(y_0),$$

where $\partial\varphi$ denotes the subdifferential of φ . Hence, $\lim_{|x|_X \rightarrow \infty} f(x) + \varphi(\Lambda x) \rightarrow \infty$ and it follows that there exists a unique minimizer $x^* \in K$ for (P).

Theorem 2.1. *The necessary and sufficient condition for $x^* \in K$ to be the minimizer of (P) is given by*

$$\langle f'(x^*), x - x^* \rangle + \varphi(\Lambda x) - \varphi(\Lambda x^*) \geq 0 \text{ for all } x \in K. \quad (1)$$

Proofs to the results of this section can be found in [IK1]. Next a Lagrangian associated to the nonsmooth summand φ in (P) will be introduced, while the condition $x \in K$ is kept as explicit constraint. For this purpose we consider

$$f(x) + \varphi_c(\Lambda x, \lambda) \text{ over } x \in K, \quad (P_c)$$

where the regularization φ_c of φ is defined as the shifted inf-convolution

$$\varphi_c(y, \lambda) = \inf \left\{ \varphi(y - u) + (\lambda, u) + \frac{c}{2} |u|^2 \right\} \quad \text{over } u \in H, \tag{2}$$

for $y, \lambda \in H$ and $c > 0$.

Before we return to the necessary optimality condition, properties of the smooth approximation $\varphi_c(x, \lambda)$ to φ are addressed. For $\lambda > 0$ let $J_\lambda = (I + \lambda \partial\varphi)^{-1}$ denote the resolvent of $\partial\varphi$ and let

$$A_\lambda x = \lambda^{-1} (x - J_\lambda x).$$

stand for the Yosida approximation of $\partial\varphi$.

Theorem 2.2. *For $x, \lambda \in H$ the infimum in (2) is attained at a unique point $u_c(x, \lambda)$ where $u_c(x, \lambda) = x - J_{1/c}(x + c^{-1}\lambda)$. Further $\varphi_c(x, \lambda)$ is convex, (Lipschitz-) continuously Fréchet differentiable in x and $\varphi'_c(x, \lambda) = \lambda + c u_c(x, \lambda) = A_{1/c}(x + c^{-1}\lambda)$. Moreover, $\lim_{c \rightarrow \infty} \varphi_c(x, \lambda) = \varphi(x)$ and*

$$\varphi(J_{1/c}(x + c^{-1}\lambda)) - \frac{1}{2c} |\lambda|^2 \leq \varphi_c(x, \lambda) \leq \varphi(x)$$

for every $x, \lambda \in H$.

In the above statement the prime denotes differentiation with respect to the primal variable x .

Theorem 2.3. *For $x, \lambda \in H$ we have*

$$\varphi_c(x, \lambda) = \sup_{y^* \in H} \left\{ (x, y^*) - \varphi^*(y^*) - \frac{1}{2c} |y^* - \lambda|^2 \right\}, \tag{3}$$

where the supremum is attained at the unique point $\lambda_c(x, \lambda) = \varphi'_c(x, \lambda)$.

Above φ^* denotes the conjugate of φ defined by

$$\varphi^*(y^*) = \sup_{y \in H} \{ (y, y^*) - \varphi(y) \} \quad \text{for } y^* \in H.$$

Remark 2.1. If $\varphi = I_{\{y=0\}}$, where I_S is the indicator function of a set S :

$$I_S(x) = \begin{cases} 0 & \text{if } x \in S \\ \infty & \text{if } x \notin S \end{cases},$$

then $\varphi_c(y, \lambda) = (\lambda, y) + \frac{c}{2} |y|^2$ which is the classical augmented Lagrangian functional associated to equality constraints, [Be, IK1].

Remark 2.2. In many applications the conjugate function φ^* is given by

$$\varphi^*(v) = I_{C^*}(v),$$

where C^* is a closed convex set in H . In this case it follows from Theorem 2.3 that for $v, \lambda \in H$

$$\varphi_c(v, \lambda) = \sup_{y^* \in C^*} \left\{ -\frac{1}{2c} |y^* - (\lambda + cv)|_H^2 \right\} + \frac{1}{2c} (|\lambda + cv|_H^2 - |\lambda|_H^2). \quad (4)$$

Hence the supremum is attained at $\lambda_c(v, \lambda) = \text{Proj}_{C^*}(\lambda + cv)$ where $\text{Proj}_{C^*}(\phi)$ denotes the projection of $\phi \in H$ onto C^* .

The following theorem provides an equivalent characterization of $\lambda \in \partial\varphi(x)$.

Theorem 2.4. *If $\lambda \in \partial\varphi(y)$ for $y, \lambda \in H$, then $\lambda = \varphi'_c(y, \lambda)$ for all $c > 0$. Conversely, if $\lambda = \varphi'_c(y, \lambda)$ for some $c > 0$, then $\lambda \in \partial\varphi(y)$.*

We return to (P_c) . Since $x \rightarrow \varphi_c(\Lambda x, \lambda)$ is bounded from below by $-\frac{1}{2c} |\lambda|_H^2$, the regularized problems (P_c) admit a unique solution $x_c \in K$. The necessary and sufficient optimality condition is given by

$$\langle f'(x_c), x - x_c \rangle + (\varphi'_c(\Lambda x_c, \lambda), \Lambda(x - x_c))_H \geq 0 \quad \text{for all } x \in K. \quad (5)$$

Theorem 2.5. *Assume that there exist $\lambda_c^* \in \partial\varphi(\Lambda x_c)$ for $c \geq 1$ such that $\{|\lambda_c^*|_H\}_{c \geq 1}$ is bounded. Then, x_c converges strongly to x^* in X as $c \rightarrow \infty$ and for each weak cluster point λ^* of $\{\lambda_c\}_{c \geq 1}$ in H*

$$\lambda^* \in \partial\varphi(\Lambda x^*) \quad \text{and} \quad \langle f'(x^*), x - x^* \rangle + (\lambda^*, \Lambda(x - x^*))_H \geq 0 \quad \text{for all } x \in K. \quad (6)$$

Conversely, if $x^ \in K$ satisfies (6) then x^* solves (P) .*

The following lemma addresses the assumption of Theorem 2.5.

Lemma 2.1. *(1) If $\text{dom}(\varphi) = H$ then $\partial\varphi(\Lambda x_c)$ is non-empty and $|\partial\varphi(\Lambda x_c)|_H$ is uniformly bounded for $c \geq 1$.*

(2) If $\varphi = \chi_C$ with C a closed convex set in H and $\Lambda x_c \in C$ for all $c > 0$, then λ_c^ can be chosen to be 0 for all $c > 0$.*

Theorem 2.6. *Assume that there exists a pair $(x^*, \lambda^*) \in K \times H$ that satisfies (6). Then the complementarity condition $\lambda^* \in \partial\varphi(\Lambda x^*)$ can equivalently be expressed as*

$$\lambda^* = \varphi'_c(\Lambda x^*, \lambda^*) \quad (7)$$

and x^ is the unique solution of*

$$\min f(x) + \varphi_c(\Lambda x, \lambda^*) \quad \text{over } x \in K \quad (8)$$

for every $c > 0$.

Note that (7) follows directly from Theorem 2.4. The importance of Theorem 2.6 is given by the fact that the complementarity condition in the form of a differential inclusion is replaced by a nonlinear equation, which is preferable for computations. In the case of Remark (2.2), $\varphi'_c(\Lambda x, \lambda)$ is a projection.

We turn to the discussion of the first order augmented Lagrangian method. Problem (P) is equivalent to

$$\begin{cases} \min f(x) + \varphi(\Lambda x - u) \\ \text{subject to } x \in K \quad \text{and} \quad u = 0 \text{ in } H. \end{cases} \tag{9}$$

To treat the constraint $u = 0$ in (9) by the augmented Lagrangian method we consider the sequential minimization over $x \in K$ and $u \in H$ of the form

$$\min f(x) + \varphi(\Lambda x - u) + (\lambda, u)_H + \frac{c}{2} |u|_H^2, \tag{10}$$

where $\lambda \in H$ is a multiplier and c is a positive scalar penalty parameter [Be, IK1]. Equivalently (10) can be expressed as

$$\min L_c(x, \lambda) = f(x) + \varphi_c(\Lambda x, \lambda) \quad \text{over } x \in K, \tag{11}$$

where $\varphi_c(v, \lambda)$ is defined in (2). The (first-order) augmented Lagrangian method is given next:

Augmented Lagrangian Method

- (i) Choose a starting value $\lambda_1 \in H$, a positive number c and set $k = 1$.
- (ii) Given $\lambda_k \in H$ determine $x_k \in K$ from

$$L_c(x_k, \lambda_k) = \min L_c(x, \lambda_k) \quad \text{over } x \in K.$$

- (iii) Update λ_k by $\lambda_{k+1} = \varphi'_c(\Lambda x_k, \lambda_k)$.
- (iv) If the convergence criterion is not satisfied then set $k = k + 1$ and go to (ii).

The following theorem asserts unconditional convergence with respect to c of the augmented Lagrangian method.

Theorem 2.7. *Assume that there exists $\lambda^* \in \partial\varphi(\Lambda x^*)$ such that (6) is satisfied. Then the sequence (x_k, λ_k) is well-defined and satisfies*

$$\frac{\sigma}{2} |x_k - x^*|_X^2 + \frac{1}{2c} |\lambda_{k+1} - \lambda^*|_H^2 \leq \frac{1}{2c} |\lambda_k - \lambda^*|_H^2, \tag{12}$$

and

$$\sum_{k=1}^{\infty} \frac{\sigma}{2} |x_k - x^*|_X^2 \leq \frac{1}{2c} |\lambda_1 - \lambda^*|_H^2, \tag{13}$$

which implies that $|x_k - x^*|_X \rightarrow 0$ as $k \rightarrow \infty$.

Example 2.1 (Obstacle problem). We consider the problem

$$\begin{cases} \min & \int_{\Omega} (\frac{1}{2} |\nabla u|^2 - \tilde{f} u) dx \quad \text{over } u \in H_0^1(\Omega) \\ \text{subject to} & \phi \leq u \leq \psi \quad \text{a.e. in } \Omega, \end{cases} \quad (14)$$

with $\tilde{f} \in L^2(\Omega)$ and ϕ, ψ given obstacles. In the context of the general framework we choose $X = H_0^1(\Omega)$, $H = L^2(\Omega)$ and $\Lambda =$ the natural injection, and define $f : X \rightarrow \mathbb{R}$ and $\varphi : H \rightarrow \mathbb{R}$ by

$$f(u) = \int_{\Omega} (|\nabla u|^2 - \tilde{f} u) dx \quad \text{and} \quad \varphi(v) = I_C,$$

where $C \subset H$ is the closed convex set defined by $C = \{v \in H : \phi \leq v \leq \psi \text{ a.e. in } \Omega\}$. For one sided constraint $u \leq \psi$ (i.e., $\phi = -\infty$) it is shown from the literature, see e.g. [GLT, IK3], that there exists a unique $\lambda^* \in \partial\varphi(u^*)$ such that (6) is satisfied provided that $\psi \in H^1(\Omega)$, $\psi|_{\Gamma} \geq 0$ and $\sup(0, \tilde{f} + \Delta\psi) \in L^2(\Omega)$. Let us set $C_{\psi} = \{v \in H : v \leq \psi \text{ a.e. in } \Omega\}$. Then we have $I_{C_{\psi}}^*(v) = (\psi, v)$ if $v \geq 0$ a.e. and $I_{C_{\psi}}^*(v) = \infty$ otherwise. By Theorems 2.2, 2.3, for example, we can argue that $\lambda_c(u, \lambda) = \max(0, \lambda + c(u - \psi))$, where \max is the pointwise a.e. operation in Ω . Therefore the optimal pair $(u^*, \lambda^*) \in (H^2 \cap H_0^1) \times L^2$ satisfies

$$\begin{cases} -\Delta u^* + \lambda^* = \tilde{f} \\ \lambda^* = \max(0, \lambda^* + c(u^* - \psi)). \end{cases} \quad (15)$$

In this case Steps 2–3 in the augmented Lagrangian method is given by

$$\begin{aligned} -\Delta u_k + \lambda_{k+1} &= \tilde{f} \\ \lambda_{k+1} &= \max(0, \lambda_k + c(u_k - \psi)). \end{aligned}$$

For bilateral constraints the existence of a multiplier is much more delicate. We refer to [IK1, IK2, IK3] and assume that $\phi, \psi \in H^1(\Omega)$ satisfy

$$\begin{aligned} \phi \leq 0 \leq \psi \quad \text{on } \Gamma, \quad \text{and} \quad \max(0, \Delta\psi + \tilde{f}), \min(0, \Delta\phi + \tilde{f}) \in L^2(\Omega), \\ S_1 = \{x \in \Omega : \Delta\psi + \tilde{f} > 0\} \cap S_2 = \{x \in \Omega : \Delta\phi + \tilde{f} < 0\} \text{ is empty,} \\ -\Delta(\psi - \phi) + c_0(\psi - \phi) \geq 0 \quad \text{a.e. in } \Omega \text{ for some } c_0 > 0. \end{aligned}$$

Once existence of a multiplier in $L^2(\Omega)$ guaranteed, see [IK1] p.123, the optimality system we can use Theorem 2.3 and Theorem 2.6 to express the optimality condition can be expressed as

$$-\Delta u^* + \lambda^* = \tilde{f}, \quad \text{with } \lambda^* \in \partial I_C(\Lambda u^*). \quad (16)$$

The latter expression is equivalent to $u^* \in \partial_C^*(\lambda^*)$. By Remark 2.2 and Theorem 2.4, this is equivalent to $u^* = \text{Proj}_C(\lambda^* + cu^*)$, which after some manipulation can be expressed as

$$\lambda^* = \max(0, \lambda^* + c(u^* - \psi)) + \min(0, \lambda^* + c(u^* - \phi)).$$

The augmented Lagrangian method for the two-sided constraint can be expressed as

$$-\Delta u_k + \lambda_{k+1} = \tilde{f}, \quad \lambda_{k+1} = \max(0, \lambda_k + c(u_k - \psi)) + \min(0, \lambda_k + c(u_k - \phi)).$$

Example 2.2 (Bingham fluid and imaging denoising). The simplified Bingham fluid problem is given by

$$\min \int_{\Omega} \left(\frac{\lambda}{2} |\nabla u|^2 - \tilde{f} u \right) dx + g \int_{\Omega} |\nabla u| dx \quad \text{over } u \in H_0^1(\Omega) \quad (17)$$

where Ω is a bounded open set in R^2 with Lipschitz boundary and $\tilde{f} \in L^2(\Omega)$. In the context of the general theory we choose

$$X = H_0^1(\Omega), \quad H = L^2(\Omega) \times L^2(\Omega), \quad K = X, \quad \text{and} \quad \Lambda = g \text{ grad},$$

and define $f : X \rightarrow R$ and $\varphi : H \rightarrow R$ by

$$f(u) = \frac{1}{g} \int_{\Omega} \left(\frac{\lambda}{2} |\nabla u|^2 - \tilde{f} u \right) dx, \quad \text{and} \quad \varphi(v_1, v_2) = \int_{\Omega} \sqrt{v_1^2 + v_2^2} dx.$$

Since $\text{dom}(\varphi) = H$ it follows from Theorem 2.5 and Lemma 2.1 that there exists λ^* such that (6) holds. Moreover $\varphi^*(v) = \chi_{C^*}(v)$, where $C^* = \{v \in H : |v(x)|_{\mathbb{R}^2} \leq 1 \text{ a.e. in } \Omega\}$. Hence it follows that the optimality system for (17) is given by

$$\begin{cases} \int_{\Omega} (\lambda \nabla u^* \nabla v - \tilde{f} v) dx + \int_{\Omega} (\lambda^* \nabla u^*) dx = 0 \text{ for all } v \in X \\ \lambda^* = \text{Proj}_{C^*}(\lambda^* + c \nabla u^*) = \frac{\lambda^* + c \nabla u^*}{|\lambda^* + c \nabla u^*|_{\mathbb{R}^2}} \text{ a.e. in } \Omega. \end{cases} \quad (18)$$

Moreover steps (ii)-(iii) in the augmented Lagrangian method are given by

$$-\lambda \Delta u_k - g \text{ div } \lambda_{k+1} = \tilde{f}, \quad (19)$$

where

$$\lambda_{k+1} = \begin{cases} \lambda_k + c \nabla u_k & \text{on } A_k = \{x : |\lambda_k(x) + c \nabla u_k(x)|_{\mathbb{R}^2} \leq 1\} \\ \frac{\lambda_k + c \nabla u_k}{|\lambda_k + c \nabla u_k|} & \text{on } \Omega \setminus A_k. \end{cases} \quad (20)$$

Equation (19) is a nonlinear equation for $u_k \in H_0^1(\Omega)$. The augmented Lagrangian method is thus closely related to the explicit duality (Uzawa-) method,

where λ_{k+1} in (19) is replaced by λ_k . The Uzawa method is conditionally convergent in the sense that there exist $0 < \underline{\rho} < \bar{\rho}$ such that it converges for $\rho \in [\underline{\rho}, \bar{\rho}]$, [ET], [GLT]. On the other hand the augmented Lagrangian method converges unconditionally by Theorem 2.7.

The image denoising problem based on BV-regularisation and an additional H^1 semi-norm regularisation term (λ much smaller than g) is given by

$$\min \int_{\Omega} \left(\frac{\lambda}{2} |\nabla u|^2 + g |\nabla u| \right) dx + \frac{1}{2} \int_{\Omega} |u - z|^2 dx \quad \text{over } u \in H^1(\Omega), \quad (21)$$

where z denotes the noise corrupted data. It can be treated analogously as the Bingham fluid problem. For a duality based treatment expressing BV-regularized problems as bilateral obstacle problems we refer to [HK1].

In the simplified friction problem the functional φ is given by $\varphi = \int_{\partial\Omega} |u| ds$. It also can be treated with the concepts of this section, see [IK4].

3. Semi-smooth Newton Method in Function Spaces

In the previous section we discussed how equations such as (16) and (18) can be solved by the augmented Lagrangian method. Due to lack of Fréchet differentiability of the involved operations they are not directly amenable for treatment by the Newton algorithm. Therefore in this section we focus on the notion of Newton differentiability.

Let X and Z be Banach spaces and consider the nonlinear equation

$$F(x) = 0, \quad (22)$$

where $F: D \subset X \rightarrow Z$, and D is an open subset of X .

Definition 3.1. *The mapping $F: D \subset X \rightarrow Z$ is called Newton differentiable in the open subset $U \subset D$ if there exists a family of mappings $G: U \rightarrow \mathcal{L}(X, Z)$ such that*

$$\lim_{h \rightarrow 0} \frac{1}{|h|_X} |F(x+h) - F(x) - G(x+h)h|_Z = 0, \quad (A)$$

for every $x \in U$.

We refer to [CNQ, K, HIK] for work related (A). In [CNQ] the term slant differentiability and in [K], for a slightly different notion, the term Newton map were used. Note that it is not required that the mapping G serving as generalized (or Newton) derivative is not required to be unique. The following convergence result is well known [CNQ, HIK].

Theorem 3.1. *Suppose that x^* is a solution to (22) and that F is Newton differentiable in an open neighborhood U containing x^* with Newton derivative*

$G(x)$. If $G(x)$ is nonsingular for all $x \in U$ and $\{\|G(x)^{-1}\| : x \in U\}$ is bounded, then the Newton-iteration

$$x^{k+1} = x^k - G(x^k)^{-1}F(x^k)$$

converges superlinearly to x^* provided that $\|x^0 - x^*\|$ is sufficiently small.

Proof. Note that the Newton iterates satisfy

$$|x^{k+1} - x^*| \leq |G(x^k)^{-1}| |F(x^k) - F(x^*) - G(x^k)(x^k - x^*)|, \tag{23}$$

provided that $x^k \in U$. Let $B(x^*, r)$ denote a ball of radius r centered at x^* contained in U and let M be such that $\|G(x)^{-1}\| \leq M$ for all $x \in B(x^*, r)$. We apply (A) with $x = x^*$. Let $\eta \in (0, 1]$ be arbitrary. Then there exists $\rho \in (0, r)$ such that

$$|F(x^* + h) - F(x^*) - G(x^* + h)h| < \frac{\eta}{M} |h| \leq \frac{1}{M} |h| \tag{24}$$

for all $|h| < \rho$. Consequently, if we choose x^0 such that $|x^0 - x^*| < \rho$, then by induction from (23), (24) with $h = x^k - x^*$ we have $|x^{k+1} - x^*| < \rho$ and in particular $x^{k+1} \in B(x^*, \rho)$. It follows that the iterates are well-defined. Moreover, since $\eta \in (0, 1]$ is chosen arbitrarily $x^k \rightarrow x^*$ converges superlinearly. \square

Here we are especially interested in applications involving the pointwise max operation when X is a function space consisting of elements defined over a bounded domain $\Omega \subset \mathbb{R}^n$ with Lipschitzian boundary $\partial\Omega$. Let $\delta \in \mathbb{R}$ be fixed arbitrarily. We introduce candidates for Newton derivatives G_m of the form

$$G_m(x)(s) = \begin{cases} 1 & \text{if } x(s) > 0, \\ 0 & \text{if } x(s) < 0, \\ \delta & \text{if } x(s) = 0, \end{cases} \tag{25}$$

where $x \in X$.

Proposition 3.1. (i) G_m can in general not serve as a Newton derivative for $\max(0, \cdot) : L^p(\Omega) \rightarrow L^p(\Omega)$, for $1 \leq p \leq \infty$.

(ii) The mapping $\max(0, \cdot) : L^q(\Omega) \rightarrow L^p(\Omega)$ with $1 \leq p < q \leq \infty$ is Newton differentiable on $L^q(\Omega)$ and G_m is a Newton derivative.

For the proof which directly verifies property (A), see [HIK]. Alternatively, if $\psi : \mathbb{R} \rightarrow \mathbb{R}$ is semi-smooth in the sense of mappings between finite-dimensional spaces, i.e. ψ is locally Lipschitz continuous and $\lim_{V \in \partial\psi(x+th'), h' \rightarrow h, t \rightarrow 0^+} Vh'$ exists for all $h \in \mathbb{R}$, then the substitution operator $F : L^q(\Omega) \rightarrow L^p(\Omega)$ defined by

$$F(x)(s) = \psi(x(s)) \text{ for a.e. } s \in \Omega$$

is Newton differentiable on $L^q(\Omega)$, if $1 \leq p < q \leq \infty$, see [U]. In particular this applies to the max operation.

The following chain rule is useful in many applications.

Proposition 3.2. *Let $f : Y \rightarrow Z$ and $g : X \rightarrow Y$ be Newton differentiable in open sets V and U , respectively, with $U \subset X, g(U) \subset V \subset Y$. Assume that g is locally Lipschitz continuous and that there exists a Newton map $G_f(\cdot)$ associated to f which is bounded on $g(U)$. Then the superposition $f \circ g : X \rightarrow Z$ is Newton differentiable in U with a Newton map $G_f G_g$.*

For the proof we refer to [HK3].

A class of nonlinear complementarity problems: The above concepts are applied to nonlinear complementarity problems of the form

$$g(x) + \lambda = 0, \quad \lambda \geq 0, \quad x \leq \psi \quad \text{and} \quad (\lambda, x - \psi)_{L^2} = 0, \tag{26}$$

where $g : X = L^2(\Omega) \rightarrow L^p(\Omega), p > 2$ is Lipschitz continuous and $\psi \in L^p(\Omega)$. If J is a continuously differentiable functional on X then (26) with $g = J'$, is the necessary optimality condition for

$$\min_{x \in L^2(\Omega)} J(x) \quad \text{subject to } x \leq \psi. \tag{27}$$

As discussed in the previous section, (26) can equivalently be expressed as

$$g(x) + \lambda = 0, \quad \lambda = \max(0, \lambda + c(x - \psi)), \tag{28}$$

for any $c > 0$, where \max denotes the pointwise max-operation, with λ the Lagrange multiplier associated to the inequality constraint.

Let us assume that (28) admits a solution $(x^*, \lambda^*) \in L^2(\Omega) \times L^2(\Omega)$. Equation (28) can equivalently be expressed as

$$F(x) = g(x) + \max(0, -g(x) + c(x - \psi)) = 0, \tag{29}$$

where F is considered as mapping from X into itself. The semi-smooth Newton iteration for this reduced equation is given by

$$\begin{aligned} g'(x^k)(x^{k+1} - x^k) + G(-g(x^k) + c(x^k - \psi))(-g'(x)(x^{k+1} - x^k) \\ + c(x^{k+1} - x^k)) + g(x^k) + \max(0, -g(x^k) + c(x^k - \psi)) = 0, \end{aligned} \tag{30}$$

where G_m was defined in (25). To investigate local convergence of (30) we denote for any partition $\Omega = \mathcal{A} \cup \mathcal{I}$ into measurable sets \mathcal{I} and \mathcal{A} by $R_{\mathcal{I}} : L^2(\Omega) \rightarrow L^2(\mathcal{I})$ the canonical restriction operator and by $R_{\mathcal{I}}^* : L^2(\mathcal{I}) \rightarrow L^2(\Omega)$ its adjoint. Further we set

$$g'(x)_{\mathcal{I}} = R_{\mathcal{I}} g'(x) R_{\mathcal{I}}^*.$$

Proposition 3.3. *Assume that (28) admits a solution x^* , that $x \rightarrow g(x) - c(x - \psi)$ is a C^1 function from $L^2(\Omega)$ to $L^p(\Omega)$ in a neighborhood U of x^* for some $c > 0$ and $p > 2$, and that*

$$\{g'(x)_{\mathcal{I}}^{-1} \in \mathcal{L}(L^2(\mathcal{I})) : x \in U, \Omega = \mathcal{A} \cup \mathcal{I}\} \text{ is uniformly bounded.}$$

Then the iterates x^k defined by (30) converge superlinearly to x^* , provided that $|x^* - x^0|$ is sufficiently small. Here x^0 denotes the initialization of the algorithm.

Proof. By Propositions 3.1, 3.2 the mapping $x \rightarrow \max(0, -g(x) + c(x - \psi))$ is Newton differentiable in U as mapping from $L^2(\Omega)$ into itself and $G_m(-g(x) + c(x - \psi))(-g'(x) + cI)$ is a Newton-derivative. Consequently F is Newton differentiable in U . Moreover $g'(x) + G_m(-g(x) + c(x - \psi))(-g' + cI)$ is invertible in $\mathcal{L}(L^2(\Omega))$ with uniformly bounded inverses for $x \in U$. In fact, setting

$$z = -g(x) + c(x - \psi), \mathcal{A} = \{z > 0\}, \mathcal{I} = \Omega \setminus \mathcal{A}, h_{\mathcal{I}} = \chi_{\mathcal{I}}h, h_{\mathcal{A}} = \chi_{\mathcal{A}}h,$$

this follows from the fact that for given $f \in L^2(\Omega)$ the solution to the equation

$$g'(x)h + G(z)(-g'(x)h + ch) = f$$

is given by

$$ch_{\mathcal{A}} = f_{\mathcal{A}} \text{ and } h_{\mathcal{I}} = g'_{\mathcal{I}}(x)^{-1} \left(f_{\mathcal{I}} - \frac{1}{c} \chi_{\mathcal{I}} g'(x) f_{\mathcal{A}} \right).$$

From Theorem 3.1 we conclude that $x^k \rightarrow x^*$ superlinearly, provided that $|x^* - x^0|$ is sufficiently small. □

It can be observed that the semi-smooth Newton step can be equivalently expressed as

$$\begin{aligned} g'(x^k)(x^{k+1} - x^k) + g(x^k) + \lambda^{k+1} &= 0 \\ x^{k+1} &= \psi \text{ in } \mathcal{A}_k = \{s : -g(x^k)(s) + c(x^k(s) - \psi(s)) > 0\} \\ \lambda^{k+1} &= 0 \text{ in } \mathcal{I}_k = \{s : -g(x^k)(s) + c(x^k(s) - \psi(s)) \leq 0\}. \end{aligned} \tag{31}$$

Remark 3.1. We refer to (31) as the primal-dual active set strategy for the reduced equation. If the semi-smooth Newton step is applied to (28) rather than to the reduced equation, then the resulting algorithm differs in the update of the active/inactive sets. In fact, in this case the update for the active set is given by $\mathcal{A}_k = \{s : \lambda^k(s) + c(x^k(s) - \psi(s)) > 0\} = \{s : -g(x^{k-1})(s) - g'(x^{k-1})(x^k - x^{k-1})(s) + c(x^k(s) - \psi(s)) > 0\}$. In case g is linear the two updates coincide.

If we consider regularized least squares problems of the form

$$\min J(x) = \frac{1}{2} |Tx - z|_Y^2 + \frac{\alpha}{2} |x|_{L^2}^2, \text{ subject to } x \leq \psi, \tag{32}$$

where Y is a Hilbert space, $T \in \mathcal{L}(L^2(\Omega), Y)$, $\alpha > 0$ and $z \in Y, \psi \in L^p(\Omega)$, then $g(x) = T^*(Tx - z) + \alpha x$ and $g(x) - \alpha(x - \psi) = T^*(Tx - z) + \alpha\psi$. Hence Proposition 3.3 with $c = \alpha$ is applicable if $T^* \in \mathcal{L}(Y, L^p(\Omega))$, for some $p > 2$.

The optimality condition (29) is given by

$$\alpha(x - \psi) + \max(0, T^*(Tx - z) + \alpha\psi) = 0, \tag{33}$$

in this case.

So far we addressed local convergence. The following result gives a sufficient condition for global convergence.

Proposition 3.4. *Consider (32) and assume that $\|T\|_{\mathcal{L}(L^2(\Omega_1), L^2(\Omega_2))}^2 < \alpha$. Then the semi-smooth Newton algorithm converges independently of the initialisation to the unique solution of (32).*

The proof is based an argument using

$$M(x, \lambda) = \alpha^2 \int_{\Omega} |(x - \psi)^+|^2 ds + \int_{\mathcal{A}(x)} |\lambda^-|^2 ds$$

as a merit function, where $\mathcal{A}(x) = \{s : x(s) \geq \psi(s)\}$. It decays along the iterates (x^k, λ^k) of the semi-smooth Newton algorithm. An analogous result can be obtained in case of bilateral constraints and for nonlinear mappings g , if additional requirements are met, [IK3].

Propositions 3.3 and 3.4 are applicable to optimal control problems with control constraints, for example. This is the contents of the following section.

4. Optimal Dirichlet Boundary Control

Let us consider the Dirichlet boundary optimal control problem with point-wise constraints on the boundary, formally given by

$$\left\{ \begin{array}{l} \min \frac{1}{2} \|y - z\|_{L^2(Q)}^2 + \frac{\alpha}{2} \|u\|_{L^2(\Sigma)}^2 \\ \text{subject to} \\ \partial_t y - \kappa \Delta y + b \cdot \nabla y = f \quad \text{in } Q \\ y = u, \quad u \leq \psi \quad \text{on } \Sigma \\ y(0) = y_0 \quad \text{in } \Omega, \end{array} \right. \tag{34}$$

where $Q = (0, T] \times \Omega$, $\Sigma = (0, T] \times \partial\Omega$ and Ω a bounded domain in \mathbb{R}^n , $n \geq 2$ with C^2 boundary $\partial\Omega$. This guarantees that the Laplacian with homogenous Dirichlet boundary conditions, denoted by Δ , is an isomorphism from $H^2(\Omega) \cap H_0^1(\Omega)$ to $L^2(\Omega)$. Further $\kappa > 0$, $y_0 \in H^{-1}(\Omega)$, $z \in L^2(Q)$, $f \in L^2(H^{-2}(\Omega))$, $u \in L^2(\Sigma)$ and $b \in \mathbb{L}^\infty(Q)$, $\text{div } b \in L^\infty(L^{\hat{n}}(\Omega))$ where $\hat{n} = \max(n, 3)$, and $\mathbb{L}^\infty(Q) = \bigotimes_{i=1}^n L^\infty(Q)$.

Under these conditions there exists a unique very weak solution $y \in L^2(Q) \cap H^1(H^{-2}(\Omega)) \cap C(H^{-1}(\Omega))$ satisfying for a.e. $t \in (0, T)$

$$\left\{ \begin{array}{l} \langle \partial_t y(t), v \rangle - \kappa \langle y(t), \Delta v \rangle - \langle y(t), \text{div}(b(t))v \rangle - \langle y(t), b(t)\nabla v \rangle \\ \quad = \langle f(t), v \rangle - \kappa \langle u(t), \frac{\partial}{\partial n} v \rangle_{\partial\Omega} \text{ for all } v \in H^2(\Omega) \cap H_0^1(\Omega), \\ y(0) = y_0, \end{array} \right. \tag{35}$$

where $\langle \cdot, \cdot \rangle = \langle \cdot, \cdot \rangle_{H^{-2}(\Omega), H^2(\Omega) \cap H_0^1(\Omega)}$ denotes the canonical duality pairing, (\cdot, \cdot) and $(\cdot, \cdot)_{\partial\Omega}$ stand for the inner products in $L^2(\Omega)$ and $L^2(\partial\Omega)$ respectively. Moreover

$$|y|_{L^2(Q) \cap H^1(H^{-2}(\Omega)) \cap C(H^{-1}(\Omega))} \leq C(|y_0|_{H^{-1}(\Omega)} + |f|_{L^2(H^{-2}(\Omega))} + |u|_{L^2(\Sigma)}), \tag{36}$$

where C depends continuously on $\kappa > 0$, $|b|_{L^\infty(Q)}$ and $|\operatorname{div} b|_{L^\infty(L^n(\Omega))}$, and is independent of $f \in L^2(H^{-2}(\Omega))$, $u \in L^2(\Sigma)$ and $y_0 \in H^{-1}(\Omega)$.

Utilizing the a-priori bound (36) it is straightforward to argue the existence of a unique solution $u^* \in L^2(\Sigma)$ of (34). It can be shown that it is characterized by the optimality system

$$\begin{cases} \partial_t y - \kappa \Delta y + b \cdot \nabla y = f \text{ in } Q, \\ y = u \text{ on } \Sigma, \quad y(0) = y_0 \text{ in } \Omega, \\ -\partial_t p - \kappa \Delta p - \operatorname{div} b p - b \cdot \nabla p = -(y - z) \text{ in } Q, \\ p = 0 \text{ on } \Sigma, \quad p(T) = 0 \text{ in } \Omega, \\ \kappa \frac{\partial p}{\partial n} + \alpha u + \lambda = 0 \text{ on } \Sigma, \\ \lambda = \max(0, \lambda + c(u - \psi)) \text{ on } \Sigma, \end{cases} \tag{37}$$

where the primal must be interpreted in the very weak form. In terms of (32) we have that the operator $T : L^2(\Sigma) \rightarrow L^2(Q)$ is given as the control to state operator for (35). Its adjoint $T^* \in \mathcal{L}(L^2(Q), L^2(\Sigma))$ is the solution of the adjoint equation, i.e. the third and fourth equations in (37), with right hand side $\varphi \in L^2(Q)$. In [KV] we verified that the adjoint satisfies

$$\left\| \frac{\partial p}{\partial n} \right\|_{L^{q_n}(\Sigma)} \leq C_1 \|p\|_{L^2(H^2(\Omega)) \cap H^1(L^2(\Omega))} \leq C_2 \|\varphi\|_{L^2(Q)},$$

with an embedding constant C_1, C_2 , where

$$q_n = \begin{cases} \frac{2(n+1)}{n}, & \text{if } n \geq 3, \\ 3 - \varepsilon, & \text{if } n = 2, \end{cases}$$

for every $\varepsilon > 0$, so that in particular $q_n > 2$ for every n . Equation (33) is given by

$$\alpha u - \psi + \max \left(0, \kappa \frac{\partial p}{\partial n} + \alpha \psi \right) = 0, \tag{38}$$

in this case, and Propositions 3.3 and 3.4 imply that the semi-smooth Newton method applied to (38) converges locally superlinearly, as well as globally, if $\alpha > \|T\|_{\mathcal{L}(L^2(\Sigma), L^2(\Omega))}^2$.

5. Sparse Controls

The control cost in optimal control problems is most frequently chosen to be of the form $\frac{\alpha}{2}|u|^2$, where u denotes the control. In this way the control cost is differentiable, in some applications the term can be interpreted as energy. It is indispensable in the stochastic interpretation of the linear quadratic regulator theory. However, it also has drawbacks, most notably, it does not put proportional weight on the control. The purpose of this section is to sketch a framework for the use of $\alpha|u|$ as control cost. For this choice the cost of the control is proportional to its “size”. Moreover it has the feature of being sparse. To get an appreciation for this latter property let us consider the non-differentiable problem in $L^2(\Omega)$ given by

$$\min \frac{1}{2}|u - z|_{L^2}^2 + \alpha|u|_{L^1}. \quad (39)$$

The solution to (39) is given in the a.e. sense by

$$u^* = \begin{cases} 0 & \text{if } |z| < \alpha^{-1} \\ z - \alpha^{-1} \operatorname{sgn} z & \text{if } |z| \geq \alpha^{-1}. \end{cases} \quad (40)$$

In particular the solution is 0 where z is small relative to $1/\alpha$. The space of $L^1(\Omega)$ -controls, however, does not lend itself to weak* compactness arguments which are needed to guarantee existence in the context of optimal control. Consequently the control space is enlarged to measure-valued controls. We consider the model problem

$$\begin{cases} \min_{u \in \mathcal{M}} \frac{1}{2}|y - z|_{L^2}^2 + \alpha|u|_{\mathcal{M}} \\ \text{s.t. } Ay = u, \end{cases} \quad (\mathcal{P}_{\mathcal{M}})$$

where \mathcal{M} denotes the vector space of all bounded Borel measures on Ω , that is the space of all bounded σ -additive set functions $\mu : \mathcal{B}(\Omega) \rightarrow \mathbb{R}$ defined on the Borel algebra $\mathcal{B}(\Omega)$ satisfying $\mu(\emptyset) = 0$. The total variation of $\mu \in \mathcal{M}$ is defined for all $B \in \mathcal{B}(\Omega)$ by $|\mu|(B) := \sup \{ \sum_{i=0}^{\infty} |\mu(B_i)| : \bigcup_{i=0}^{\infty} B_i = B \}$, where the supremum is taken over all partitions of B . Endowed with the norm $|\mu|_{\mathcal{M}} = |\mu|(\Omega)$, \mathcal{M} is a Banach space. By the Riesz representation theorem, \mathcal{M} can be isometrically identified with the topological dual of $C_0(\Omega)$. This leads to the following equivalent characterization of the norm on \mathcal{M} :

$$|\mu|_{\mathcal{M}} = \sup_{\substack{\phi \in C_0(\Omega), \\ |\phi|_{C_0} \leq 1}} \int_{\Omega} \phi d\mu. \quad (41)$$

Further A is a second order elliptic operator with homogenous Dirichlet boundary conditions in the bounded domain $\Omega \subset \mathbb{R}^n$ with $n \in \{2, 3\}$, and such that

$$\|A \cdot\|_{L^2} \text{ and } \|A^* \cdot\|_{L^2} \text{ are equivalent norms on } H^2(\Omega) \cap H_0^1(\Omega),$$

where A^* denotes the adjoint of A with respect to the inner product in L^2 . For $u \in \mathcal{M}$, the equation $Ay = u$ has a unique weak solution $y \in W_0^{1,p}(\Omega)$, for all $1 \leq p < \frac{n}{n-1}$. Furthermore, there exists a constant $C > 0$ such that $|y|_{W_0^{1,p}} \leq C|u|_{\mathcal{M}}$.

Since $W_0^{1,p}(\Omega)$ is compactly embedded in $L^2(\Omega)$, $(\mathcal{P}_{\mathcal{M}})$ is well-defined, and standard arguments imply the existence of a unique solution (y^*, u^*) . Next we aim for a formulation of the problem that is appropriate for computational purposes. By Fenchel duality theory the predual to $(\mathcal{P}_{\mathcal{M}})$ is given by

$$\begin{cases} \min_{p \in H^2 \cap H_0^1} \frac{1}{2}|A^*p + z|^2 - \frac{1}{2}|z|_{L^2}^2 \\ \text{s.t. } |p|_{C_0} \leq \alpha, \end{cases} \tag{\mathcal{P}_{\mathcal{M}}^*}$$

which can be considered as a bilaterally constraint problem. Existence of a unique solution p^* can readily be verified and the relationship between solutions to the original and the predual problem are given by:

$$\begin{cases} Ay^* = u^*, \\ A^*p^* = z - y^*, \\ 0 \leq \langle u^*, p^* - p \rangle_{(H^2 \cap H_0^1)^*, H^2 \cap H_0^1} \text{ for all } p \in H^2 \cap H_0^1, |p|_{C_0} \leq \alpha. \end{cases} \tag{42}$$

The inequality in (42) can be interpreted as the larger α , the smaller is the support of the control u^* .

While $(\mathcal{P}_{\mathcal{M}}^*)$ is of bilateral constraint type, some further consideration is required before Newton methods can be used efficiently. Comparing to (32) and the optimal control problem treated in Section 4, the operator appearing in $(\mathcal{P}_{\mathcal{M}}^*)$ is not of smoothing type. Note that if we were to discretize (32) and $(\mathcal{P}_{\mathcal{M}}^*)$ then these problems have the same structure. But this is not the case on the continuous level. Computationally this becomes apparent in the context of mesh independence. Applying the semi-smooth Newton method to the discretized form of (32) with T satisfying $T^* \in \mathcal{L}(Y, L^p(\Omega))$ will result in mesh-independent iteration numbers of the semi-smooth Newton method, while this is not the case for $(\mathcal{P}_{\mathcal{M}}^*)$. For an analysis of mesh-independence of the semi-smooth Newton method we refer to [HU].

To obtain a formulation which is appropriate for a super-linear and mesh-independent behavior of the semi-smooth Newton method some type of regularization is required. For example an additional regularization term of the form $\frac{\beta}{2}|u|_{L^2}^2$ can be added to the cost in $(\mathcal{P}_{\mathcal{M}})$, see e.g. [St]. Here we go a different way and consider the Moreau-Yosida approximation, see (15), of the inequality constraints leading to

$$\min_{p \in H^2 \cap H_0^1} \frac{1}{2}|A^*p + z|_{L^2}^2 - \frac{1}{2}|z|_{L^2}^2 + \frac{c}{2}|\max(0, p - \alpha)|_{L^2}^2 + \frac{c}{2}|\min(0, p + \alpha)|_{L^2}^2, \tag{\mathcal{P}_{\mathcal{M},c}^*}$$

where the max- and min- operations are taken pointwise in Ω . For $c > 0$ let p_c denote the solutions to $(\mathcal{P}_{\mathcal{M},c}^*)$. They satisfy the optimality system

$$\begin{cases} AA^*p_c + Az + \lambda_c = 0, \\ \lambda_c = \max(0, c(p_c - \alpha)) + \min(0, c(p_c + \alpha)), \end{cases} \quad (43)$$

where $\lambda_c \in W^{1,\infty}$ approximates the Lagrange multiplier associated to the constraint $|p|_{C_0} \leq \alpha$. Let $(p^*, \lambda^*) \in H^2 \cap H_0^1 \times (H^2 \cap H_0^1)^*$ denote the unique solution to the optimality system for $(\mathcal{P}_{\mathcal{M}}^*)$:

$$\begin{cases} AA^*p^* + Az + \lambda^* = 0, \\ \langle \lambda^*, p - p^* \rangle \leq 0, \end{cases} \quad (44)$$

for all $p \in H^2 \cap H_0^1$ with $|p|_{C_0} \leq \alpha$. Then, see [CK], as $c \rightarrow \infty$:

$$p_c \rightarrow p^* \quad \text{in } H^2 \cap H_0^1, \quad \lambda_c \rightarrow \lambda^* \quad \text{in } (H^2 \cap H_0^1)^*. \quad (45)$$

The regularized optimality system $(\mathcal{P}_{\mathcal{M},c}^*)$ can be solved efficiently by the semi-smooth Newton method with G_m as in (25) and appropriately adapted for the min term. For this purpose we express $(\mathcal{P}_{\mathcal{M},c}^*)$ as a nonlinear equation $F(p) = 0$ with $F : H^2 \cap H_0^1 \rightarrow (H^2 \cap H_0^1)^*$, where

$$F(p) := AA^*p + \max(0, c(p - \alpha)) + \min(0, c(p + \alpha)) + Az. \quad (46)$$

Due to the regularity gap between the domain and the range of F the following result can be obtained quite readily from Theorem 3.1, and Proposition 3.1, [CK].

Theorem 5.1. *If $|p_c - p^0|_{H^2 \cap H_0^1}$ is sufficiently small, the iterates p^k of the semi-smooth Newton algorithm converge superlinearly in $H^2 \cap H_0^1$ to the solution p_c of $(\mathcal{P}_{\mathcal{M},c}^*)$ as $k \rightarrow \infty$.*

For this application, let us give the algorithm in detail in Algorithm 1. The stopping criterion is typically met without any need for globalization. If it applies then the algorithm stops at the solution of (43). For actual computations a discretisation of the infinite dimensional spaces is required. This is not within the scope of this paper.

The question also arises how to choose c in practice. Large c implies that we can be close to the solution of the unregularized problem at the expense of possible ill-conditioning of the regularized one. We have only rarely experienced that ill-conditioning actually occurs. In practice it is certainly advisable to utilize a continuation principle, applying the Algorithm with a moderate value for c , and utilizing the solution thus obtained as initialization for a computation with a larger value for c . This procedure can be put onto solid ground for unilateral constraints by means of path following techniques as detailed in

Algorithm 1 Semismooth Newton method for (43)

- 1: Set $k = 0$, Choose $p^0 \in H^2 \cap H_0^1$
- 2: **repeat**
- 3: Set

$$\mathcal{A}_{k+1}^+ = \{x\} p^k(x) > \alpha, \quad \mathcal{A}_{k+1}^- = \{x\} p^k(x) < -\alpha, \quad \mathcal{A}_{k+1} = \mathcal{A}_{k+1}^+ \cup \mathcal{A}_{k+1}^-$$
- 4: Solve for $p^{k+1} \in H^2 \cap H_0^1$:

$$(A^* p^{k+1}, A^* v)_{L^2} + c(p^{k+1} \chi_{\mathcal{A}_k}, v)_{L^2} = -(z, A^* v)_{L^2} + c\alpha(\chi_{\mathcal{A}_k^+} - \chi_{\mathcal{A}_k^-}, v)_{L^2}$$

for all $v \in H^2 \cap H_0^1$
- 5: Set $k = k + 1$
- 6: **until** $(\mathcal{A}_{k+1}^+ = \mathcal{A}_k^+)$ and $(\mathcal{A}_{k+1}^- = \mathcal{A}_k^-)$

[HK2]. Since the infinite dimensional problem always needs to be discretized a natural stopping criterion for the increase of c is given once the error due to regularization is smaller than that of discretization. For certain obstacle type problems which satisfy a maximum principle the L^∞ error due to regularization can be estimated, see [IK5]. – Concerning regularization let us stress that discretization also has a regularizing effect. In this case staggered grid strategies applied to the original unregularized formulation, i.e. $(\mathcal{P}_{\mathcal{M}}^*)$ in our case, correspond to the increase of the regularisation parameter c and can be very effective in numerical computations. The formal analysis of this procedure has not been carried out yet.

6. Time Optimal Control

This section is devoted to time optimal control problems for a class of nonlinear ordinary differential equations. The techniques are applicable to much wider class of problems, but the detailed analysis yet needs to be carried out. While the computation of time optimal controls and trajectories has a long history, the use of Newton-type methods is a very recent one. We refer to [IK4] for a detailed description of the procedure that we describe here.

$$\begin{cases} \min_{\tau \geq 0} \int_0^\tau dt & \text{subject to} \\ \frac{d}{dt} x(t) = Ax(t) + Bu(t), |u(t)|_{\ell^\infty} \leq 1, x(0) = x_0, x(\tau) = x_1, \end{cases} \quad (P_T)$$

where $A \in \mathbb{R}^{n \times n}$, $B \in \mathbb{R}^{n \times m}$, $x_0 \neq x_1$ are given vectors in \mathbb{R}^n , $u(t) \in \mathbb{R}^m$, u is measurable, and $|\cdot|_{\ell^\infty}$ denotes the infinity-norm on \mathbb{R}^m . It is assumed that x_1 can be reached in finite time by an admissible control. Then (P_T) admits a solution with optimal time denoted by τ^* , and associated optimal state x^* and optimal control u^* .

It is wellknown that under appropriate conditions [HL] the optimal solution is related to the adjoint equation

$$p(t) = \exp(A^T(\tau^* - t))q, \text{ with } q \in \mathbb{R}^n,$$

through

$$u^*(t) = -\sigma(B^T p(t)) = -\sigma(B^T \exp(-A^T(\tau^* - t))q), \tag{47}$$

for $t \in [0, \tau^*]$, where $q \in \mathbb{R}^n$ and σ denotes the coordinate-wise operation

$$\sigma(s) \in \begin{cases} \text{sgn } s & \text{if } s \neq 0 \\ [-1, 1] & \text{if } s = 0. \end{cases} \tag{48}$$

This operation prohibits the use of superlinear Newton-type methods for solving (P_T) numerically. Therefore a family of regularized problems given by

$$\begin{cases} \min_{\tau \geq 0} \int_0^\tau (1 + \frac{\varepsilon}{2} |u(t)|^2) dt & \text{subject to} \\ \frac{d}{dt}x(t) = Ax(t) + Bu(t), |u(t)|_{\ell^\infty} \leq 1, x(0) = x_0, x(\tau) = x_1, \end{cases} \tag{P_\varepsilon}$$

with $\varepsilon > 0$ is considered. The norm $|\cdot|$ used in the cost-functional denotes the Euclidean norm. It is straightforward to argue the existence of a solution $(u_\varepsilon, x_\varepsilon, \tau_\varepsilon)$. Convergence of the solutions $(x_\varepsilon, u_\varepsilon, \tau_\varepsilon)$ of (P_ε) to a solution (x^*, u^*, τ^*) of (P_T) was analysed in [IK4]. Note that τ^* is unique.

Proposition 6.1. *For every $0 < \varepsilon_0 < \varepsilon_1$ and any solution (τ^*, u^*) of (P) we have*

$$\tau^* \leq \tau_{\varepsilon_0} \leq \tau_{\varepsilon_1} \leq \tau^* \left(1 + \frac{\varepsilon_1}{2}\right), \tag{49}$$

$$|u_{\varepsilon_1}|_{L^2(0, \tau_{\varepsilon_1})} \leq |u_{\varepsilon_0}|_{L^2(0, \tau_{\varepsilon_0})} \leq |u^*|_{L^2(0, \tau^*)}. \tag{50}$$

If u^* is a bang-bang solution, then

$$0 \leq |u^*|_{L^2(0, \tau^*)}^2 - |u_\varepsilon|_{L^2(0, \tau_\varepsilon)}^2 \leq \text{meas } \{t \in [0, \tau^*] : |u_\varepsilon(t)| < 1\} \tag{51}$$

for every $\varepsilon > 0$. Moreover, if (A, B_i) is controllable for each column B_i of B , then the solution u^* is unique, it is bang-bang and $u_\varepsilon \rightarrow u^*$ in L^2 as $\varepsilon \rightarrow 0^+$.

Recall that a control is called bang-bang if $|u_i(t)| = 1$ for all $t \in [0, \tau^*]$ and $i = 1, \dots, m$. Concerning a necessary optimality condition for (P_ε) we have the following result:

Theorem 6.1. *Let $(x_\varepsilon, u_\varepsilon, \tau_\varepsilon)$ be a solution of (P_ε) . Assume that there exists some \tilde{i} such that*

$$(A, B_{\tilde{i}}) \text{ is controllable,} \tag{H1}$$

and such that exist $\eta > 0$ and an interval $I_{\tilde{i}} \subset (0, 1)$ satisfying

$$|(\hat{u}_\varepsilon)_{\tilde{i}}(t)|_{\ell^\infty} \leq 1 - \eta \text{ for a.e. } t \in I_{\tilde{i}}. \tag{H2}$$

Then there exists an adjoint state p_ε such that

$$\begin{cases} \dot{x}_\varepsilon = Ax_\varepsilon + Bu_\varepsilon, & x_\varepsilon(0) = x_0, & x_\varepsilon(\tau_\varepsilon) = x_1, \\ -\dot{p}_\varepsilon = A^T p_\varepsilon, \\ u_\varepsilon = -\sigma_\varepsilon(B^T p_\varepsilon), \\ 1 + \frac{\varepsilon}{2}|u_\varepsilon(\tau_\varepsilon)|^2_{\mathbb{R}^m} + p_\varepsilon(\tau_\varepsilon)^T(Ax_\varepsilon(\tau_\varepsilon) + Bu_\varepsilon(\tau_\varepsilon)) = 0, \end{cases} \tag{52}$$

where

$$\sigma_\varepsilon(s) \in \begin{cases} \text{sgn } s & \text{if } s \leq -\varepsilon \\ \frac{s}{\varepsilon} & \text{if } |s| < \varepsilon. \end{cases} \tag{53}$$

System (52) can readily be treated by a semi-smooth Newton method. In a first step the method of mappings is used to transform the system to a fixed time domain. The transformation $t \rightarrow \frac{t}{\tau}$ transforms (52) to

$$\begin{cases} \dot{x} = \tau(Ax + Bu), & x(0) = x_0, & x(1) = x_1, \\ -\dot{p} = \tau A^T p, \\ u = -\sigma_\varepsilon(B^T p), \\ 1 + \frac{\varepsilon}{2}|u(1)|^2 + p(1)^T(Ax(1) + Bu(1)) = 0. \end{cases} \tag{54}$$

To investigate the semi-smooth Newton method we require an additional assumption

$$|B_i^T p_\varepsilon(1)| \neq \varepsilon, \text{ for all } i = 1, \dots, m, \tag{H3}$$

where we now fix ε and a solution $(x_\varepsilon, u_\varepsilon, \tau_\varepsilon) \in W^{1,2}(0, 1) \times L^2(0, 1) \times \mathbb{R}$ of (P_ε) with associated adjoint $p_\varepsilon \in W^{1,2}(0, 1)$. With (H2) and (H3) holding there exists a neighborhood $\mathcal{U}_{p_\varepsilon}$ of p_ε in $W^{1,2}(0, 1; \mathbb{R}^n)$, $\bar{t} \in (0, 1)$, and a nontrivial interval $(\alpha, \alpha + \delta) \subset (0, 1)$ such that for $p \in \mathcal{U}_{p_\varepsilon}$ we have

$$|B_i^T p(t)| \neq \varepsilon \text{ for all } t \in [\bar{t}, 1], \text{ and } i = 1, \dots, m, \text{ and } |B_i^T p(t)| < \varepsilon \text{ for } t \in (\alpha, \alpha + \delta).$$

Equation (54) suggests to introduce

$$F(x, p, u, \tau) = \begin{pmatrix} \dot{x} - \tau Ax - \tau Bu \\ -\dot{p} - \tau A^T p \\ u + \sigma_\varepsilon(B^T p) \\ x(1) - x_1 \\ 1 + \frac{\varepsilon}{2}|u(1)|^2 + p(1)^T(Ax(1) + Bu(1)) \end{pmatrix}. \tag{55}$$

where

$$F : D_F \subset X \rightarrow L^2(0, 1; \mathbb{R}^n) \times L^2(0, 1; \mathbb{R}^n) \times U \times \mathbb{R}^n \times \mathbb{R},$$

and

$$D_F = W^{1,2}(0, 1) \times \mathcal{U}_{p_\varepsilon} \times U \times \mathbb{R}, \quad X = W^{1,2}(0, 1; \mathbb{R}^n) \times W^{1,2}(0, 1; \mathbb{R}^n) \times U \times \mathbb{R}.$$

Here we have set $U = \{u \in L^2(0, 1; \mathbb{R}^m) : u|_{[\bar{t}, 1]} \in W^{1,2}(\bar{t}, 1; \mathbb{R}^m)\}$ endowed with the norm $|u|_U = (|u|_{L^2(0,1)}^2 + |\dot{u}|_{L^2(\bar{t},1)}^2)^{\frac{1}{2}}$. The only equation that requires special attention in (55) is the third one which contains the operator σ_ε . We use

$$G\sigma_\varepsilon(s) := \begin{cases} \frac{1}{\varepsilon} & \text{if } |s| < \varepsilon \\ 0 & \text{if } |s| \geq \varepsilon \end{cases} \tag{56}$$

as generalized derivative in the sense of Definition 3.1 for σ_ε . It is now straightforward to argue that F is Newton differentiable. To apply Theorem 3.1 it remains to argue that the inverse of the Newton derivative of F is uniformly bounded in a neighborhood of $(x_\varepsilon, u_\varepsilon, \tau_\varepsilon, p_\varepsilon)$. For this purpose the Newton system is considered for reduced unknowns $(p(1), \tau)^T \in \mathbb{R}^{n+1}$. In terms of these variables the system matrix becomes:

$$\mathcal{A} = \begin{pmatrix} A_{11} & A_{12} \\ A_{21} & 0 \end{pmatrix},$$

where

$$A_{11} = \varepsilon^{-1} \tau \int_0^1 e^{\tau A(1-t)} B \chi_I B^T e^{\tau A^T(1-t)} dt \in \mathbb{R}^{n \times n}, \tag{57}$$

$$A_{12} = \varepsilon^{-1} \tau \int_0^1 e^{\tau A(1-t)} B \chi_I B^T \int_t^1 e^{-\tau A^T(t-s)} A^T p(s) ds dt - \int_0^1 e^{\tau A(1-t)} (Ax + Bu) dt \in \mathbb{R}^n, \tag{58}$$

$$A_{21} = (Ax(1) + Bu(1))^T - (p^T(1)B + \varepsilon u^T(1)) G\sigma_\varepsilon(B^T p(1)) B^T \in (\mathbb{R}^n)^T, \tag{59}$$

with $\chi_I = \text{diag}(\chi_{I_1}, \dots, \chi_{I_m})$ and χ_{I_i} the characteristic function of the set

$$I_i = I_i(p) = \{t : |(B^T p)_i| < \varepsilon\}, \quad i = 1, \dots, m$$

which is nonempty for $p \in \mathcal{U}_{p_\varepsilon}$ and $i = \tilde{i}$. The controllability assumption (H1) together with (H2) imply that the symmetric matrix A_{11} is invertible with bounded inverse uniformly with respect to $p \in \mathcal{U}_{p_\varepsilon}$ and τ in compact subsets of $(0, \infty)$.

To guarantee uniform boundedness of the inverse of the Newton derivative we require that the Schur complement $A_{21} A_{11}^{-1} A_{12} \in \mathbb{R}$ of \mathcal{A} for (x, p, u, τ) in a neighborhood of $(x_\varepsilon, p_\varepsilon, u_\varepsilon, \tau_\varepsilon)$ is nontrivial. We therefore assume that

$$\left\{ \begin{array}{l} \text{there exists a bounded neighborhood} \\ \mathcal{U} \subset D_F \subset X \text{ of } (x_\varepsilon, p_\varepsilon, u_\varepsilon, \tau_\varepsilon) \text{ and } c > 0 \text{ such that} \\ |A_{21} A_{11}^{-1} A_{12}| \geq c \text{ for all } (x, p, u, \tau) \in \mathcal{U}. \end{array} \right. \tag{H4}$$

Theorem 6.2. *If (H1)–(H4) hold and $(x_\varepsilon, u_\varepsilon, \tau_\varepsilon)$ denotes a solution to (P_ε) with associated adjoint p_ε , then the semi-smooth Newton algorithm converges superlinearly, provided that the initialization is sufficiently close to $(x_\varepsilon, p_\varepsilon, u_\varepsilon, \tau_\varepsilon)$.*

7. L^1 -data Fitting

Here we treat the data fitting problem with robust $L^1(\Omega)$ fit-to-data term and consider

$$\min_{x \in L^2} \left\{ \mathcal{J}_\alpha(x) \equiv |Kx - y^\delta|_{L^1} + \frac{\alpha}{2}|x|^2 \right\}, \tag{P_{L^1}}$$

where $K : L^2(\Omega) \rightarrow L^2(\Omega)$ is a compact linear operator, and $y^\delta \in L^2$ are measurements corrupted by noise. For every α there exists a unique minimizer x_α . For the value function

$$F(\alpha) = |Kx_\alpha - y^\delta|_{L^1} + \frac{\alpha}{2}|x_\alpha|^2,$$

it can be shown that

$$F'(\alpha) = \frac{1}{2}|x_\alpha|^2, \tag{60}$$

[CJK, IK6]. Fenchel duality theory implies that the dual to (P_{L^1}) is given by

$$\begin{cases} \min_{p \in L^2} \frac{1}{2\alpha} |K^*p|_{L^2}^2 - (p, y^\delta)_{L^2} \\ \text{s.t. } |p|_{L^\infty} \leq 1. \end{cases} \tag{P_{L^1}^*}$$

The dual problem has at least one solution p_α and the relationship between x_α and p_α is given by

$$K^*p_\alpha = \alpha x_\alpha, \quad 0 \leq (Kx_\alpha - y^\delta, p - p_\alpha)_{L^2}, \quad \text{for all } p \in L^2 \text{ with } |p|_{L^\infty} \leq 1. \tag{61}$$

Problem $(P_{L^1}^*)$ does not lend itself to treatment with a superlinearly convergent semi-smooth Newton algorithm. In fact the optimality system for $(P_{L^1}^*)$ is given by

$$\frac{1}{\alpha} KK^*p_\alpha - y^\delta + \lambda_\alpha = 0, \quad (\lambda_\alpha, p - p_\alpha)_{L^2} \leq 0, \quad \text{for all } |p|_{L^\infty} \leq 1, \tag{62}$$

where λ_α denotes the Lagrange multiplier associated to the inequality constraint. This system does not admit a reformulation such that Theorem 3.1(ii) is applicable. We therefore introduce the family of regularized problems

$$\begin{cases} \min_{p \in H^1} \frac{1}{2\alpha} |K^*p|_{L^2}^2 + \frac{\beta}{2} |\nabla p|_{L^2}^2 - (p, y^\delta) \\ \text{s.t. } |p|_{L^\infty} \leq 1, \end{cases} \tag{P_{\beta}^*}$$

for $\beta > 0$, and finally for the numerical realisation the Moreau-Yosida regularization of the box constraints:

$$\min_{p \in H^1} \frac{1}{2\alpha} |K^*p|_{L^2}^2 + \frac{\beta}{2} |\nabla p|_{L^2}^2 - (p, y^\delta) + \frac{1}{2c} |\max(0, c(p-1))|_{L^2}^2 + \frac{1}{2c} |\min(0, c(p+1))|_{L^2}^2, \quad (\mathcal{P}_{\beta,c}^*)$$

for $c > 0$. It is assumed that $\ker K^* \cap \ker \nabla = \emptyset$. Then (\mathcal{P}_β^*) and $(\mathcal{P}_{\beta,c}^*)$ admit unique solutions in H^1 denoted by p_β and p_c respectively. At the end of this section we comment on the choice of the regularization parameters.

The optimality system for $(\mathcal{P}_{\beta,c}^*)$ is given by

$$\begin{cases} \frac{1}{\alpha} K K^* p_c - \beta \Delta p_c + \lambda_c = y^\delta, \\ \lambda_c = \max(0, c(p_c - 1)) + \min(0, c(p_c + 1)), \end{cases} \quad (63)$$

where $\lambda_c \in H^1(\Omega)$. It can be shown by techniques which are by now quite standard [IK5, CJK] that for each fixed $\beta > 0$ we have

$$(p_c, \lambda_c) \rightarrow (p_\beta, \lambda_\beta) \text{ in } H^1(\Omega) \times H^1(\Omega)^*,$$

where $\lambda_\beta \in H^1(\Omega)^*$ is the Lagrange multiplier associated to the inequality constraint in (\mathcal{P}_β^*) . Moreover, for every sequence $\beta_n \rightarrow 0$ there exists a subsequence such that $p_{\beta_k} \rightharpoonup p_\alpha$ in $L^2(\Omega)$, where p_α is a solution of $(\mathcal{P}_{L^1}^*)$. Analogously, if c is fixed then the solutions to $(\mathcal{P}_{\beta,c}^*)$, now denoted by $p_{\beta_n,c}$ converge to a solution of $p_{0,c}$ of $(\mathcal{P}_{\beta,c}^*)$ with $\beta = 0$.

To solve the optimality system (63) for the regularized problem we consider the nonlinear operator equation $F(p) = 0$ for $F : H^1(\Omega) \rightarrow H^1(\Omega)^*$, where

$$F(p) := \frac{1}{\alpha} K K^* p - \beta \Delta p + \max(0, c(p-1)) + \min(0, c(p+1)) - y^\delta. \quad (64)$$

In view of Section 3 we use as Newton derivative for the projection operator $P(p) := \max(0, (p-1)) + \min(0, (p+1))$ the mapping

$$G_P(p)h := h\chi_{\{|p|>1\}} = \begin{cases} h(x) & \text{if } |p(x)| > 1, \\ 0 & \text{if } |p(x)| \leq 1. \end{cases}$$

It can readily be verified that the update $p^{k+1} \in H^1(\Omega)$ of the Newton equation $G_P(p^k)(p^{k+1} - p^k) = -F(p^k)$ is the solution to the equation

$$\frac{1}{\alpha} K K^* p^{k+1} - \beta \Delta p^{k+1} + c\chi_{\mathcal{A}_k} p^{k+1} = y^\delta + c(\chi_{\mathcal{A}_k^+} - \chi_{\mathcal{A}_k^-}), \quad (65)$$

where the active sets are given by

$$\mathcal{A}_k^+ := \{x\} p^k(x) > 1, \quad \mathcal{A}_k^- := \{x\} p^k(x) < -1, \quad \mathcal{A}_k := \mathcal{A}_k^+ \cup \mathcal{A}_k^-.$$

Moreover we can use the techniques of Section 3 to establish the following result.

Theorem 7.1. *If $|p_c - p^0|_{H^1}$ is sufficiently small, then the iterates p^k of the semi-smooth Newton algorithm converge superlinearly in $H^1(\Omega)$ to the solution p_c of $(\mathcal{P}_{\beta,c}^*)$ as $k \rightarrow \infty$.*

We turn to a discussion of the choice of the parameters α, β and c in problem $(\mathcal{P}_{\beta,c}^*)$. Clearly β and c , which are used in the inner loop of an iterative procedure, should be taken close to 0 and ∞ , respectively. The choice of α , which is different from 0 in general, is the most delicate one and we turn to it first.

Choice of α by model function approach: The model function approach proposed in [IK6] approximates the value function $F(\alpha)$ by rational polynomials. Here we consider a model function of the form

$$m(\alpha) = b + \frac{d}{t + \alpha}. \tag{66}$$

Noting that $x_\alpha \rightarrow 0$ for $\alpha \rightarrow \infty$ and $\alpha|x_\alpha|^2 \rightarrow 0$ by (61), we fix $b = |y^\delta|_{L^1}$. The parameters d and t are determined by interpolation conditions according to

$$m(\alpha) = F(\alpha), \quad m'(\alpha) = F'(\alpha), \tag{67}$$

which together with the definition of $m(\alpha)$ gives

$$b + \frac{d}{t + \alpha} = F(\alpha), \quad -\frac{d}{(t + \alpha)^2} = F'(\alpha). \tag{68}$$

We recall from (60) that $F'(\alpha) = -\frac{1}{2}|x_\alpha|_{L^2}^2$, and this expression can be calculated without any extra computational effort. Note that $F(\alpha)$, just like $m(\alpha)$, is monotonically increasing. In case the L^1 fit-to-data term is replaced by an L^2 term, then $F''(\alpha) = -(x_\alpha, (\alpha I + K^*K)^{-1}x_\alpha) \leq 0$. In particular, in this case, F is concave, just as m . One of the important features of our approach lies in not requiring knowledge of the noise level. The rationale for noise level estimation is that $F(0)$ represents a lower bound on the noise level and consequently, if m approximates well F , then $m(0)$ can be taken as an approximation of the noise level.

To analyse the sequence $\{\alpha_k\}$ determined by Algorithm (2) one can argue [CJK] that if this sequence converges then its limit α^* satisfies

$$(\sigma - 1)\phi(\alpha^*) - \alpha^*F'(\alpha^*) = 0, \tag{69}$$

where $\varphi(\alpha) = |Kx_\alpha - x^\delta|_{L^1}$. The intuitive interpretation of the iteration is clear: it balances the weighted data-fitting term $(\sigma - 1)\phi(\alpha) = (\sigma - 1)|Kx_\alpha - y^\delta|_{L^1}$ and the penalty term $\alpha F'(\alpha) = \frac{\alpha}{2}|x_\alpha|_{L^2}^2$. The scalar σ controls the relative weighting between the two terms.

Algorithm 2 Fixed-point algorithm for adaptively determining α

- 1: Set $k = 0$, choose $\alpha_0 > 0$, $b \geq |y|_{L^1}$ and $\sigma > 1$
- 2: **repeat**
- 3: Compute x_{α_k} by a path-following semismooth Newton method
- 4: Compute $F(\alpha_k)$ and $F'(\alpha_k)$
- 5: Construct the model function $m_k(\alpha) = b + \frac{d_k}{t_k + \alpha}$ by solving the interpolation condition at α_k

$$d_k = -\frac{(b - F(\alpha_k))^2}{F'(\alpha_k)}, \quad t_k = \frac{b - F(\alpha_k)}{F'(\alpha_k)} - \alpha_k.$$

- 6: Calculate the m -intercept \hat{m} of the tangent of $m_k(\alpha)$ at $(\alpha_k, F(\alpha_k))$ by

$$\hat{m} = F(\alpha_k) - \alpha_k F'(\alpha_k),$$

- 7: Solve for α_{k+1} by setting $m_k(\alpha_{k+1}) = \sigma \hat{m}$, i.e. $\alpha_{k+1} = \frac{c_k}{\sigma - b} - t_k$
 - 8: Set $k = k + 1$
 - 9: **until** the stopping criterion is satisfied.
-

From [CJK] we now quote the following result.

Theorem 7.2. (a) If σ is sufficiently close to 1 and $y^\delta \neq 0$, then (69) has at least one solution. (b) If in addition $\alpha_0 F'(\alpha_0) - (\sigma - 1)\varphi(\alpha_0) > 0$, then the iterates $\{\alpha_k\}$ converge monotonically from above to a solution of (69).

Choice of β within a path-following semismooth Newton method: The introduction of the H^1 smoothing alters the structure of the problem and therefore the value of β should be as small as possible. However, the regularized dual problem $(\mathcal{P}_{\beta,c}^*)$ becomes increasingly ill-conditioned as β decreases to zero due to the ill-conditioning of discretized KK^* and rank-deficiency of the diagonal matrix corresponding to the (discrete) active set, see (65). Therefore, the respective system matrix will eventually become numerically singular for vanishing β .

One remedy is a continuation strategy: Starting with a large β , e.g. $\beta_0=1$, we reduce its value, e.g. geometrically, as long as the system is still solvable, and take the solution corresponding to the smallest such value. The question remains how to automatically select the stopping index without a priori knowledge or expensive computations for estimating the condition number or smallest singular value by e.g. singular value decomposition. To select an appropriate stopping index, we exploit the structure of the (infinite-dimensional) bound constraint problem: the correct solution should be nearly feasible for c sufficiently large, i.e. $\|p\|_{L^\infty} \leq \tau$ for some $\tau \approx 1$. Recall that for the linear system (65), the right hand side f satisfies $\|f\|_{L^\infty} \approx c \gg 1$, which should be balanced by the diagonal matrix $c\chi_{\mathcal{A}}$ in order to verify the feasibility condition. If the matrix is nearly singular, this will no longer be the case, and the solution p blows up and

violates the feasibility condition, i.e. $\|p\|_{L^\infty} \gg 1$. Once this happens, we take the last iterate which is still (close to) feasible and return it as the solution. This procedure provides an efficient and simple strategy to achieve the conflicting goals of minimizing the effect on the primal problem and maintaining the numerical stability of the dual problem $(\mathcal{P}_{\beta,c}^*)$ for sufficient accuracy.

For the choice of c it appears to be worthwhile to also investigate path-following techniques as introduced in [HK2] but this remains to be done in future work.

8. Mathematical Programming

In this section we discuss a nonsmooth mathematical programming problem, which only in part relies on convexity assumptions. Let X be a Banach space, Y a Hilbert space and Z a Hilbert lattice with an ordering induced by a cone K with vertex at 0, i.e. $x \leq y$ if $x - y \in K$. Consider the minimization

$$\min F(y) \quad \text{subject to } G_1(y) = 0, \quad G_2(y) \leq 0, \quad y \in \mathcal{C}, \tag{70}$$

where $G_1 : X \rightarrow Y$ is C^1 , $G_2 : X \rightarrow Z$ is convex, and $\mathcal{C} \subset X$ is a closed convex set. We assume that $F = F_0(y) + F_1(y)$ where F_0 is C^1 and $F_1(y)$ is convex. Then we have the following necessary optimality condition.

Theorem 8.1. *Let $y^* \in \mathcal{C}$ is a minimizer of (70). Then there exists a nontrivial $(\lambda_0, \mu_1, \mu_2) \in \mathbb{R}^+ \times Y^* \times Z^*$ such that*

$$\begin{aligned} \lambda_0 (F'_0(y - y^*) + F_1(y) - F_1(y^*)) + \langle \mu_1, G'_1(y^*)(y - y^*) \rangle + \langle \mu_2, G_2(y) - G_2(y^*) \rangle \geq 0 \\ \mu_2 \geq 0, \quad \langle \mu_2, G_2(y^*) \rangle = 0, \quad \text{for all admissible } y \in \mathcal{C}. \end{aligned} \tag{71}$$

Proof. For $\epsilon > 0$ define the functional

$$J_\epsilon(u, \tau) = (((F(y) - F(y^*) + \epsilon)^+)^2 + |G_1(y)|_Y^2 + |\max(0, G_2(y))|_Z^2)^{\frac{1}{2}}.$$

Then, $J_\epsilon(y^*) = \epsilon$ and $J_\epsilon(y^*) \leq \inf J_\epsilon + \epsilon$. For any $y \in \mathcal{C}$ define the metric

$$d(y, y^*) = |y - y^*|_X.$$

By the Ekeland variational principle there exists a y^ϵ such that

$$\begin{aligned} J_\epsilon(y^\epsilon) &\leq J_\epsilon(y^*) \\ J_\epsilon(y) - J_\epsilon(y^\epsilon) &\geq -\sqrt{\epsilon} d(y, y^\epsilon) \quad \text{for all } y \in \mathcal{C} \\ d(y^\epsilon, y^*) &\leq \sqrt{\epsilon}. \end{aligned} \tag{72}$$

Let

$$\mu_1^\epsilon = 2 G_1(y^\epsilon), \quad \mu_2^\epsilon = 2 \max(0, G_2(y^\epsilon)).$$

Letting $y = y_t = y^\epsilon + t(\hat{y} - y^\epsilon)$, $t \in (0, 1)$ with $\hat{y} \in \mathcal{C}$ in (72), we have

$$\begin{aligned}
 -\sqrt{\epsilon}d(y_t, y^\epsilon) \leq J_\epsilon(y_t) - J_\epsilon(y^\epsilon) &\leq \frac{1}{J_\epsilon(y^\epsilon) + J_\epsilon(y_t)} (\alpha^{\epsilon,t} (F(y_t) - F(y^\epsilon))) \\
 + \langle \mu_1^\epsilon, G_1'(y^\epsilon)(y_t - y^\epsilon) \rangle + \langle \mu_2^\epsilon, G_2(y_t) - G_2(y^\epsilon) \rangle + |\mu^\epsilon|o(|y_t - y^\epsilon|),
 \end{aligned} \tag{73}$$

where

$$\alpha^{\epsilon,t} = ((F(y_t) - F(y^*) + \epsilon)^+ + (F(y^\epsilon) - F(y^*) + \epsilon)^+).$$

and we used for $t > 0$ sufficiently small

$$(F(y_t) - F(y^*) + \epsilon)(F(y^\epsilon) - F(y^*) + \epsilon) \geq 0.$$

Since F_1 and G_2 are convex,

$$F(y_t) - F(y^\epsilon) \leq tF_0'(y^\epsilon)(\hat{y} - y^\epsilon) + t(F_2(\hat{y}) - F_2(y^\epsilon)) + o(|y_t - y^\epsilon|)$$

$$G_2(y_t) - G_2(y^\epsilon) \leq t(G_2(\hat{y}) - G_2(y^\epsilon)).$$

Let

$$\tilde{\mu}^{t,\epsilon} = \frac{\mu^\epsilon}{J_\epsilon(y_t) + J_\epsilon(y^\epsilon)}, \quad \tilde{\alpha}^{\epsilon,t} = \frac{\alpha^{\epsilon,t}}{J_\epsilon(y_t) + J_\epsilon(y^\epsilon)}.$$

Since $(\tilde{\alpha}^{\epsilon,t}, \tilde{\mu}^{\epsilon,t})$ is bounded, there exists a subsequence such that $\tilde{\mu}^{\epsilon,t} \rightarrow \mu \in (Y \times Z)^*$ (weakly star) and $\tilde{\alpha}^{\epsilon,t} \rightarrow \lambda^0 \geq 0$ as $\epsilon \rightarrow 0^+$, $t \rightarrow 0^+$. Dividing (73) by t and letting $t \rightarrow 0^+$ and subsequently $\epsilon \rightarrow 0^+$, we obtain

$$\lambda_0 (F_0'(\hat{y} - y^*) + F_1(\hat{y}) - F_1(y^*)) + \langle \mu_1, G_1'(y^*)(\hat{y} - y^*) \rangle + \langle \mu_2, G(\hat{y}) - G(y^*) \rangle \geq 0,$$

for all $\hat{y} \in \mathcal{C}$. Since $\tilde{\mu}_2^\epsilon \geq 0$ and $\langle \tilde{\mu}_2^\epsilon, G_2(y^\epsilon) \rangle \geq 0$, it follows that $\mu_2 \geq 0$ and $\langle \mu_2, G_2(y^*) \rangle \geq 0$ and since $G_2(y^*) \leq 0$, thus $\langle \mu_2, G_2(y^*) \rangle = 0$. \square

Corollary 8.1. *Assume there exists a nontrivial $(\lambda_0, \mu_1, \mu_2) \in \mathbb{R}^+ \times Y^* \times Z^*$ such that (71) holds. If the regular point condition:*

$$0 \in \text{int} \left\{ \begin{array}{l} G_1'(y^*)(\mathcal{C} - y^*) \\ G_2(y) - G_2(y^*) - K + G_2(y^*) \end{array} \right\}. \tag{74}$$

is satisfied at y^ , then one can take $\lambda_0 = 1$.*

Proof. As a consequence of the regular point condition, there exists for all $(\tilde{\mu}_1, \tilde{\mu}_2)$ belonging to a neighborhood of 0 in $Y \times Z$, elements $y \in \mathcal{C}$, $k \in K$ such that

$$(\tilde{\mu}_1, \tilde{\mu}_2) = (G_1'(y^*)(y - y^*), G_2(y) - G_2(y^*) - k + G_2(y^*)).$$

Consequently $\langle \mu_1, \tilde{\mu}_1 \rangle + \langle \mu_2, \tilde{\mu}_2 \rangle = \langle \mu_1, G_1'(y^*)(y - y^*) \rangle + \langle \mu_2, G_2(y) - G_2(y^*) - k + G_2(y^*) \rangle$. Note that $\langle \mu_2, k - G_2(y^*) \rangle = \langle \mu_2, k \rangle \leq 0$. If $\lambda_0 = 0$ then the first equation in (71) implies that $\langle \mu_1, \tilde{\mu}_1 \rangle + \langle \mu_2, \tilde{\mu}_2 \rangle \geq 0$ for all $(\tilde{\mu}_1, \tilde{\mu}_2)$ in a neighborhood of 0 and thus $\mu_1 = \mu_2 = 0$, which is a contradiction. That is, $\lambda_0 \neq 0$ and thus the problem is strictly normal and one can set $\lambda^0 = 1$. \square

L¹-minimum norm control: Consider the optimal exit problem with minimum *L¹* norm

$$\begin{cases} \min_{u,\tau} \int_0^\tau (f(x(t)) + \delta|u(t)|) dt & \text{subject to} \\ \frac{d}{dt}x = b(x(t), u(t)), \quad x(0) = x, \\ g(x(\tau)) = 0, \quad |u(t)|_{\mathbb{R}^m} \leq \gamma \text{ for a.e. } t, \end{cases} \tag{75}$$

where $\delta > 0$, $f : \mathbb{R}^n \rightarrow \mathbb{R}$, $b : \mathbb{R}^n \times \mathbb{R}^m \rightarrow \mathbb{R}^n$, $g : \mathbb{R}^n \rightarrow \mathbb{R}^k$ are smooth functions. We have two motivations to consider (75). In the context of sparse controls, compare Section 5, the pointwise norm constraints, allow us avoid controls in measure space. In the context of time optimal controls the term $\delta \int_0^\tau |u| dt$ can be considered as regularisation term. We shall see from the optimality condition (82) below that this determines the control as a function of the adjoint by mean of an equation rather than an inclusion as in (47), where no such regularisation was used.

One can transform (75) to the fixed interval $s \in [0, 1]$ via the change of variable $t = \tau s$

$$\begin{cases} \min_{u,\tau} \int_0^1 \tau (f(x(t)) + \delta|u(t)|) dt & \text{subject to} \\ \frac{d}{dt}x = \tau b(x(t), u(t)), \quad x(0) = x, \\ g(x(1)) = 0, \quad u \in U_{ad} = \{u \in L^\infty(0, 1; \mathbb{R}^m) : |u(t)| \leq \gamma\}. \end{cases} \tag{76}$$

Let $y = (u, \tau)$ and define

$$\begin{aligned} F_0(y) &= \tau \int_0^1 f(x(t)) dt, & F_1(u) &= \delta \int_0^1 |u(t)| dt, \\ F(y) &= F_0(y) + \tau F_1(u), & G(y) &= g(x(1)), \end{aligned}$$

where $x = x(\cdot; u, \tau)$ is the solution to the initial value problem in (76), given $u \in U_{ad}$ and $\tau \geq 0$. Then the control problem can equivalently formulated as

$$\min_{(u,\tau) \in U_{ad} \times \mathbb{R}^+} F(y) \quad \text{subject to } G(y) \in K. \tag{77}$$

Assume that $y^* = (u^*, \tau^*)$ is an optimal solution to (77) and suppose that the regular point condition

$$0 \in \text{int} \{G_u(y^*)(v - u^*) + G_\tau(\tau - \tau^*) : v \in U_{ad}, \tau > 0\} \tag{78}$$

holds. Since $\tau_1 F_1(u_1) - \tau_2 F_1(u_2) = (\tau_1 - \tau_2) F_1(u_1) + \tau_2 (F_1(u_1) - F_1(u_2))$, it is easy to modify the proof of Theorem 8.1 to obtain the necessary optimality: there exist a Lagrange multiplier $\mu \in \mathbb{R}^k$ such that

$$\begin{aligned} &\tau^* (F_1(u) - F_1(u^*)) + (\tau - \tau^*) F_1(u^*) \\ &+ ((F_0)_u + G_u^* \mu)(u - u^*) + ((F_0)_\tau + G_\tau^* \mu)(\tau - \tau^*) \geq 0 \end{aligned} \tag{79}$$

for all $u \in U_{ad}$ and $\tau \geq 0$, where $F_0 = F_0(y^*), G_0 = G_0(y^*)$. Note that for $v \in L^\infty(0, 1; \mathbb{R}^m)$

$$G_u(v) = (g_x(x^*(1)), h(1))_{\mathbb{R}^n}, \quad G_\tau = (g_x(x^*(1)), \xi(1))_{\mathbb{R}^n},$$

$$(F_0)_u(v) = \tau \int_0^1 (f'(x(t)), h(t))_{\mathbb{R}^n} dt, \quad (F_0)_\tau(v) = \int_0^1 ((\tau f'(x(t)), \xi(t))_{\mathbb{R}^n} + f(x(t))) dt,$$

where (h, ξ) satisfies

$$\begin{aligned} \frac{d}{dt}h(t) &= \tau^* (b_x(x^*(t), u^*(t))h(t) + b_u(x^*(t), u^*(t))v(t)), \quad h(0) = 0 \\ \frac{d}{dt}\xi(t) &= \tau^* b_x(x^*(t), u^*(t))\xi(t) + b(x^*(t), u^*(t)), \quad \xi(0) = 0. \end{aligned} \tag{80}$$

Let $p \in H^1(0, 1; \mathbb{R}^n)$ satisfy the adjoint equation

$$-\frac{d}{dt}p(t) = \tau^* (b_x(x^*(t), u^*(t))^t p(t) + f_x(x^*(t))), \quad p(1) = \mu g_x(x^*(1)), \tag{81}$$

then

$$\begin{aligned} (h(1), p(1))_{\mathbb{R}^n} &= \tau^* \int_0^1 (b_u(x^*(t), u^*(t))v(t) - (f'(x^*(t)), h(t))) dt \\ (\xi(1), p(1))_{\mathbb{R}^n} &= \int_0^1 (b(x^*(t), u^*(t)), p(t))_{\mathbb{R}^n} dt. \end{aligned}$$

From (79) therefore for all $u \in U_{ad}$ and $\tau \geq 0$

$$\begin{aligned} &(\tau - \tau^*) \int_0^1 (f(x^*(t)) + \delta|u^*(t)| + (b(x^*(t), u^*(t)), p(t))) dt \\ &+ \int_0^1 (b_u(x^*(t), u^*(t))^t p(t), u(t) - u^*(t)) + \delta|u(t)| - \delta|u^*(t)| dt \geq 0. \end{aligned}$$

Hence we obtain the optimality condition

$$u^*(t) = \begin{cases} 0 & \text{if } |b_u(x^*(t), u^*(t))^t p(t)| \leq \delta \\ -\gamma \frac{b_u(x^*(t), u^*(t))^t p(t)}{|b_u(x^*(t), u^*(t))^t p(t)|} & \text{if } |b_u(x^*(t), u^*(t))^t p(t)| \geq \delta, \end{cases} \tag{82}$$

and

$$\int_0^1 (f(x^*(t)) + \delta|u^*(t)| + (b(x^*(t), u^*(t)), p(t))_{\mathbb{R}^n}) dt = 0.$$

This, together with the fact that the Hamiltonian \mathcal{H} is constant along (u^*, x^*, p) , implies the transversality condition

$$\mathcal{H}(t) := f(x^*(t)) + \delta|u^*(t)| + (b(x^*(t), u^*(t)), p(t))_{\mathbb{R}^n} = 0 \text{ on } [0, 1]. \tag{83}$$

Acknowledgement

Many colleagues have joined us over the last years in the development of semi-smooth Newton methods in function spaces for a wide range of different application. We appreciate specifically the collaboration with M. Hintermüller, M. Bergounioux, C. Clason, C. DeLosReyes, R. Herzog, B. Jin, V. Kovtunenکو, G. Stadler, and B. Vexler.

References

- [Be] D.P. Bertsekas, *Constraint Optimization and Lagrange Multiplier Methods*, Academic Press, Paris, 1982.
- [CJK] C. Clason, B. Jin and K. Kunisch: A semismooth Newton method for L^1 data fitting with automatic choice of regularization parameters and noise calibration, to appear in *SIAM Journal on Imaging Sciences*.
- [CK] C. Clason and K. Kunisch: A duality-based approach to elliptic control problems in non-reflexive Banach spaces, *ESAIM: COCV*, to appear.
- [CNQ] X. Chen, Z. Nashed and L. Qi: Smoothing methods and semi-smooth methods for nondifferentiable operator equations, *SIAM J. on Numerical Analysis*, 38(2000), 1200–1216.
- [ET] I. Ekeland and R. Temam, *Convex Analysis and Variational Problems*, North Holland, Amsterdam, 1976.
- [ETu] I. Ekeland and T. Turnbull, *Infinite Dimensional Optimization and Convexity*, The University of Chicago Press, Chicago, 1983.
- [GLT] R. Glowinski, J.L. Lions and R. Tremolieres, *Numerical Analysis of Variational Inequalities*, North Holland, Amsterdam (1981).
- [HIK] M. Hintermüller, K. Ito and K. Kunisch: The primal–dual active set strategy as a semi–smooth Newton method, *SIAM Journal on Optimization*, 13(2002), 865–888. doi:10.1137/S1052623401383558
- [HK1] M. Hintermüller and K. Kunisch: Total bounded variation regularization as bilaterally constrained optimization problem, *SIAM J. Appl. Mathematics*, 64(2004), 1311–1333. doi:10.1137/S0036139903422784
- [HK2] M. Hintermüller and K. Kunisch: Feasible and non-interior path-following in constrained minimization with low multiplier regularity, *SIAM J. Control and Optim.*, 45(2006), 1198–1221. doi:10.1137/050637480
- [HK3] M. Hintermüller and K. Kunisch: PDE-Constrained optimization subject to pointwise control and zero- or first-order state constraints, *SIAM Journal on Optimization*, 20 (2009), 1133–1156. doi:10.1137/080737265
- [HL] H. Hermes and J. LaSalle: *Functional Analysis and Time Optimal Control*, Academic Presse, 1969.
- [HU] M. Hintermüller and M. Ulbrich: Semismooth Newton methods for operator equations in function spaces. *SIAM J. Optim.* 13(2002), 805–842

- [IK1] K. Ito and K. Kunisch: *On the Lagrange Multiplier Approach to Variational Problems and Applications*, SIAM, Philadelphia, 2008.
- [IK2] K. Ito and K. Kunisch: Augmented Lagrangian methods for nonsmooth convex optimization in Hilbert spaces, *Nonlinear Analysis, Theory, Methods and Applications*, 41(2000), 591–616. doi:10.1016/S0362-546X(98)00299-5
- [IK3] K. Ito and K. Kunisch: The primal-dual active set method for nonlinear optimal control problems with bilateral constraints, *SIAM J. on Control and Optimization*, 43(2004), 357–376. doi:10.1137/S0363012902411015
- [IK4] K. Ito and K. Kunisch: Semi-smooth Newton Methods for time-optimal control for a class of ODEs, to appear in *SIAM Journal on Control and Optimization*.
- [IK5] K. Ito and K. Kunisch: Semi-smooth Newton methods for variational inequalities of the first kind, *Mathematical Modelling and Numerical Analysis, R.A.R.I.O, Mathematical Modelling and Numerical Analysis*, 37(2002), 41–62.
- [IK6] K. Ito and K. Kunisch: On the choice of the regularization parameter in nonlinear inverse problems, *SIAM J. on Optimization*, 2(1992), 1–29. doi:10.1137/0802019
- [K] B. Kummer: Generalized Newton and NCP methods: convergence, regularity, actions, *Discuss. Math. Differ. Incl. Control Optim.*, 20, (2000), 209–244.
- [KV] K. Kunisch and B. Vexler: Constrained Dirichlet boundary control in L^2 for a class of evolution equations, *SIAM J. Control and Optimization*, 46 (2007), 1726–1753. doi:10.1137/060670110
- [St] G. Stadler: Elliptic optimal control problems with L^1 -control cost and applications for the placement of control devices, *Computational Optimization and Applications*, 2007. doi = 10.1007/s10589-007-9150-9
- [U] M. Ulbrich: Semi-smooth Newton methods for operator equations in function spaces, *SIAM J. on Optimization*, 13(2003), 805–841.

Modelling Aspects of Tumour Metabolism

Philip K. Maini*, Robert A. Gatenby†, and Kieran Smallbone‡

Abstract

We use a range of mathematical modelling techniques to explore the acid-mediated tumour invasion hypothesis. The models make a number of predictions which are experimentally verified. The therapeutic implications, namely either buffering acid or manipulating the phenotypic selection process, are described.

Mathematics Subject Classification (2010). 92C50

Keywords. Carcinogenesis – Glycolytic phenotype – Mathematical modelling

1. Biological Background

Cancer cell populations are extremely heterogeneous, displaying a wide range of genotypic and phenotypic differences [7]. For example, studies of clinical breast cancers have shown that every cell line exhibited a novel genotype, meaning no prototypic cancer cell can be defined. It is likely that several of the lethal phenotypic traits of cancer are not the direct result of genetic changes, but rather arise from the unique physiological environments of tumours.

The tumour microenvironment is significantly different from that of normal tissue. Marked fluctuations can be seen in glucose, lactate, acidic pH and

*Philip K. Maini, Centre for Mathematical Biology, Mathematical Institute, 24-29 St Giles', Oxford, OX1 3LB, UK and Oxford Centre for Integrative Systems Biology, Department of Biochemistry, South Parks Road, Oxford OX1 3QU. E-mail: maini@maths.ox.ac.uk.

†Robert A. Gatenby, Moffitt Cancer Center, 12902 Magnolia Drive, Tampa, FL 33612, USA. E-mail: robert.gatenby@moffitt.org.

‡Kieran Smallbone, Manchester Centre for Integrative Systems Biology, Manchester Interdisciplinary Biocentre, 131 Princess Street, Manchester, M1 7DN, UK.
E-mail: kieran.smallbone@manchester.ac.uk.

PKM was partially supported by a Royal Society-Wolfson Merit award. PKM and RAG were partially supported by NIH grant 1U54CA143970-01. KS acknowledges the support of the BBSRC/EPSRC grant BB/C008219/1 "The Manchester Centre for Integrative Systems Biology (MCISB)".

oxygen tensions. These variations have their roots both in poor perfusion and metabolic changes. The chaotic vasculature of tumours creates an unbalanced blood supply. As a consequence, many regions within tumours are found to be transiently or chronically hypoxic (poorly oxygenated). Cells respond to periods of hypoxia by converting to anaerobic metabolism, or glycolysis, which in turn produces lactic acid and brings about lower tissue pH. However, the pioneering work of Warburg [22] showed that tumour acidification can occur independently of hypoxia. The increased reliance on glycolysis to produce energy in many aggressive tumours occurs even in the presence of sufficient oxygen. Thus acidification is an intrinsic property of both poor vasculature and altered tumour cell metabolism.

The constitutive adoption of increased aerobic glycolysis is known as the glycolytic phenotype. The inefficiency of this anaerobic metabolism is compensated for through a several-fold increase in cellular glucose consumption. This phenomenon is now routinely exploited for tumour imaging through fluorodeoxyglucose positron emission tomography (FDG-PET). PET has confirmed that the vast majority (> 90%) of human primary and metastatic tumours demonstrate increased glucose uptake indicating abnormal metabolism. Furthermore, PET has been used to show a direct correlation between tumour aggressiveness and the rate of glucose consumption [4].

The presence of the glycolytic phenotype in the malignant phenotype of such a wide range of cancers seems inconsistent with an evolutionary model of carcinogenesis. Due to the Darwinian dynamics at play, it is reasonable to assume the common appearance of a specific phenotype within a large number of different cancer populations is evidence that it must confer a significant growth advantage. However, the proliferative advantages gained from altered glucose metabolism are far from clear. Firstly, anaerobic metabolism is more than an order of magnitude less efficient than its aerobic counterpart, producing only 2 ATP molecules per glucose molecule in comparison to approximately 36 ATP molecules. Secondly, the hydrogen ions produced as a result of glycolysis cause a consistent acidification of the extracellular space that is toxic [17]. Intuitively, one would expect the Darwinian forces prevailing during carcinogenesis to select against this inefficient and environmentally toxic phenotype, in favour of more optimal metabolic regimes.

Gatenby and Gillies [11] propose that evolution of aerobic glycolysis is the result of environmental constraints imposed by the morphology of the ducts in which premalignant lesions evolve. Initially, normal epithelial cells grow along a basement membrane, with the epithelial layer at most a few cells thick. Homeostasis mechanisms do not normally allow growth of these cells away from the basement membrane. However, following initial genetic events in the carcinogenesis pathways such as those depicted by the Fearon-Vogelstein model [6], the cells become hyperplastic, leading to a thickening of the epithelial layer, pushing cells into the lumen and away from the membrane. Since the blood vessels remain outside the basement membrane, nutrients and waste must diffuse over

longer and longer distances. As a result, it is likely that hyperplastic cells beyond the Thomlinson–Gray limit of 100–150 μm [21] from the basement membrane will experience profound hypoxia, which will initiate a sequence of critical cellular adaptations and environmental changes. Specifically, it is proposed that hypoxia leads to constitutive upregulation of glycolysis which, in turn, results in increased H^+ production and acidification of the microenvironment. This decreased extracellular pH (pH_X) is toxic to the local populations, in turn selecting for cells that are resistant to acid-induced toxicity. Acidosis also selects for motile cells that eventually breach the basement membrane, gaining access to existing and newly formed blood and lymphatic routes for metastasis.

Gatenby and Gawlinski [10] point out that the tumour phenotype that emerges from the sequence above, constitutively increasing acid production and becoming resistant to acid-induced toxicity, has a powerful growth advantage over its normal counterparts. They propose that acidity may play a key role in mediating tumour invasion. The key idea is that the transformed tumour metabolism with increased use of glycolysis and acid secretion alters the microenvironment by substantially reducing tumour extracellular pH, usually by more than 0.5 pH units. The H^+ ions produced by the tumour then diffuse along concentration gradients into the adjacent normal tissue. This acidification leads to death of normal cells; tumour cells, however, are relatively resistant to acidic pH_X . Whilst normal cells die in environments with a persistent pH below about 7, tumour cells typically exhibit a maximum proliferation rate in a relatively acidic medium (pH 6.8) [3]. As a result, the tumour edge can be seen as forming a travelling wave progressing into normal tissue, preceded by another travelling wave of increased microenvironmental acidity.

2. Continuum Modelling Approaches

Population ecology methods provide a means for examining tumours, not as an isolated collection of transformed cells, but rather as an invading species in a previously stable multicellular population. Gatenby and Gawlinski [10] model the tumour-host interface as a network of interacting normal and malignant cell populations, using coupled, non-linear differential equations. The interactions are then explored to define the crucial parameters that control tumourigenesis and to demonstrate the limitations of traditional therapeutic strategies.

Tumour cell populations, as with any invading population in biology, must directly perturb their environment in such a way as to facilitate their own growth while inhibiting the growth of the original community. The commonality of altered tumour metabolism, in particular the adoption of the glycolytic phenotype in most cancers, led Gatenby and Gawlinski to propose the acid-mediated tumour invasion hypothesis outlined above. The authors propose that tumour cells' increased acid secretion, coupled with their resistance to low extracellular pH, may provide a simple but complete mechanism for cancer invasion.

The hypothesis is modelled as a system of three coupled partial differential equations (PDEs), determining the spatio-temporal distribution of three fields: the normal tissue density N_1 , the tumour tissue density N_2 , and the concentration of excess hydrogen ions L . The model includes: (1) logistic cellular growth; (2) normal cell death due to exposure to acid; (3) acid production by tumour cells; (4) acid reabsorption and buffering; and (5) spatial diffusion of acid and cells. It takes the form

$$\frac{\partial N_1}{\partial t} = r_1 N_1 \left(1 - \frac{N_1}{K_1}\right) - d_1 L N_1, \quad (1)$$

$$\frac{\partial N_2}{\partial t} = r_2 N_2 \left(1 - \frac{N_2}{K_2}\right) + D_2 \nabla \cdot \left[\left(1 - \frac{N_1}{K_1}\right) \nabla N_2 \right], \quad (2)$$

$$\frac{\partial L}{\partial t} = r_3 N_2 - d_3 L + D_3 \nabla^2 L, \quad (3)$$

where r_1 and r_2 are the growth rates of the normal and tumour cell populations, respectively, K_1 and K_2 their carrying capacities, D_2 scales the diffusion coefficient for tumour cells, d_1 is the normal cell susceptibility to acid, r_3 the rate of hydrogen ion production by tumour cells, d_3 the combined rate of acid removal by blood vessels and buffering, and D_3 the diffusion coefficient for hydrogen ions in tissue. Notice that there is no normal cell diffusion within the model, in recognition of the fact that healthy tissue is well-regulated and participating normally in an organ. Notice also that the tumour diffusion coefficient is constructed such that when normal tissue is at its carrying capacity, the diffusion coefficient for tumour tissue is zero and the tumour is confined. This final assumption is at the heart of the model: tumour tissue is unable to spread without first diminishing the surrounding healthy tissue from its carrying capacity.

In non-dimensional form, Eqs. (1)–(3) become

$$\frac{\partial \eta_1}{\partial \tau} = \eta_1(1 - \eta_1) - \delta_1 \Lambda \eta_1, \quad (4)$$

$$\frac{\partial \eta_2}{\partial \tau} = \rho_2 \eta_2(1 - \eta_2) + \Delta_2 \nabla_\xi \cdot [(1 - \eta_1) \nabla_\xi \eta_2], \quad (5)$$

$$\frac{\partial \Lambda}{\partial \tau} = \delta_3(\eta_2 - \Lambda) + \nabla_\xi^2 \Lambda. \quad (6)$$

The system has four spatially-homogeneous steady states:

- $\eta_1 = 0, \eta_2 = 0$: the trivial solution.
- $\eta_1 = 1, \eta_2 = 0$: corresponding to normal healthy tissue with no tumour cells present.
- $\eta_1 = 1 - \delta_1, \eta_2 = 1$: corresponding to tissue consisting of both normal and tumour cells at an intermediate level, which may be interpreted as a less aggressive (but invasive) tumour. (Note that this is only biologically realistic for non-negative values of the density).
- $\eta_1 = 0, \eta_2 = 1$: corresponding to total tumour invasion.

Linear stability analysis [15] shows us that the trivial state and the state corresponding to normal cells alone are unconditionally unstable. Both the invasive state and the coexisting state are conditionally, but mutually exclusively, stable. The critical parameter is found to be $\delta_1 = d_1 r_3 K_2 / d_3 r_1$. Depending on the value of this dimensionless parameter, either the steady state for total destruction of normal tissue ($\delta_1 > 1$) or the steady state with the tumour and normal cells coexisting ($\delta_1 < 1$) is stable. Thus as the value of δ_1 passes through the critical value of 1, the entire system will change from a less aggressive to a more aggressive invasive pattern. For example, increased tumour vascularity will increase K_2 and push the system to an unstable steady state. A detailed analytical study of this system reveals a rich variety of wave propagation dynamics with fast and slow waves [5].

Late-time travelling wave solutions [15] to Eqs. (4)–(6) are computed in Fig. 1. The first point of note is that the model predicts a smooth pH gradient extending from the tumour edge into the peritumoural tissue. The authors re-analyse data presented by Martin and Jain [12] relating to *in vivo* interstitial pH profiles for the VX2 rabbit carcinoma and its surrounding normal tissue, demonstrating that the data are consistent with the presence and approximate range of the pH gradient predicted by the model. Most significantly, however, the model predicts that (when $\delta_1 > 1$) there exists a previously unrecognised acellular gap separating the advancing tumour and receding host tissue fronts. In subsequent *in vitro* experiments, the authors found that, of 21 specimens of human squamous cell carcinoma of the head and neck, 14 were judged to show such a gap. Naked nuclei and morphologically disrupted cells were frequently observed scattered within the gap, or at its edge, as predicted by the model [10].

The GG model focuses on malignant invasion and not transition from benign to malignant states. This issue is addressed in [19, 20], in a model in which this transition occurs as a critical parameter breaches a bifurcation value. This is consistent with data [8] showing that the acquisition of the angiogenic phenotype radically and abruptly alters the tumour growth pattern from non-invasive, slow growth to rapidly expanding, invasive growth.

3. Hybrid Modelling Approaches

Despite the apparent success of Gatenby and Gawlinski's model in examining large, clinically apparent tumours, its relevance to early tumour growth is not clear. Continuous partial differential equation models are well suited to modelling large populations, but individual-based models such as cellular automata (CA) are more appropriate when the evolutionary dynamics of individual cells must be considered. However, traditional CA methods lack the ability to deal with continuously varying elements such as substrate diffusion and utilisation. Thus, hybrid CA have been developed to investigate early cancer development [2, 16].

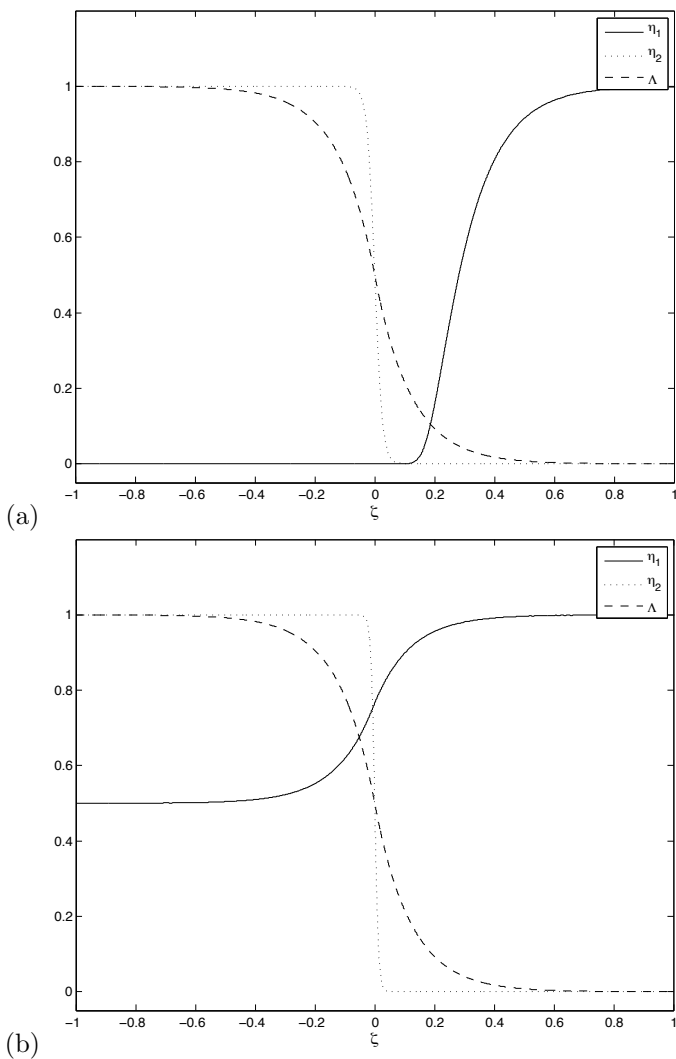


Figure 1. (From Eqs. (4)–(6).) Late-time travelling wave solutions to Gatenby and Gawlinski's model, with respect to the moving coordinate $\zeta = \xi - c\tau$. Waves are propagating from left to right and parameter values used are $\rho_2 = 1$, $\Delta_2 = 4 \times 10^{-5}$ and $\delta_3 = 70$. (a) The invasive case with $\delta_1 = 12.5 > 1$. Notice the formation of an acellular gap separating the advancing tumour (η_2) and receding host tissue (η_1) fronts. (b) The “benign” case with $\delta_1 = 0.5 < 1$. Notice the coexistence of tumour and host tissue behind the wave front. In both cases there is a smooth pH gradient (Λ) extending from the tumour edge into the surrounding normal tissue.

The cellular automaton model used here [18] is composed of an $M \times N$ array of automaton elements with a specific rule-set governing their evolution, as well as glucose (g), oxygen (c) and H^+ (h) fields, each satisfying reaction-diffusion equations. A two-dimensional automaton is used as we focus on growth away from the basement membrane, rather than along the duct. In the model we reflect the avascular geometry of premalignant epithelia by assuming that one edge of the array represents the basement membrane.

We consider the selective pressures placed on a number of different possible tumour phenotypes. Initially, the automaton consists of a single layer of normal epithelial tissue. As well as proliferation and death, these cells may randomly undergo three possible heritable changes, either through mutations or epigenetic changes such as alterations in the methylation patterns of promoters. The cells may become hyperplastic (allowing growth away from the basement membrane), glycolytic (increasing their rate of glucose uptake and utilisation) or acid-resistant (requiring a lower extracellular pH to induce toxicity). These three changes give rise to $2^3 = 8$ different phenotype combinations, and thus eight competing cellular populations.

Cellular metabolism Suppose that a cell consumes glucose and oxygen at rates ϕ_g and ϕ_c , respectively, and that they are used to produce ATP and H^+ at rates ϕ_a and ϕ_h . In non-dimensional form, we have

$$\phi_g = \begin{cases} g & \text{in a normal cell,} \\ kg & \text{in a glycolytic cell,} \end{cases} \quad (7)$$

$$\phi_c = c, \quad (8)$$

$$\phi_a = c + n(\phi_g - c), \quad (9)$$

$$\phi_h = \phi_g - c, \quad (10)$$

subject to the condition $\phi_g \geq c$.

Metabolite profiles After each automaton generation, the known rates of metabolite consumption and production for each cell are used to calculate the corresponding metabolite profiles. Note that metabolite diffusion time-scales (\sim minutes) are much shorter than the cellular proliferation time-scale (\sim days), and thus we may assume that metabolites are in equilibrium at all times. Assuming that diffusion is the primary method for metabolite movement within the tissue, profiles are given in non-dimensional form by

$$d_g^2 \nabla_\xi^2 g = \phi_g, \quad (11)$$

$$d_c^2 \nabla_\xi^2 c = \phi_c, \quad (12)$$

$$\nabla_\xi^2 h = -\phi_h, \quad (13)$$

which may be solved on the square grid using a finite difference approximation. As boundary conditions, we assume zero flux at the edge furthest from the basement membrane (as there are no sources or sinks beyond this point), and periodic boundary conditions at the two sides. At the membrane, we assume glucose and oxygen are fixed at their normal levels $g_{0,j} = c_{0,j} = 1$ (as the stroma is well-vascularised); H^+ is also fixed, $h_{0,j} = h_X$, where the parameter h_X reflects the level of systemic acidosis.

Cell dynamics Cells may proliferate, adapt or die, and cells with different phenotypic patterns respond to the microenvironmental pressures in different ways. As such, competition is incorporated into the model: for a new population to progress and grow, it must successfully compete for space and resources with existing populations. The rules governing the evolution of the automaton elements are as follows:

- If the amount of ATP produced by a cell (ϕ_a) falls below a critical threshold value, a_0 , it dies, and the element becomes empty; a_0 represents the level of ATP required for normal cellular maintenance.
- The local H^+ level may also induce cellular death, with probability p_{dea} , defined by

$$p_{\text{dea}} = \begin{cases} h/h_N & \text{in a normal cell} \\ h/h_T & \text{in an acid-resistant cell} \end{cases} \quad (14)$$

where $h_N < h_T$. Thus the probability of cell death increases with acidity, and the cell will always die if the H^+ level is greater than h_N or h_T , dependent on the cell type under consideration.

- If the cell is not attached to the basement membrane, and is not hyperplastic, it dies.
- If the cell does not die through any of the mechanisms above, it either attempts to divide, with probability p_{div} , or becomes quiescent. The probability of division is a function of the cellular ATP production

$$p_{\text{div}} = (\phi_a - a_0)/(1 - a_0). \quad (15)$$

Hence we assume that the probability of division is proportional to the ATP generated that is not needed for maintenance. If there is more than one neighbouring empty space, the new cell goes to the element with the largest oxygen concentration (following [1]).

- If a cell divides, each of the two daughter cells has probability p_a of randomly acquiring one of the three heritable characteristics (hyperplasia,

glycolysis and acid-resistance). In order to avoid bias in the model, we assume these changes are reversible. For example, a cell displaying constitutive up-regulation of glycolysis may revert to normal glucose metabolism; if this metabolism is most appropriate for the current microenvironmental conditions, the cell will successfully compete for resources with its neighbours.

Fig. 2 presents a typical result from our hybrid CA model. Initially, normal epithelial cells line the basement membrane (a). Acquisition of the hyperplastic phenotype allows growth away from the membrane towards the oxygen diffusion limit (b). Beyond this point, cells cannot exist as the oxygen levels are insufficient to meet cellular ATP demands. This drives adaptation to a glycolytic phenotype, less reliant on oxygen for ATP production (c). The increased ATP levels within glycolytic cells give a competitive advantage over the existing population, thus glycolytic cells dominate the system. Note, however, that the total number of cells within the system has decreased; the increased reliance on glycolysis has resulted in higher levels of acidity, in turn inducing cell death. Further adaptation occurs to an acid-resistant phenotype (d). Increased use of glycolysis allows growth well beyond the oxygen diffusion limit, whilst the cells are more resistant to the resulting acidosis.

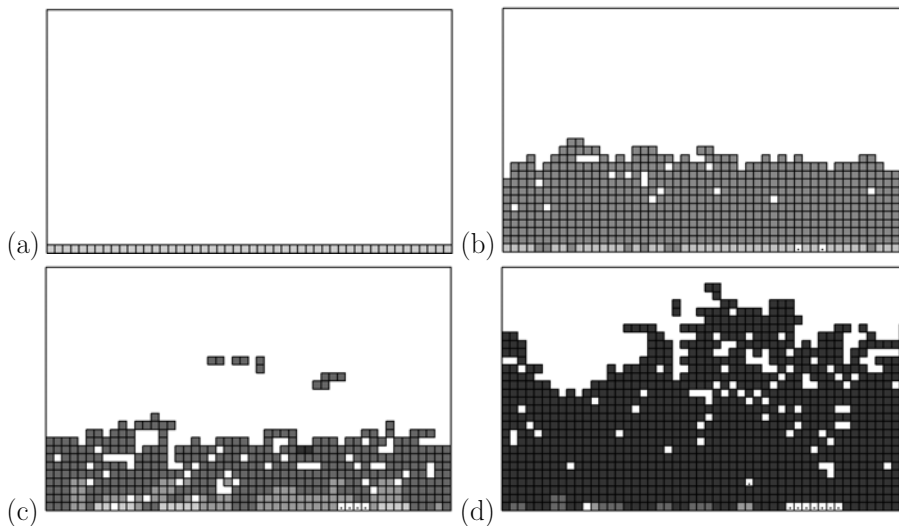


Figure 2. The temporal evolution of a typical cellular automaton after (a) $t = 0$, (b) $t = 100$, (c) $t = 250$ and (d) $t = 300$ generations. Darker denotes a more aggressive phenotype. Shown are normal epithelial (\blacksquare), hyperplastic (\blacksquare), hyperplastic-glycolytic (\blacksquare) and hyperplastic-glycolytic-acid-resistant (\blacksquare) cells. Cells with other phenotypic patterns are shown as \square . Parameter values used are $N = 50$, $n = 5.6 \times 10^{-2}$, $k = 10$, $d_g = 1.3 \times 10^2$, $d_c = 5$, $a_0 = 0.1$, $h_N = 9.3 \times 10^2$, $h_T = 8.6 \times 10^3$, $p_a = 10^{-3}$ and $h_X = 0$.

It is interesting to note that throughout the simulations performed, the heritable changes within the dominant population are accumulated in this same order. Within our model, the underlying environmental selection parameters drive the cells to always follow this adaptive pathway – escaping in turn from the constraints of limited proliferation (hyperplasia), substrate availability (glycolysis) and waste removal (acid-resistance). The same order of progression occurs despite allowing phenotypic reversibility within our model. This means mutations are not a necessary mechanism for phenotypic variation within tumour tissue; rather the model demonstrates that reversible, epigenetic changes are sufficient to drive global change. Of course reversibility is not necessary to observe this adaptation; if irreversible, we would see the same phenotype emerge on a slightly shorter time-scale.

4. Potential Applications I: Bicarbonate Treatment

Recently, we have used compartmental models to predict the effect of bicarbonate treatment on humans and shown, through a sensitivity analysis, that this could best be made more effective by combination with proton inhibitors [13].

5. Potential Applications II: Exercise

There is accumulating evidence that regular physical activity is an effective cancer prevention strategy. Friedenrich and Orenstein [9] recently reviewed over 170 epidemiological studies and concluded that evidence for decreased cancer risk with increased physical activity was convincing for breast and colon cancer, probable for prostate cancer and possible for lung cancer. We hypothesise that exercise produces toxicity within *in situ* cancers through transient decreases in serum pH and, by doing so, will also transiently cause significant further decrease in extracellular pH in the already acidic regions within *in situ* cancers. This abrupt increase in acid concentrations will result in tumour cell death and interrupt the adaptive mechanisms necessary for subsequent evolution to the malignant phenotype. To test the hypothesis, we extend the CA to include variations in systemic pH.

When investigating transient acidosis, each time-step is split into two parts: a proportion of time $\tau \in [0, 1]$ spent at high acidity $h_X \gg 1$, followed by a proportion of time $1 - \tau$ at normal acidity $h_X = 0$. Letting p_0 denote the probability of death p_{dea} , division p_{div} , or mutation p_a during one time unit (as defined previously), the corresponding probability p of occurrence during the acidic phase is given by

$$p(\tau) = 1 - (1 - p_0)^\tau, \quad (16)$$

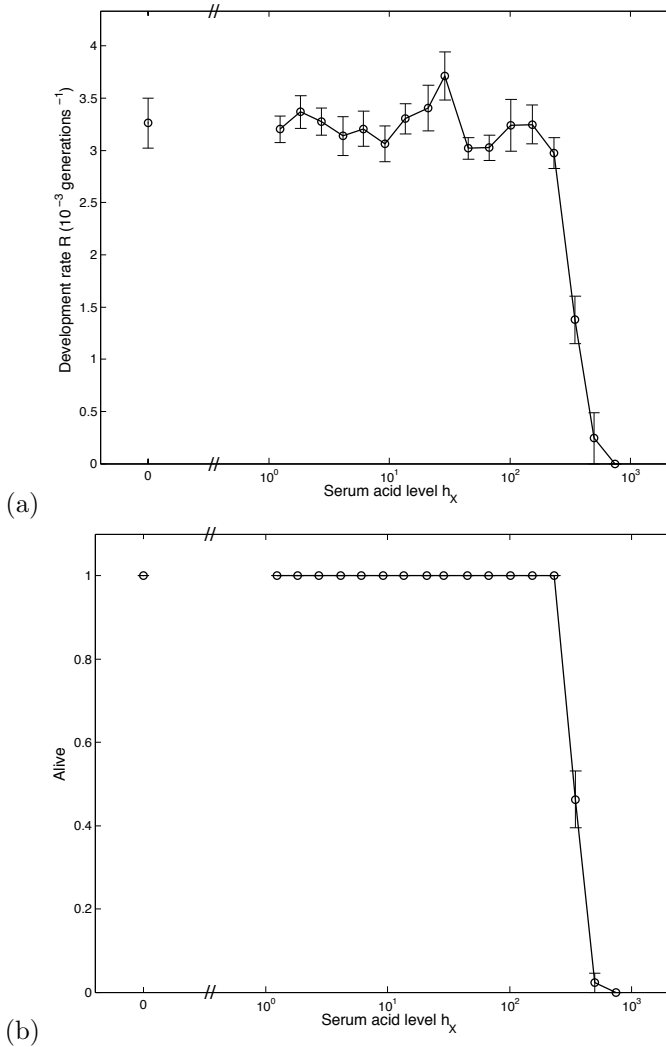


Figure 3. Effect of sustained acidosis. (a) Variation in the development rate R with serum acid level h_X (plotted on a log scale). Each data point is the mean value of R calculated over 50 simulations, whilst the accompanying error bars show the standard errors of these means. (b) Variation in epithelium survival with h_X .

whilst the probability of occurrence during the normal phase is given by

$$p(1 - \tau) = 1 - (1 - p_0)^{1-\tau}. \tag{17}$$

In order to examine the effects of parameter changes on system dynamics, we define a measure of the ‘fitness’ of a specific parameter set. Let ‘invasive’

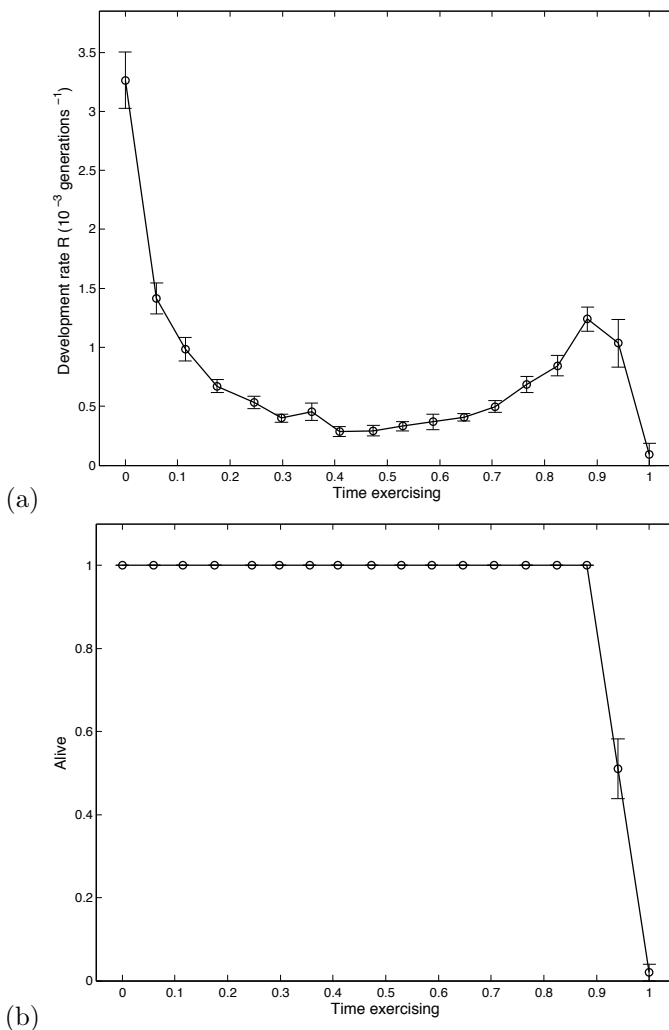


Figure 4. Effect of transient acidosis. (a) Variation in the development rate R with proportion of time under exercise. Exercise is assumed to correspond to high acidity ($h_X = 400$), whilst during rest acidity drops to normal levels ($h_X = 0$). (b) Variation in epithelium survival with exercise time.

be used to describe cells displaying all three heritable changes – hyperplasia, glycolysis and acid resistance. For a particular automaton, let T denote the number of generations after which 95% of the cells in the system display the invasive phenotype; thus T is representative of the amount of time taken for full carcinogenesis to occur. Now let the development rate $R = T^{-1}$, where we take $R = 0$ if $T \geq 5000$ (equivalent to approximately 20 years) – i.e. assume no

carcinogenesis occurs. Automata with a higher value of R proceed more quickly through the carcinogenesis pathway.

From Fig. 3 (a) we see how the development rate R varies with changes in serum acidity h_X . We vary the external acid levels from $h_X = 0$ (normal) to $h_X \sim 1000$, equivalent to pH 6.8, corresponding to the threshold for normal cell survival [16]. Development rate R remains fairly constant until $h \sim 100$ (a drop of around 0.1 pH units), when a marked decrease is observed. Looking further however, we see (Fig. 3 (b)) that this result follows simply because the harsher conditions lead to death of the entire epithelium; normal cells die out before having the opportunity to turn cancerous.

Since the model predicts that permanent acidosis cannot arrest cancer development, we move on to investigate transient acidosis, allowing the system to spend a certain proportion of time at high acidity and a certain proportion at normal acidity; this transient acidosis mimics what occurs when engaging in rigorous exercise followed by rest. In Fig. 4 (a) we see how the development rate R varies with the amount of time exercising. We see that only a small proportion of time spent at low pH ($h = 400$, a drop of around 0.25 pH units) leads to a significant reduction in R . By contrast to the previous figure, the behaviour is not due to total epithelial death (Fig. 4 (b)).

6. Discussion

We have used a range of mathematical modelling techniques to explore the acid-mediated tumour invasion hypothesis. The models have made a number of predictions which have been experimentally verified. The therapeutic implications, namely either buffering acid or manipulating the phenotypic selection process, have been described. It should be noted that while we have focussed here on the competitive interaction between tumour and normal cells, there is also a cooperative interaction between them in the production of enzymes to degrade extracellular material. We have recently extended the Gatenby and Gawlinski model mentioned in this paper to show that invasion may behave in a biphasic way in response to acid [14], suggesting that more subtle therapeutic approaches may be necessary.

References

- [1] T. Alarcon, H. M. Byrne, P. K. Maini, *A cellular automaton model for tumour growth in inhomogeneous environment.*, J. Theor. Biol. **225** (2003), 257–274.
- [2] A. R. A. Anderson, *A hybrid mathematical model of solid tumour invasion: the importance of cell adhesion.*, Math. Med. Biol. **22** (2005), 163–186.
- [3] J. Casciari, S. Sotirchos, R. Sutherland, *Variations in tumor cell growth rates and metabolism with oxygen concentration, glucose concentration, and extracellular pH*, J. Cell. Physiol. **151** (1992), 386–394.

- [4] G. Di Chiro, J. Hatazawa, D. Katz, H. Rizzoli, D. D. Michele, *Glucose utilization by intracranial meningiomas as an index of tumor aggressivity and probability of recurrence: a PET study*, *Radiology* **164** (1987), 521–526.
- [5] A. Fasano, M. A. Herrero, M. R. Rodrigo, *Slow and fast invasion waves in a model of acid-mediated tumour growth*, *Math. Biosci.* **220** (2009), 45–56.
- [6] E. Fearon, B. Vogelstein, *A genetic model for colorectal tumorigenesis*, *Cell* **61** (1990), 759–767.
- [7] I. Fidler, I. Hart, *Biological diversity in metastatic neoplasms: origins and implications*, *Science* **217** (1982), 998–1003.
- [8] J. Folkman, *The role of angiogenesis in tumor growth*, *Semin. Cancer Biol.* **3** (1992), 65–71.
- [9] C. M. Friedenreich, M. R. Orenstein, *Physical activity and cancer prevention: etiologic evidence and biological mechanisms.*, *J. Nutr.* **132** (2002), 3456S–3464S.
- [10] R. Gatenby, E. Gawlinski, *A reaction-diffusion model of cancer invasion*, *Cancer Res.* **56** (1996), 5745–5753.
- [11] R. Gatenby, R. Gillies, *Why do cancers have high aerobic glycolysis?*, *Nature Rev. Cancer* **4** (2004), 891–899.
- [12] G. Martin, R. Jain, *Noninvasive measurement of interstitial pH profiles in normal and neoplastic tissue using fluorescence ratio imaging microscopy*, *Cancer Res.* **54** (1994), 5670–5674.
- [13] N. Martin, E. Gaffney, R. Gatenby, R. Gillies, I. Robey, P. Maini, *A mathematical model of tumour and blood pH*, submitted.
- [14] N. Martin, E. Gaffney, R. Gatenby, P. Maini, *Tumour-stromal interactions in acid-mediated invasion: A mathematical model*, submitted.
- [15] J. Murray, *Mathematical Biology I. An Introduction*, Springer-Verlag, 2002, 3rd edition.
- [16] A. A. Patel, E. T. Gawlinski, S. K. Lemieux, R. A. Gatenby, *A cellular automaton model of early tumor growth and invasion.*, *J. Theor. Biol.* **213** (2001), 315–331.
- [17] P. Schornack, R. Gillies, *Contributions of cell metabolism and H^+ diffusion to the acidic pH of tumors*, *Neoplasia* **5** (2003), 135–145.
- [18] K. Smallbone, R. A. Gatenby, R. J. Gillies, P. K. Maini, D. J. Gavaghan, *Metabolic changes during carcinogenesis: potential impact on invasiveness.*, *J. Theor. Biol.* **244** (2007), 703–713.
- [19] K. Smallbone, R. A. Gatenby, P. K. Maini, *Mathematical modelling of tumour acidity.*, *J. Theor. Biol.* **255** (2008), 106–112.
- [20] K. Smallbone, D. Gavaghan, R. Gatenby, P. K. Maini, *The role of acidity in solid tumour growth and invasion*, *J. Theor. Biol.* **235** (2005), 476–484.
- [21] R. Thomlinson, L. Gray, *The histological structure of some human lung cancers and the possible implications for radiotherapy*, *Br. J. Cancer* **9** (1955), 539–549.
- [22] O. Warburg, *The Metabolism of Tumours*, Constable Press, London, 1930.

On Markov State Models for Metastable Processes

Natasa Djurdjevac, Marco Sarich, and Christof Schütte*

Abstract

We consider Markov processes on large state spaces and want to find low-dimensional structure-preserving approximations of the process in the sense that the longest timescales of the dynamics of the original process are reproduced well. Recent years have seen the advance of so-called Markov state models (MSM) for processes on very large state spaces exhibiting metastable dynamics. It has been demonstrated that MSMs are especially useful for modelling the interesting slow dynamics of biomolecules (cf. Noe et al, PNAS(106) 2009) and materials. From the mathematical perspective, MSMs result from Galerkin projection of the transfer operator underlying the original process onto some low-dimensional subspace which leads to an approximation of the dominant eigenvalues of the transfer operators and thus of the longest timescales of the original dynamics. Until now, most articles on MSMs have been based on full subdivisions of state space, i.e., Galerkin projections onto subspaces spanned by indicator functions. We show how to generalize MSMs to alternative low-dimensional subspaces with superior approximation properties, and how to analyse the approximation quality (dominant eigenvalues, propagation of functions) of the resulting MSMs. To this end, we give an overview of the construction of MSMs, the associated stochastic and functional-analysis background, and its algorithmic consequences. Furthermore, we illustrate the mathematical construction with numerical examples.

Mathematics Subject Classification (2010). Primary 65C50; Secondary 60J35.

Keywords. Markov process, metastability, transition path theory, milestoning, eigenvalue problem, transfer operator, propagation error, Markov state models, committor, Galerkin approximation

*Supported by the DFG research center MATHEON “Mathematics for key technologies” in Berlin.

Fachbereich Mathematik und Informatik, Institut für Mathematik, Freie Universität Berlin. E-mails: djurdjev@mi.fu-berlin.de, sarich@mi.fu-berlin.de, and schuette@mi.fu-berlin.de.

1. Introduction

We consider Markov processes on large state spaces that have a unique invariant measure. We are interested in the question of whether we can find a low-dimensional approximation of the process in the sense that the longest timescales of the dynamics of the original process are reproduced well and the low-dimensional model inherits the essential structural properties of the original process: the dynamics transports probability distribution (or densities, respectively) into probability distributions (or densities), i.e., non-negativity and normalization are preserved. This is a rather old problem that has been answered in many different ways some belonging to classical themes in the literature [1, 2] that have been discussed in hundreds of articles, e.g., Markov chain decomposition for nearly reducible chains (for example, by aggregation-disaggregation techniques [3, 2, 4], stochastic complementation [5, 6], Peron Cluster Cluster Analysis (PCCA) [7, 8]), or network partition problems [9]. In these classical topics most contributions consider finite state spaces and have been based on linear algebra and associated stochastic analysis approaches.

Recent years have seen the advance of so-called Markov state models (MSM) as low-dimensional models for processes on very large, mostly on continuous state spaces exhibiting metastable dynamics [10, 7, 11, 12, 13]. Recently the interest in MSMs has drastically increased since it could be demonstrated that MSMs can be constructed even for very high dimensional systems [11] and have been especially useful for modelling the interesting slow dynamics of biomolecules [14, 15, 16, 17, 18, 19] and materials [20] (there under the name “kinetic Monte Carlo”). Metastable dynamics means that one can subdivide state space into metastable sets in which the system remains for *long* periods of time before it exits *quickly* to another metastable set; here the words “long” and “quickly” mainly state that the typical residence time has to be much longer than the typical transition time so that the jump process between the metastable sets is approximately Markovian. An MSM then just describes the Markov process that jumps between the sets with the aggregated statistics of the original process.

The approximation quality of a MSM on large time scales has been rigorously studied for many different systems, e.g., for diffusion processes, or Glauber dynamics and Ising models in the limit of vanishing smallness parameters (noise intensity, temperature) where the analysis can be based on large deviation estimates and variational principles [21, 22] and/or potential theory and capacities [23, 24]. In these cases the effective dynamics is governed by some MSM with exponentially small transition probabilities and its states label the different attracting sets of the underlying Markov process. Other, quite general, rigorous approaches to the construction of MSM involve the exploitation of spectral properties, where the relation between dominant eigenvalues and eigenvectors, exit times and rates, and metastable sets has been studied

in detail, in some cases even without assumptions about smallness parameters [25, 26, 12, 13, 7, 19].

In this contribution we will use the approach via Galerkin discretization of the *transfer operator* of the original Markov process as developed in [12, 11, 7, 10]; here “transfer operator” just refers to a generalization of the transition matrix on finite discrete state spaces to general, e.g., continuous state spaces. In this approach the low-dimensional approximation results from orthogonal projection of the transfer operator onto some low-dimensional subspace. For so-called *full partition MSM* this subspace is spanned by indicator functions of n sets that partition state space. Then, the Galerkin approach has a direct stochastic interpretation since the resulting n -dimensional approximation simply exhibits jumps between the sets with aggregated statistics as mentioned above.

However in many cases indicator ansatz spaces do not allow to achieve good approximation quality for reasonably small numbers of sets [19]. Therefore other ansatz spaces, e.g., fuzzy ansatz spaces, have also been discussed [27]. This obviously raises the questions of (a) how to find good ansatz functions, (b) what may be the associated stochastic interpretation, and (c) what can be said about the resulting approximation quality. Let $D = \text{span}\{q_1, \dots, q_n\}$ denote the low-dimensional ansatz space in state space S . We will be interested in ansatz functions q_i that are non-negative functions with $\sum_{x \in S} q_i(x) = 1$ or $\int_S q_i(x) dx = 1$ for any i so that $q_i(x)$ can be interpreted as the probability (density) that state x belongs to MSM state i . We will herein discuss an approach that allows to identify such ansatz functions and answers the above three questions jointly for full partition and fuzzy ansatz functions. The key idea will be that we consider n sets C_1, \dots, C_n that (in general) do *not* partition the state space but are just the very cores of the different attracting sets of the underlying Markov process. These *core sets* are then used as *milestones* in the sense of the milestoneing approach as introduced in [28]: The approximating m -dimensional milestoneing process is assigned to state i whenever the last core entered by the original process has been C_i . We will see that we can relate the milestoneing process to transition path theory [29, 30, 31, 14] and use it to construct good fuzzy ansatz functions. The resulting low-dimensional MSM will prove to have very good approximation properties whenever the core sets have been chosen appropriately.

The remainder of the paper is organized as follows. In Section 2 we introduce the setting, define transfer operators, introduce full-partition MSM and relate them to Galerkin projections. Then, in Sec. 3 we introduce the milestoneing process, relate it to transition path theory, and analyse its transition statistics. Section 4 then discusses Galerkin projection of the transfer operator in general, gives rigorous approximation results for long-term behavior and for eigenvalues and related timescales, and then shows how to use the milestoneing process to compute the resulting MSMs efficiently. Finally, the results are illustrated with numerical experiments in Section 5.

2. Setting the Scene

We consider a Markov process $(X_t)_{t \in T}$ on a discrete state space S and its associated family of transition matrices $(P_t)_{t \in \mathbb{N}}$ with entries

$$p_t(x, y) = \mathbb{P}[X_t = y | X_0 = x]. \quad (1)$$

We restrict our considerations to discrete state spaces just for simplicity of presentation; all statements made in the following can be generalized to continuous state spaces as well (see Remark 2.1):

Because (X_t) is a Markov process, the transition matrices have the *semi-group property*

$$P_t P_s = P_{t+s}. \quad (2)$$

If (X_t) is a time-discrete process, i.e. a Markov chain with $T = \mathbb{N}$, we will only consider $P := P_1$, because (2) implies

$$P_t = P^t. \quad (3)$$

If (X_t) is time-continuous, it is usually referred to as a Markov jump process. In this case, the dynamics of the process is given by its *generator* L with entries $l(x, y)$ such that

$$P_t = e^{Lt}. \quad (4)$$

The generator L can also be defined explicitly

$$L = \lim_{t \rightarrow \infty} \frac{P_t - Id}{t}, \quad (5)$$

so that its entries form a rate matrix

$$l(x, y) \geq 0, \quad x \neq y, \quad l(x, x) = - \sum_{y \neq x} l(x, y). \quad (6)$$

For the time-discrete case, we define an analog for the rate matrix by setting $L_d = P - Id$ which we call *discrete generator*.

In the following we always assume that (X_t) has a unique invariant measure μ , that is given by

$$(P_t \mu)(y) = \sum_{x \in S} p_t(x, y) \mu(x) = \mu(y). \quad (7)$$

Now we introduce the family of *transfer operators* (T_t) that describes the propagation of densities in L_μ^2

$$(T_t f)(y) \mu(y) = \sum_x f(x) p_t(x, y) \mu(x) \quad (8)$$

and set $T := T_1$ for discrete time.

In analogy, we define on L_μ^2

$$(\mathcal{L}f)(y)\mu(y) = \sum_x l(x, y)f(x)\mu(x) \tag{9}$$

and for the discrete case

$$\mathcal{L}_d = T - Id. \tag{10}$$

In the following we will only consider the scalar product in L_μ^2 , the induced 2-norm and the 1-norm

$$\langle f, g \rangle = \sum_x f(x)g(x)\mu(x), \quad \|f\|^2 = \langle f, f \rangle, \quad \|f\|_1 = \sum_x |f(x)|\mu(x). \tag{11}$$

In the theory of building standard Markov state models (MSM) one chooses a partitioning of state space, i.e. sets A_1, \dots, A_n , such that

$$A_i \cap A_j = \emptyset, \quad i \neq j, \quad \bigcup_{i=1}^n A_i = S \tag{12}$$

and a certain *lag time* $\tau > 0$. Then one can compute the transition probabilities

$$\mathbb{P}[X_\tau \in A_j | X_0 \in A_i]$$

and use the corresponding Markov chain on the index space $\{1, \dots, n\}$ to approximate the original dynamics, switching between those sets. The approximation quality of such MSMs is discussed in [19]. A key feature is, that the Markov chain on the index space represents the dynamics of a projection of the transfer operator, that is $QT_\tau Q$, where Q is the orthogonal projection onto

$$D = \text{span} \{ \mathbb{1}_{A_1}, \dots, \mathbb{1}_{A_n} \}.$$

As outlined above, we will not restrict our attention to full partitionings of state space. However, we will return to the analysis of Galerkin projections of transfer operators $QT_\tau Q$, also to projections onto step-function spaces.

Remark 2.1. *On continuous state space the transfer operator $T_t : L_\mu^2 \rightarrow L_\mu^2$ is defined via*

$$\int_C T_t f(y)\mu(dy) = \int_S \mathbb{P}[X_t \in C | X_0 = x] f(x)\mu(dx), \quad \text{for all measurable } C \subset S,$$

for the general case where the transition function $p(t, x, C) = \mathbb{P}[X_t \in C | X_0 = x]$ as well as the invariant measure may contain singular as well as absolutely continuous parts. Then, all of the above and subsequent sums have to be replaced by respective integrals. Further details, in particular regarding the respective generators for, e.g., diffusion processes, can be found in [12].

3. Milestoning and Transition Path Theory

We will now follow the approach first introduced in [32]. While the approach in [32] is restricted to reversible processes with generators, we will herein present the general framework for non-reversible processes.

3.1. Core sets and committors. Motivated by standard Markov state model approach we define sets $C_1, \dots, C_n \subset S$, that we will call *core sets*, such that

$$C_i \cap C_j = \emptyset, \quad i \neq j. \quad (13)$$

That is, we relax the full partition constraint in (12). We denote the region that is not assigned to any core set by

$$C = S \setminus \bigcup_{k=1}^n C_k.$$

For analyzing the switching dynamics of the original process between the core sets we introduce the *milestoning process* (\hat{X}_t)

$$\hat{X}_t = i \Leftrightarrow X_{\sigma(t)} \in C_i, \quad \text{with } \sigma(t) = \sup_{s \leq t} \left\{ X_s \in \bigcup_{k=1}^n C_k \right\}, \quad (14)$$

i.e. the milestoning process is in state i , if the original process came last from core set C_i , cf. [28].

Now let $q_i^+(x)$ denote the probability that the process (X_t) will visit the core set C_i next, conditional on being in state x . q_i^+ is usually referred to as the *forward committor* and, as for example in [30], one can derive that q_i^+ is the solution of

$$\begin{aligned} (Lq_i^+)(x) &= 0, \quad \forall x \in C, \\ q_i^+(x) &= 1, \quad \forall x \in C_i, \\ q_i^+(x) &= 0, \quad \forall x \in C_j, j \neq i. \end{aligned} \quad (15)$$

In a similar way it can be shown, that the *backward committor* $q_i^-(x) = \mathbb{P}[\hat{X}_t = i | X_t = x]$, i.e. the probability that the process (X_t) came last from core set C_i , conditional on being in state x , solves

$$\begin{aligned} (\mathcal{L}q_i^-)(x) &= 0, \quad \forall x \in C, \\ q_i^-(x) &= 1, \quad \forall x \in C_i, \\ q_i^-(x) &= 0, \quad \forall x \in C_j, j \neq i. \end{aligned} \quad (16)$$

In the time-discrete case one has to replace L by the discrete generator L_d and \mathcal{L} by \mathcal{L}_d . Moreover one can show, that (15) and (16) have a unique solution under the assumption that the invariant measure is unique and not vanishing on all core sets.

Note that \mathcal{L} or in time-discrete setting \mathcal{L}_d generate the family of transition matrices (P_t^b) with entries

$$p_t^b(y, x) = \frac{\mu(x)}{\mu(y)} p_t(x, y), \tag{17}$$

which describe the dynamics of the process (X_t) running backward in time.

For more details on the definition and properties of committors we refer to [29, 30, 31, 14]; the discrete setting studied herein is worked out in [30].

3.2. Jump statistics of milestoning process.

Time-discrete case: Transition probabilities. When observing a time-discrete process (X_n) , we can define the transition matrix \hat{P} of the milestoning process (\hat{X}_n) , with entries $\hat{p}(i, j) = \mathbb{P}_\mu(\hat{X}_{n+1} = j | \hat{X}_n = i)$. Since in general the milestoning process will not be a Markov process, we cannot assume that it is essentially characterized by its transition matrix \hat{P} ; this also holds true for the generator \hat{L}_d whose definition therefore should be understood as a formal one at this point. We will see that it is *not* the crucial point whether the dynamics of the milestoning process is governed by \hat{P} or not.

Based on the introduced quantities we have

$$\mathbb{P}_\mu(\hat{X}_n = i, X_n = x) = \mathbb{P}_\mu(\hat{X}_n = i | X_n = x) \mathbb{P}_\mu(X_n = x) = q_i^-(x) \mu(x).$$

Therefore, the total probability that the milestoning process is assigned to state i , i.e. the invariant measure of the milestoning process is

$$\hat{\mu}(i) = \mathbb{P}_\mu(\hat{X}_n = i) = \sum_x \mathbb{P}_\mu(\hat{X}_n = i, X_n = x) = \sum_x q_i^-(x) \mu(x) = \|q_i^-\|_1.$$

The following theorem gives us the entries of the discrete generator.

Theorem 3.1. *For a time-discrete process (X_n) , the entries of the discrete generator \hat{L}_d of the milestoning process (\hat{X}_n) are given with*

$$\hat{l}_d(i, j) = \frac{1}{\|q_i^-\|_1} \langle q_j^+, \mathcal{L}_d q_i^- \rangle. \tag{18}$$

Proof. Using that

$$\begin{aligned} &\mathbb{P}_\mu(X_{n+1} = y, \hat{X}_n = i, X_n = x) \\ &= \mathbb{P}_\mu(X_{n+1} = y | \hat{X}_n = i, X_n = x) \mathbb{P}_\mu(\hat{X}_n = i, X_n = x) \\ &= p(x, y) q_i^-(x) \mu(x), \end{aligned}$$

we can calculate

$$\begin{aligned} &\mathbb{P}_\mu(\hat{X}_{n+1} = j, X_{n+1} = y, \hat{X}_n = i, X_n = x) = \\ &= \mathbb{P}_\mu(\hat{X}_{n+1} = j | X_{n+1} = y, \hat{X}_n = i, X_n = x) \mathbb{P}_\mu(X_{n+1} = y, \hat{X}_n = i, X_n = x) \\ &= \begin{cases} \mathbb{1}_{C_j}(y) p(x, y) q_i^-(x) \mu(x), & \text{if } i \neq j \\ \mathbb{1}_{C_i \cup C}(y) p(x, y) q_i^-(x) \mu(x), & \text{if } i = j. \end{cases} \end{aligned}$$

Therefore, the one-step transition probability $\hat{p}(i, j)$ from i to $j \neq i$ is given by

$$\begin{aligned} \hat{p}(i, j) &= \mathbb{P}_\mu(\hat{X}_{n+1} = j | \hat{X}_n = i) = \frac{\mathbb{P}_\mu(\hat{X}_{n+1} = j, \hat{X}_n = i)}{\mathbb{P}_\mu(\hat{X}_n = i)} \\ &= \frac{1}{\mathbb{P}_\mu(\hat{X}_n = i)} \sum_{x, y} \mathbb{P}_\mu(\hat{X}_{n+1} = j, X_{n+1} = y, \hat{X}_n = i, X_n = x) \\ &= \frac{1}{\|q_i^-\|_1} \sum_{x, y} \mathbf{1}_{C_j}(y) p(x, y) q_i^-(x) \mu(x) = \frac{1}{\|q_i^-\|_1} \langle Tq_i^-, \mathbf{1}_{C_j} \rangle. \end{aligned}$$

In addition, when $i = j$

$$\begin{aligned} \hat{p}(i, i) &= \mathbb{P}_\mu(\hat{X}_{n+1} = i | \hat{X}_n = i) = \frac{\mathbb{P}_\mu(\hat{X}_{n+1} = i, \hat{X}_n = i)}{\mathbb{P}_\mu(\hat{X}_n = i)} \\ &= \frac{1}{\mathbb{P}_\mu(\hat{X}_n = i)} \sum_{x, y} \mathbb{P}_\mu(\hat{X}_{n+1} = i, X_{n+1} = y, \hat{X}_n = i, X_n = x) \\ &= \frac{1}{\|q_i^-\|_1} \sum_{x, y} \mathbf{1}_{C_i \cup C}(y) p(x, y) q_i^-(x) \mu(x) \\ &= \frac{1}{\|q_i^-\|_1} \langle Tq_i^-, \mathbf{1}_{C_i \cup C} \rangle. \end{aligned}$$

Using the properties of committors on core sets for $i \neq j$, we get that

$$\begin{aligned} \langle Tq_i^-, \mathbf{1}_{C_j} \rangle &= \langle Tq_i^-, q_j^+ \rangle - \langle Tq_i^-, q_j^+ \mathbf{1}_C \rangle = \langle Tq_i^-, q_j^+ \rangle - \langle q_i^-, q_j^+ \mathbf{1}_C \rangle \\ &= \langle (T - Id)q_i^-, q_j^+ \rangle = \langle \mathcal{L}_d q_i^-, q_j^+ \rangle, \end{aligned}$$

which yields

$$\hat{l}_d(i, j) = \hat{p}(i, j) = \frac{1}{\|q_i^-\|_1} \langle q_j^+, \mathcal{L}_d q_i^- \rangle, \quad i \neq j.$$

Similarly, for $i = j$, we get

$$\begin{aligned} \langle Tq_i^-, \mathbf{1}_{C_i \cup C} \rangle &= \langle Tq_i^-, \mathbf{1}_{C_i} \rangle + \langle Tq_i^-, \mathbf{1}_C \rangle \\ &= \langle Tq_i^-, q_i^+ \rangle - \langle q_i^-, q_i^+ \mathbf{1}_C \rangle + \langle q_i^-, \mathbf{1}_C \rangle \\ &= \langle (T - Id)q_i^-, q_i^+ \rangle + \|q_i^-\|_1 = \langle \mathcal{L}_d q_i^-, q_i^+ \rangle + \|q_i^-\|_1, \end{aligned}$$

and

$$\hat{l}_d(i, i) = \hat{p}(i, i) - 1 = \frac{1}{\|q_i^-\|_1} (\langle q_i^+, \mathcal{L}_d q_i^- \rangle + \|q_i^-\|_1) - 1 = \frac{1}{\|q_i^-\|_1} \langle q_i^+, \mathcal{L}_d q_i^- \rangle.$$

□

Time-continuous case: Transition rates. Now we will show that all the above identities are still valid in a time-continuous case. For a given infinitely long trajectory and $i \neq j$, we define a (i, j) -reactive trajectory as a piece of this infinite long trajectory in a time interval R_{ij}^m such that for any $t \in R_{ij}^m$ we have that the next first entry into a core set is in C_j while the last first entry into a core set happened in C_i . Then, at a certain time t we are on a (i, j) -reactive trajectory if

$$t \in R_{ij} = \cup_{m=-\infty}^{\infty} R_{ij}^m.$$

The probability current from x to y generated by (i, j) -reactive trajectories is then given by

$$f_{ij}(x, y) = \lim_{s \rightarrow 0^+} \frac{1}{s} \mathbb{P}_\mu \left(X_t = x, X_{t+s} = y, t \in R_{ij}, t + s \in R_{ij} \right),$$

In order to compute this quantities we define $B_j = \bigcup_{k \neq j} C_k$ and denote the first hitting time of a set A by τ_A . Then $\mathbb{P}_y[\tau_{C_j} < \tau_{B_j}]$, is the probability to start at y and enter the core set C_j next rather than any other core set. Therefore

$$\begin{aligned} \mathbb{P}_\mu \left(X_t = x, X_{t+s} = y, t \in R_{ij}, t + s \in R_{ij} \right) &= \\ &= \mathbb{P}_\mu \left(X_{t+s} = y, t + s \in R_{ij} \mid X_t = x, t \in R_{ij} \right) \mathbb{P}_\mu \left(X_t = x, t \in R_{ij} \right) \\ &= \mathbb{P}_\mu \left(X_{t+s} = y \mid X_t = x \right) \mathbb{P}_y(\tau_{C_j} < \tau_{B_j}) \mathbb{P}_\mu \left(X_t = x, \hat{X}_t = i \right) \\ &= p_s(x, y) q_j^+(y) q_i^-(x) \mu(x). \end{aligned}$$

Since $i \neq j$ we have $l(x, y) = \lim_{s \rightarrow 0^+} \frac{1}{s} p_s(x, y)$ and thus

$$f_{ij}(x, y) = l(x, y) q_j^+(y) q_i^-(x) \mu(x).$$

Now we can compute the rate k_{ij} of transitions from i to j , which is defined as the average number of (i, j) -reactive trajectories per unit time. This quantity is given by the total probability current through a dividing surface between C_i and C_j , i.e. by the total probability current generated by (i, j) -reactive trajectories through the boundary of C_i :

$$\begin{aligned} k_{ij} &= \sum_{x \in C_i, y \in S \setminus C_i} f_{ij}(x, y) \\ &= \sum_{x \in C_i, y \in S \setminus C_i} q_j^+(y) l(x, y) q_i^-(x) \mu(x) \\ &= \sum_{x \in C_i, y \in S} q_j^+(y) l(x, y) q_i^-(x) \mu(x), \end{aligned}$$

where the last identity results from $q_j^+(y) = 0$ for all $y \in C_i$. Since additionally $q_i^-(x) = 1$ for $x \in C_i$ we find

$$k_{ij} = \langle \mathcal{L} \mathbf{1}_{C_i}, q_j^+ \rangle.$$

Therefore, the off-diagonal entries $\hat{l}(i, j)$ of the generator for the milestoning process \hat{X}_t result as

$$\hat{l}(i, j) = \frac{1}{\|q_i^-\|_1} \langle \mathcal{L}\mathbf{1}_{C_i}, q_j^+ \rangle, \tag{19}$$

such that the diagonal entries have to be

$$\begin{aligned} \hat{l}(i, i) &= - \sum_{j \neq i} \frac{1}{\|q_i^-\|_1} \langle \mathcal{L}\mathbf{1}_{C_i}, q_j^+ \rangle = - \frac{1}{\|q_i^-\|_1} \left\langle \mathcal{L}\mathbf{1}_{C_i}, \sum_{j \neq i} q_j^+ \right\rangle \\ &= - \frac{1}{\|q_i^-\|_1} \langle \mathcal{L}\mathbf{1}_{C_i}, \mathbf{1} - q_i^+ \rangle = \frac{1}{\|q_i^-\|_1} \langle \mathcal{L}\mathbf{1}_{C_i}, q_i^+ \rangle. \end{aligned}$$

Since $\langle \mathcal{L}\mathbf{1}_{C_i}, q_i^+ \rangle = \langle \mathbf{1}_{C_i}, Lq_i^+ \rangle$, we can use the same arguments as above to end up with

$$\langle \mathcal{L}\mathbf{1}_{C_i}, q_j^+ \rangle = \langle \mathcal{L}q_i^-, q_j^+ \rangle,$$

so that we have just proved the following theorem

Theorem 3.2. *For a time-continuous process (X_t) , the entries of a generator of the milestoning process (\hat{X}_t) are given with*

$$\hat{l}(i, j) = \frac{1}{\|q_i^-\|_1} \langle \mathcal{L}q_i^-, q_j^+ \rangle. \tag{20}$$

3.3. Invariant measure and self-adjointness. A Markov process (X_t) is reversible if

$$p_t(x, y)\mu(x) = p_t(y, x)\mu(y). \tag{21}$$

This condition is called *the detailed balance condition*. It obviously implies that

$$p_t^b(x, y) = p_t(x, y), \tag{22}$$

so the process running backward in time is equivalent to the process running forward in time.

Moreover, (21) implies

$$\begin{aligned} \langle Tf, g \rangle &= \sum_{x, y} p(x, y)f(x)g(y)\mu(x) \\ &\stackrel{(21)}{=} \sum_{x, y} p(y, x)f(x)g(y)\mu(y) = \langle f, Tg \rangle. \end{aligned} \tag{23}$$

This means that T is a self-adjoint operator. The same argument shows that also \mathcal{L} is self-adjoint in the reversible case. Further, (22), (15) and (16) yield the identity of forward and backward committors, i.e.

$$q_i^- = q_i^+ \quad \forall i = 1, \dots, n. \tag{24}$$

Hence, in the following we will use the shorthand notation $q_i := q_i^- = q_i^+$.

First we note some properties of the milestoning generator \hat{L} .

Lemma 3.3. *Let (X_t) be a reversible Markov process with unique invariant measure μ . Then the milestoning generator \hat{L} has the invariant measure*

$$\hat{\mu}(i) = \sum_x q_i(x)\mu(x)$$

and the according operator in $L^2(\hat{\mu})$

$$(\hat{\mathcal{L}}v)(j)\hat{\mu}(j) = \sum_{i=1}^n \hat{l}(i, j)v(i)\hat{\mu}(i)$$

is self-adjoint. Therefore it also defines a reversible jump process.

Proof. We have

$$\begin{aligned} \sum_{i=1}^n \hat{l}(i, j)\hat{\mu}(i) &= \sum_{i=1}^n \langle q_i, \mathcal{L}q_j \rangle \\ &= \langle \mathbb{1}, \mathcal{L}q_j \rangle = 0. \end{aligned}$$

Moreover,

$$\begin{aligned} \hat{l}(i, j)\hat{\mu}(i) &= \langle q_i, \mathcal{L}q_j \rangle \\ &= \langle \mathcal{L}q_i, q_j \rangle = \hat{l}(j, i)\hat{\mu}(j), \end{aligned}$$

which implies reversibility and self-adjointness. □

4. Galerkin Approximation

We will now discuss Galerkin projections of transfer operators. For the sake of simplicity we will restrict our considerations to reversible Markov processes. Before we enter into the details of Galerkin projections we will shortly address the properties of the milestoning process induced by reversible Markov processes.

4.1. Galerkin projection and eigenvalues. In this section we will only consider discrete processes (X_n) . If (X_t) is time-continuous with generator \mathcal{L} , we will fix a lag time $\tau > 0$ and just consider the snapshot dynamics of $(X_{n\tau})$ with the semi-group of transfer operators (T_τ^n) . In this case the eigenvalues of the transfer operator T_τ will be given by

$$\lambda_{i,\tau} = e^{\Lambda_i\tau}, \tag{25}$$

where $\Lambda_i < 0$ is an eigenvalue of the generator \mathcal{L} . Now we want to approximate the dynamics of (X_n) by its projection to some low-dimensional subspace D in terms of density propagation. Therefore we will denote the orthogonal

projection onto D by Q . Assume that the process (X_n) is initially distributed according to

$$\rho_0(x)\mu(x) = \mathbb{P}[X_0 = x], \quad (26)$$

where $\rho_0(x)$ is a distribution with respect to μ . Then at any time n the distribution of X_n is given by

$$\rho_n(y)\mu(y) = \sum_x p_n(x, y)\rho_0(x)\mu(x) \quad (27)$$

or in matrix notation

$$\rho_n = T^n \rho_0. \quad (28)$$

Next, consider

$$\tilde{\rho}_n = Q\rho_n. \quad (29)$$

If we assume that $\rho_0 = Q\rho_0 \in D$ is a consistent initial distribution, i.e. it belongs to the subspace $D \subset S$, we find

$$\tilde{\rho}_n = QT^n Q\rho_0. \quad (30)$$

So the operator QT^nQ describes the propagation of the initial density $\rho_0 \in D$ to $\tilde{\rho}_n = Q\rho_n$, but we do not have a semi-group property anymore, i.e.

$$QT^nQ \neq (QTQ)^n. \quad (31)$$

Subsequently we will consider subspaces $D \subset L_\mu^2$ such that $\mathbf{1} \in D$, i.e., the invariant measure with density $\mathbf{1}$ in L_μ^2 is still contained in D . In this case we find in [19] an error bound for the approximation error from (31) $\|QT^nQ - (QTQ)^n\|$. We now cite Theorem 3.3 from section 3.4 of [19].

Theorem 4.1. *Let $T = T_\tau$ be a transfer operator of a time-continuous reversible Markov process with generator \mathcal{L} for lag time $\tau > 0$, or the transfer operator of some time-discrete reversible process. Let $1 = \lambda_0 > \lambda_1 > \dots > \lambda_{m-1}$ be the m dominant eigenvalues of T , i.e. for every other eigenvalue λ it holds $\lambda \leq r \leq \lambda_{m-1}$ such that r is the upper bound on the remaining spectrum. Furthermore, set $\eta = r/\lambda_1 < 1$. Whenever we have a generator, its eigenvalues Λ_i then satisfy: $\Lambda \in \text{spec}(\mathcal{L}), \Lambda \leq \Lambda_{m-1} \Rightarrow \Lambda \leq R < 0$ with $r = \exp(\tau R)$. Then, $\eta(\tau) = \exp(-\tau\Delta) < 1$ with*

$$\Delta = \Lambda_1 - R > 0, \quad (32)$$

as a τ -independent measure for the spread in the spectrum between the first non-trivial eigenvalue and the part of the spectrum that is not taken into account. Let u_0, u_1, \dots, u_{m-1} be the corresponding normalized eigenvectors. Let Q denote the projection onto some subspace $D \subset S$ with $\mathbf{1} \in D$ and define

$$\delta := \max_{j=1, \dots, m-1} \|Q^\perp u_j\| \quad (33)$$

where $Q^\perp = \text{Id} - Q$. Finally, define the projected transfer operator $P = QTQ$. Then the error $E(k)$ satisfies $\|QT^kQ - P^k\|_1 \leq E(k) = \|QT^kQ - P^k\|$ and is bounded from above by

$$E(k) \leq \min[2; C(\delta, \eta(\tau), k)] \cdot \lambda_1^k, \tag{34}$$

with a leading constant of following form

$$C(\delta, \eta, k) = ((m - 1)\delta + \eta)[C_{sets}(\delta, k) + C_{spec}(\eta, k)] \tag{35}$$

$$C_{sets}(\delta, k) = (m - 1)^{1/2}(k - 1) \delta \tag{36}$$

$$C_{spec}(\eta, k) = \frac{\eta}{1 - \eta}(1 - \eta^{k-1}). \tag{37}$$

The bound of Theorem 4.1 consists of two prefactors. C_{spec} depends on the lag time and the gap Δ in the spectrum of the generator. It will go to zero, if we increase the lag time τ , or, alternatively, the number m of eigenvectors that we have to approximate. The approximation or *projection error* δ of eigenvectors that we take into account governs the second part of the bound C_{sets} . More precisely, for fixed k , i.e., time span $k\tau$, the prefactor C_{sets} will be small, if the maximal projection error δ is small.

The next question is, how well the eigenvalues of the projected operator approximate the original eigenvalues of T . Because of self-adjointness of the transfer operator we can use the results from [33] to show

Theorem 4.2. *Let $1 = \lambda_0 > \lambda_1 > \dots > \lambda_{m-1}$ be the m dominant eigenvalues of T , i.e. for every other eigenvalue λ it holds $\lambda < \lambda_{m-1}$. Let u_0, u_1, \dots, u_{m-1} be the corresponding normalized eigenvectors, $D \subset S$ a subspace with*

$$\mathbf{1} \in D \quad \dim(D) =: n \geq m \tag{38}$$

and Q the orthogonal projection onto D .

Moreover, let $1 = \hat{\lambda}_0 > \hat{\lambda}_1 > \dots > \hat{\lambda}_{m-1}$ be the dominating eigenvalues of the projected operator QTQ . Then

$$E(\delta) = \max_{i=1, \dots, m-1} |\lambda_i - \hat{\lambda}_i| \leq \lambda_1(m - 1)\delta^2, \tag{39}$$

where

$$\delta = \max_{i=1, \dots, m-1} \|Q^\perp u_i\|$$

is the maximal projection error of the eigenvectors to the space D .

For the proof we refer to [34].

Remark 4.1. *Inserting (25) into (39), we get the lag time depended eigenvalue estimate*

$$E(\tau, \delta) = \max_{i=1, \dots, m-1} |\lambda_i - \hat{\lambda}_i| \leq e^{\Lambda_1 \tau}(m - 1)\delta^2, \tag{40}$$

where (λ_i) are the dominant eigenvalues of the transfer operator T_τ and $(\hat{\lambda}_i)$ the dominant eigenvalues of the projection $QT_\tau Q$.

Since $\Lambda_1 < 0$,

$$E(\tau, \delta) \rightarrow 0, \text{ for } \tau \rightarrow \infty. \tag{41}$$

Furthermore, for the relative eigenvalue error we have, at least for the first non-trivial eigenvalue

$$\frac{|\lambda_1 - \hat{\lambda}_1|}{|\lambda_1|} \leq (m - 1)\delta^2, \tag{42}$$

from which we see that by decreasing the maximal projection error we will have control even over the relative eigenvalue error.

4.2. Estimating the eigenvalues from trajectories. In this section we choose the special subspace D that is spanned by the committors associated with some core sets $C_1 \dots, C_n$, i.e. $D = \text{span}\{q_1, \dots, q_n\}$. Because $\mathbf{1} \in D$, Theorem 4.1 and 4.2 apply. Moreover we will see that this subspace allows us to compute the projected operator QTQ , its eigenvalues and all other related quantities from a trajectory. The first step is

Theorem 4.3. *Let D be a the subspace spanned by the committors*

$$D = \text{span}\{q_1, \dots, q_n\} \tag{43}$$

and let $\hat{\lambda}$ be an eigenvalue of the operator QTQ . Then $\hat{\lambda}$ solves the generalized eigenvalue problem

$$\hat{T}r = \hat{\lambda}Mr, \tag{44}$$

with

$$\hat{T}_{ij} = \frac{\langle q_i, Tq_j \rangle}{\hat{\mu}(i)}, \tag{45}$$

$\hat{\mu}(i) = \|q_i\|$, and the mass matrix

$$M_{ij} = \frac{\langle q_i, q_j \rangle}{\hat{\mu}(i)}. \tag{46}$$

Proof. Let D be as in (43). Then the orthogonal projection Q can be written as

$$(Qv)(y) = \sum_{i,j=1}^n S_{ij}^{-1} \langle v, q_j \rangle q_i, \tag{47}$$

with $S_{ij} = \langle q_i, q_j \rangle$.

Since

$$\hat{T}_{ij} = \frac{\langle q_i, Tq_j \rangle}{\hat{\mu}(i)} = \frac{\langle q_i, (Id + \mathcal{L}_d)q_j \rangle}{\hat{\mu}(i)} = (\hat{L}_d)_{ij} + M_{ij}, \tag{48}$$

(44) is equivalent to

$$\hat{L}_d r = (\hat{\lambda} - 1)Mr. \tag{49}$$

Let ϕ be an eigenvector of QTQ with respect to $\hat{\lambda}$, i.e.

$$\begin{aligned} QTQ\phi &= \hat{\lambda}\phi \Leftrightarrow Q(\mathcal{L}_d + Id)Q\phi = \hat{\lambda}\phi \\ \Leftrightarrow Q\mathcal{L}_dQ\phi &= (\hat{\lambda} - 1)\phi. \end{aligned}$$

This is equivalent to

$$\begin{aligned} \langle Q\mathcal{L}_dQ\phi, q_i \rangle &= (\hat{\lambda} - 1)\langle \phi, q_i \rangle \quad \forall i = 1, \dots, n \\ \Leftrightarrow \langle \mathcal{L}_dQ\phi, q_i \rangle &= (\hat{\lambda} - 1)\langle \phi, q_i \rangle \quad \forall i = 1, \dots, n \\ \Leftrightarrow \sum_{j,k=1}^n S_{jk}^{-1} \langle \phi, q_k \rangle \langle \mathcal{L}_dq_j, q_i \rangle &= (\hat{\lambda} - 1)\langle \phi, q_i \rangle \quad \forall i = 1, \dots, n. \end{aligned} \tag{50}$$

Introducing

$$r_j = \sum_{k=1}^n S_{jk}^{-1} \langle \phi, q_k \rangle$$

(50) can be written as

$$\sum_{j=1}^n r_j \langle \mathcal{L}_dq_j, q_i \rangle = (\hat{\lambda} - 1)\langle \phi, q_i \rangle = (\hat{\lambda} - 1) \sum_{j,k=1}^n S_{ij} S_{jk}^{-1} \langle \phi, q_k \rangle = (\hat{\lambda} - 1) \sum_{j=1}^n S_{ij} r_j. \tag{51}$$

Deviding both sides by $\hat{\mu}(i)$ completes the proof. □

Theorem 4.3 states, that we can compute the eigenvalues of the projected transfer operator QTQ by solving the generalized eigenvalue problem (44). In general, Theorem 4.3 does not depend on the special choice of D being the subspace spanned by the committors. The advantage is, that for D as in (43) the entries $\hat{l}_d(i, j)$ and M_{ij} have a stochastic interpretation: We have already seen that $\hat{L}_d = \hat{P} - Id$ with

$$\hat{p}(i, j) = \mathbb{P}[\hat{X}_{n+1} = j | \hat{X}_n = i].$$

As well-known, we can approximate the transition probabilities $\hat{p}(i, j)$ of the process (\hat{X}_n) from a (long enough) realization via the maximum likelihood estimator \hat{p}_N^* with entries

$$\hat{p}_N^*(i, j) = \frac{n_{ij}(N)}{N_i(N)},$$

where $n_{ij}(N)$ is the number of transition from i to j observed in the finite trajectory $\hat{X}_n, n = 0, \dots, N$, while $N_i(N) = \sum_j n_{ij}(N)$ is the total number of visits to state i in the trajectory. Since we are dealing with ergodic processes, we know that $\hat{P}_N^* \rightarrow \hat{P}$ in the limit of arbitrarily long trajectories, i.e., for $N \rightarrow \infty$ (law of large numbers).

Similarly, we can approximate the mass matrix M . We find

Lemma 4.4. *Let i, j be arbitrary and, as above, let $B_j = \bigcup_{k \neq j} C_k$ and let τ_A denote the first hitting time into set A . Then M_{ij} can be written as*

$$M_{ij} = \mathbb{P}[X_n \in C, \tau_{C_j} < \tau_{B_j} | \hat{X}_n = i],$$

that is, the probability to be outside of the core sets and enter the core set C_j next rather than any other core set, under the condition, that the last core set hit was C_i .

Thus, the entry M_{ij} of the mass matrix includes only those parts of (i, j) -reactive trajectories that are *outside* of core i and go to core j next, that is, in general, a typical (i, j) reactive trajectory will be much longer than those parts of it which enter into the mass matrix.

Proof. By definition

$$\begin{aligned} \mathbb{P}[X_n \in C, \tau_{C_j} < \tau_{B_j} | \hat{X}_n = i] &= \frac{\mathbb{P}[\hat{X}_n = i, X_n \in C, \tau_{C_j} < \tau_{B_j}]}{\mathbb{P}[\hat{X}_n = i]} \\ &= \sum_{x \in C} \mathbb{P}[\hat{X}_n = i, X_n = x] \mathbb{P}_x[\tau_{C_j} < \tau_{B_j}] \frac{1}{\hat{\mu}(i)} \\ &= \sum_{x \in C} \mathbb{P}[\hat{X}_n = i | X_n = x] \mathbb{P}[X_n = x] \mathbb{P}_x[\tau_{C_j} < \tau_{B_j}] \frac{1}{\hat{\mu}(i)} \\ &= \sum_{x \in C} q_i(x) \mu(x) q_j(x) \frac{1}{\hat{\mu}(i)} = \frac{\langle q_i, q_j \rangle}{\hat{\mu}(i)} = M_{ij}. \end{aligned}$$

□

Lemma 4.4 implies that we can estimate the mass matrix M by

$$M_N^*(i, j) = \frac{r_{ij}(N)}{R_i(N)}, \quad i \neq j$$

where $r_{ij}(N)$ is the total number of time steps during which the finite trajectory $\hat{X}_n, n = 0, \dots, N$ is reactive from i to j , i.e. the number of time steps the process spend in C coming from C_i and going to C_j , while $R_i(N)$ is the total number of time steps during which the finite trajectory resides in i , i.e., $\hat{X}_n = i$.

So we can estimate \hat{P} and M from a realization, i.e. a trajectory of the process (X_n) , compute \hat{T} by

$$\hat{T} \stackrel{(48)}{=} L_d + M = \hat{P} - Id + M \quad (52)$$

and solve the generalized eigenvalue problem in order to estimate the eigenvalues of the projected transfer operator QTQ .

Special case: full partition. When the core sets are chosen such that they form a full partition of state space (12), the definition of the committors directly yield

$$q_i(x) = \mathbb{1}_{C_i}(x). \tag{53}$$

That is, the committors are given by the characteristic functions on the coresets. This is exactly the standard MSM setting, such that the operator QTQ has a special interpretation, because

$$\hat{p}(i, j) = \mathbb{P}[X_n \in A_j | X_0 \in A_i] \tag{54}$$

is a matrix representation of the operator. Because of orthogonality of the stepfunctions we have

$$M_{ij} = \frac{\langle q_i, q_j \rangle}{\hat{\mu}(i)} = \begin{cases} 1, & i = j \\ 0, & i \neq j \end{cases}. \tag{55}$$

Now Theorem 4.2 states, that the eigenvalues of the matrix (54) are close to the eigenvalues of the transfer operator T , if the corresponding eigenvectors are well approximated by step-functions on the partitioning sets.

5. Illustrative Examples

5.1. Double well potential with diffusive transition region.

We consider the diffusion process

$$\gamma dX_t = -\nabla V(X_t)dt + \sqrt{2\beta^{-1}\gamma}dB_t \tag{56}$$

with B_t denoting Brownian motion in a potential V with two wells that are connected by an extended transition region. The potential V and its unique invariant measure μ are shown in Figure 1, we set the noise intensity $\sigma = \sqrt{2\beta^{-1}\gamma} = 0.8$ with $\gamma = 1$. We observe that the transition region between the two main wells contains four smaller wells that will have their own, less pronounced metastability each. The minima in the two main wells are located at $x_0 = -1$ and $x_1 = 6.62$, the respective saddle points that separate the main wells from the rest of the landscape at $x_0^\pm = x_0 \pm 1$, and $x_1^\pm = x_1 \pm 1$, respectively.

In order to find the transfer operator for this process we start with the Fokker-Planck equation $\partial_t u = \mathcal{L}u$, $u(t = 0, x) = f(x)$ that governs the propagation of a function f by the diffusion process. In the weighted Hilbert space L^2_μ the generator in the Fokker-Planck equation reads $\mathcal{L} = -\nabla V(x) \cdot \nabla_x + \beta^{-1}\Delta_x$, where ∇_x denotes the first derivative wrt. x and Δ_x the associated Laplacian. Thus, the transfer operator reads

$$T_t = \exp(t\mathcal{L}) \tag{57}$$

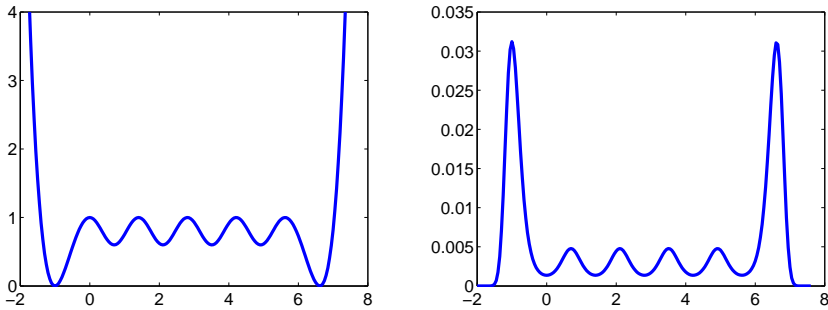


Figure 1. The potential V with extended transition region and the associated invariant measure for $\sigma = 0.8$.

This operator is self-adjoint since the diffusion process is reversible. The dominant eigenvalues of \mathcal{L} take the following values:

$$\begin{array}{cccccccc} \Lambda_0 & \Lambda_1 & \Lambda_2 & \Lambda_3 & \Lambda_4 & \Lambda_5 & \Lambda_6 & \Lambda_7 \\ +0.0000 & -0.0115 & -0.0784 & -0.2347 & -0.4640 & -0.7017 & -2.9652 & -3.2861 \end{array}$$

The main metastability has a corresponding implied timescale (ITS) $|1/\Lambda_1| \approx 88$ related to the transitions from one of the main wells to the other. Four other, minor metastable timescales related to the interwell switches between the main and the four additional small wells exist in addition. The eigenvalues have been computed by solving the eigenvalue problem for the partial differential operator \mathcal{L} by an adaptive finite element (FE) discretization with an accuracy requirement of $\text{tol} = 1e - 8$.

5.2. Two core sets.

In the following paragraphs we will compare the eigenvalues and ITS of the original process to the ones resulting from different MSM. More precisely, we first choose a lagtime τ and consider the transfer operator T_τ . Because of (25) we can compute the implied timescale

$$|1/\Lambda_1| = -\frac{\tau}{\ln(\lambda_{1,\tau})}, \quad (58)$$

where $\lambda_{1,\tau} < 1$ is the largest non-trivial eigenvalue of T_τ .

Next we choose two core sets of the form $C_0^s = (-\infty, x_0 + s]$ and $C_1^s = [x_1 - s, \infty)$ for some parameter s . Then we compare the ITS from (58) to the one, which corresponds to the largest non-trivial eigenvalue $\hat{\lambda}_{i,\tau}$ of the projected operator $QT_\tau Q$

$$|1/\hat{\Lambda}_1| = -\frac{\tau}{\ln(\hat{\lambda}_{1,\tau})}. \quad (59)$$

Since the process under investigation is just one-dimensional, we can compute the committor functions from the already mentioned FE discretization of \mathcal{L}

and just compute very accurate FE approximations of \hat{T}_τ and M , which allows to compute the eigenvalues of $QT_\tau Q$ as in Theorem 4.3. Figure 2 shows the dependence of the non-trivial eigenvalue on the core set size s for different values of the lagtime τ .

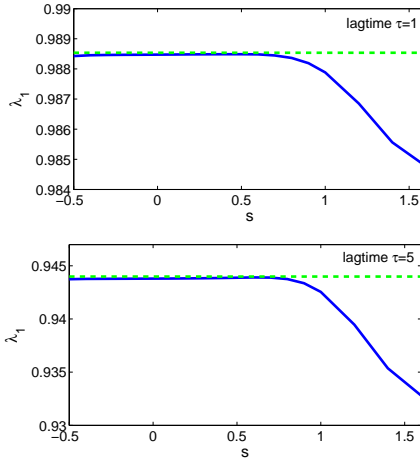


Figure 2. Non-trivial eigenvalues $\lambda_{1,\tau}^s < 1$ of the generalized eigenvalue problem $\hat{T}_\tau r = \hat{\lambda}Mr$ versus cores set size parameter s for lagtimes $\tau = 1$ (left) and $\tau = 5$ (right) in comparison to the exact first non-trivial eigenvalue $\exp(\tau\Lambda_1)$.

We observe that the for small enough core sets the approximation of the exact first non-trivial eigenvalue of T_τ , $\exp(\tau\Lambda_1)$, is good, while for too large core sets the approximation quality decreases. This can be understood since for $s > 1$ the core sets contain parts of the transition regions of the process where recrossing events lead to an overestimation of the transition probability between the cores. Moreover, Theorem 4.2 connected this error to the projection error $\|Q^\perp u_1\|$ and Figure 3 shows that this error behaves exactly like the approximation quality of the eigenvalues.

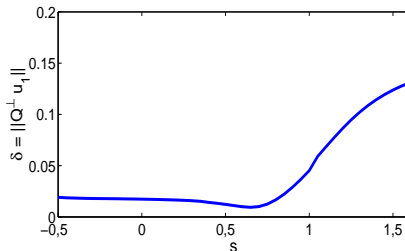


Figure 3. Projection error $\|Q^\perp u_1\|$ versus size of core sets, i.e., the parameter s .

Let us finally compare the effect of our choice of (two) core sets on the approximation error of dominant eigenvalues with the statements of Theorem 4.2

(with $m = 2$). To this end we will study the relative error

$$E_{rel}(\tau, \delta) = \frac{|\lambda_{1,\tau} - \hat{\lambda}_{1,\tau}|}{\lambda_{1,\tau}} \quad (60)$$

for different core set sizes s , see Figure 4. We observe that for small lagtimes the real relative error is significantly smaller than the upper bound (here given by the τ -independent square of the projection error $\delta = \|Q^\perp u_1\|$) but for larger lagtimes the upper bound and the real error are very close. As to be expected from Figure 3 the error for good core sets ($s = 0.5$) is two orders of magnitude smaller than the “not so good” core sets for $s = 2$.

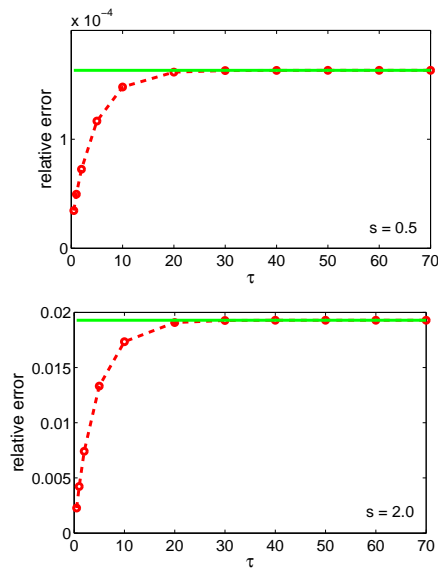


Figure 4. Relative error $E_{rel}(\tau, s)$ versus lagtime τ (dashed line) compared to the upper bound δ^2 given by Theorem 4.2 (solid line), for $s = 0.5$ (left hand panel) and $s = 2$ (right).

5.3. Estimation from data.

The computation of the committor functions will only be possible via FE discretization of the generator, which is infeasible in higher dimensions. This motivates to follow the instructions of Sec. 4.2 to estimate the eigenvalues from a trajectory.

We study the milestoning process $(\hat{X}_{n\tau})$ on state space $\{0, 1\}$ induced by the time-discrete process given by T_τ and the cores sets C_i^s , $i = 0, 1$.

Therefore we compute a very long trajectory $x(t)$, $t \in [0, t_{max}]$ of the diffusion process (for example based on Euler-Maruyama discretization of the SDE (56)). From this, we get discrete trajectories of the process $X_{n\tau}$ and of the

milestoning process $\hat{X}_{n\tau}$, $n = 0, \dots, N_\tau$ with $N_\tau = \lfloor t_{max}/\tau \rfloor$. This was done based on a trajectory $x(t)$ in the time interval $[0, t_{max}]$ with $t_{max} = 50000$. Then we can estimate \hat{T} and M by $\hat{T}_{N_\tau}^*$ and $M_{N_\tau}^*$ respectively as described in Sec. 4.2. The resulting non-trivial eigenvalues $\hat{\lambda}_1^*$ of the generalized eigenvalues problem $\hat{T}_{N_\tau}^* r = \hat{\lambda}^* M_{N_\tau}^* r$ are compared to the ones of $\hat{T}r = \hat{\lambda}Mr$ and to the exact first non-trivial eigenvalue $\lambda_1 = \exp(\tau\Lambda_1)$ in Figure 5.

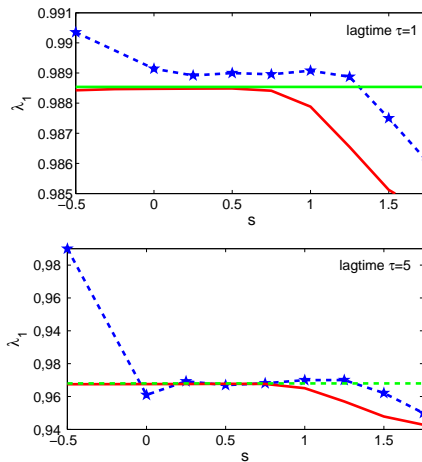


Figure 5. Comparison of the non-trivial eigenvalues λ_1^* of the trajectory-based generalized eigenvalues problem $\hat{T}_{N_\tau}^* r = \hat{\lambda} M_{N_\tau}^* r$ (dashed, stars), the ones of $\hat{T}r = \hat{\lambda}Mr$ (solid line) and the exact first non-trivial eigenvalue $\lambda_1 = \exp(\tau\Lambda_1)$ (flat line) in dependence on the core size parameter s for different lagtime $\tau = 1$ (left) and $\tau = 5$ (right).

We observe that the trajectory-based eigenvalues are overestimating the “exact” eigenvalues of the generalized eigenvalue problem, and that the approximation is getting worse for small values of s , especially for larger lagtimes. This is not surprising since for $s < 0$ and sparse undersampling of the trajectory for large lagtimes, we will miss events in which the process stays close to the minima x_i without entering the cores for some time which is not long compared to the lagtime.

Despite the good approximation quality of the trajectory-based generalized eigenvalues we should not forget that they are subject to an unknown statistical sampling error resulting from the finiteness of the trajectory. Assuming that the process (\hat{X}_n) is Markov and under additional assumptions on the prior [35, 36] one can show that the probability (density) that the given observation $\hat{X}_{n\tau}$, $n = 0, \dots, N_\tau$ results from the 2×2 stochastic transition matrix $P = p_{ij}$ is given by

$$\mathbb{P}(P|\hat{X}_{n\tau}) = p_{12}^{n_{12}}(1 - p_{12})^{n_{11}}p_{21}^{n_{21}}(1 - p_{21})^{n_{22}},$$

with $n_{ij} = n_{ij}(N_\tau)$ as defined in Sec.4.2. We have

$$\hat{P}_{N_\tau}^* = \operatorname{argmax}_{P \text{ stochastic matrix}} \mathbb{P}(P|\hat{X}_{n\tau}),$$

and for $N_\tau \rightarrow \infty$ this distribution is singularly supported in the “exact” transition matrix \hat{P}_τ of the milestoneing process.

Now, let $\nu = \nu(P)$ denote an arbitrary observable that is defined in terms of the transition matrix P , e.g., the first non-trivial eigenvalue $\nu(P) = \lambda_1(P)$ or the corresponding implied timescale $\operatorname{ITS}(P) = -\tau/\ln(\lambda_1(P))$. Then the pdf $\mathbb{P}(P|\hat{X}_{n\tau})$ on the transition matrix space and the corresponding pdf on the mass matrix space induce a pdf $\mathbb{P}(\nu|\hat{X}_{n\tau})$ on the state space of the observable. From this pdf we can compute a posteriori error indicators for the observable, e.g., the confidence intervals $I_\alpha(N_\tau)$ defined via

$$\mathbb{P}\left(\nu \in I_\alpha(N_\tau)|\hat{X}_{n\tau}\right) \geq \alpha.$$

In Figure 6, these confidence intervals are shown for the ITS for different values of s and $\tau = 1$.

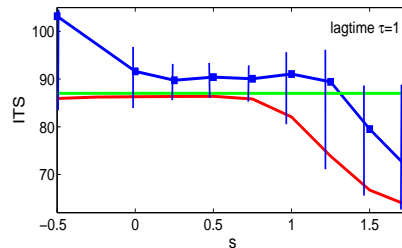


Figure 6. Implied timescale ITS and associated confidence interval I_α for $\alpha = 0.9$ of the trajectory-based generalized eigenvalue problem (solid) in comparison to the ITS of “exact” transition matrices \hat{P}_τ and to the ITS of the original process versus the cores size parameter s . Lagtime $\tau = 1$.

5.4. Full partition of state space. Let us fix $m = 2$ and observe how the relative eigenvalue error E_{rel} as defined in (60) above behaves in this case, especially how does it change for different full subdivisions of the state space and different lag times. From Theorem 4.2 we know that, as above, the bound on the relative eigenvalue error is given by the square of the projection error δ . First we choose $n = 2$ and the subdivision $A_1 = (-\infty, x]$ and $A_2 = (x, \infty)$. Figures 7 and 8 show the bound δ^2 compared to the relative error $E_{rel}(\tau, \delta)$, for two different subdivisions, i.e., different values of x . We can see that the error converges to δ^2 when increasing τ . Also, a better choice of the subdivision results not only in a smaller relative error, but in its faster convergence to the bound.

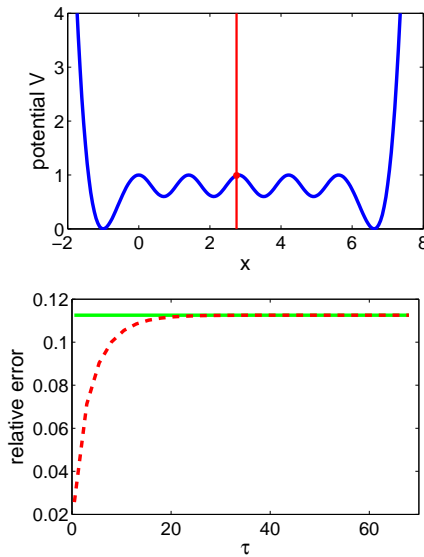


Figure 7. Relative error for eigenvalues and bound for $\tau = 0.5$, $n = 2$ and $x = 2.75$

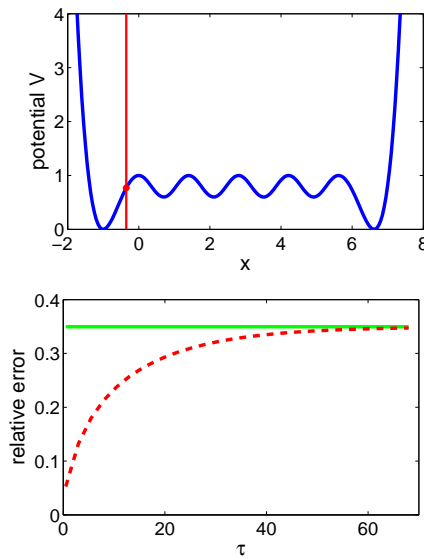


Figure 8. Relative error for eigenvalues and bound for $\tau = 0.5$, $n = 2$ and $x = -0.35$

Now we consider the full partition of a state space into $n = 6$ sets. The sets are chosen in such a way that every well belongs to one set. This choice of sets results in a smaller bound and faster convergence of the relative error to this bound, which can be seen in Figure 9.

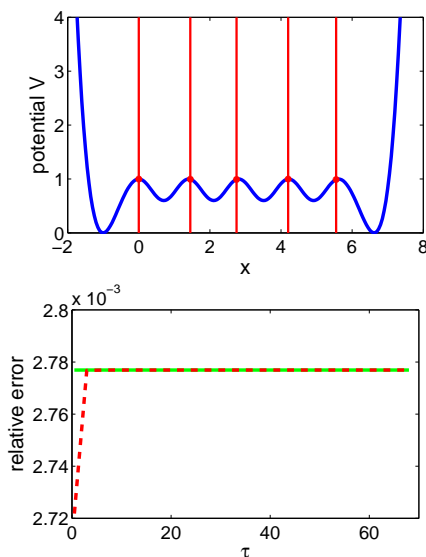


Figure 9. Relative error for eigenvalues and bound for $\tau = 0.5$ and $n = 6$

Let us finally compare the results for full subdivisions to the approximation via two core sets. We observe the following: Even the optimal full subdivision into $n = 2$ sets cannot compete with the approximation quality of the approximation based on two “reasonable/good” core sets. Good core sets result in an approximation error that is even better than the one for the optimal full subdivision into $n = 6$ sets which already resolves the well structure of the energy landscape. Thus, MSMs based on fuzzy ansatz spaces resulting from appropriate core sets and associated committor ansatz functions seem to lead to superior approximation quality than comparable full subdivision MSMs.

Conclusion

We presented a quite general approach to Markov State Models (MSM) via Galerkin projections to low-dimensional subspaces. We particularly considered the subspace spanned by the committor functions q_1, \dots, q_n defined by some core sets via the milestoning process. Our interpretation suggests that the method will work well if the space spanned by the eigenvectors corresponding to the dominant eigenvalues of the transfer operator T_t (or low-lying eigenvalues of the respective generator L) is well approximated by the subspace spanned by the committor functions. In this case, the Galerkin projection QTQ of the transfer operator $T = T_\tau$ associated with the lagtime chosen will approximate well the dominant eigenvalues of T , so that the long-time behavior will be captured,

see Theorem 4.2 as well as the propagation of functions by the dynamics, see Theorem 4.1.

Technically, these theorems *do not* require that the transfer operator of the original dynamics T possesses a *spectral gap*, i.e., a group of dominant eigenvalues which are separated from all the other ones by significant interval without eigenvalues. This is in partial contrast to the usual belief: The existence of a cluster of eigenvalues close to the largest eigenvalue $\lambda = 1$ and a spectral gap is often thought of as the fundamental condition under which MSMs can have good approximation quality.

Theorems 4.2 and 4.1 need a cluster of eigenvalues close to $\lambda = 1$ since this indicates that slow processes are taking place in the original state-space. These slow processes are what the generalized eigenvalue problem is meant to capture, in the sense that the generalized eigenvalues should be close to the small eigenvalues of the original process.

However, we do *not* need the existence of a spectral gap, at least not explicitly. What we need instead is that our committor functions are good approximations of the dominant eigenvectors, i.e., that the projection error δ is small. Since the committors depend on the choice of the core sets, smallness of the projection error can only be achieved for appropriately chosen core sets.

What our approximation theorems do not tell, however, is *how to choose* the core sets, because in general we will not be able to compute the dominant eigenvectors and committor functions that would be needed to identify the sets based on the above insight. If we assume that the original process has a cluster of eigenvalues close to 1 and a spectral gap, then general results guarantee the existence of a good collection of good core sets. What these sets are, however, is not given explicitly, except for the rather vague property that the process should oscillate inside and around each for a long time before visiting another and transitions to other core sets are significantly faster. How to use this criterion in a constructive way and whether a spectral gap is a necessary requirement here is the subject of current research, so we shall not dwell on these issues further here.

References

- [1] W.J. Stewart. *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, Princeton, NJ, 1994.
- [2] N. Madras and D. Randall. Markov chain decomposition for convergence rate analysis. *Annals of Applied Probability*, 12, 2002.
- [3] G. E. Cho and C. D. Meyer. Aggregation/disaggregation methods for nearly uncoupled Markov chains. *Technical Report NCSU no. 041600-0400, North Carolina State University*, 1999.
- [4] E. Meerbach, C. Schuette, and A. Fischer. Eigenvalue bounds on restrictions of reversible nearly uncoupled Markov chains. *Lin. Alg. Appl.*, 398, 2005.

-
- [5] C. D. Meyer. Stochastic complementation, uncoupling markov chains, and the theory of nearly reducible systems. *SIAM Rev.*, 31, 1989.
- [6] R.B. Mattingly. A revised stochastic complementation algorithm for nearly completely decomposable markov chains. *ORSA Journal on Computing*, 7(2), 1995.
- [7] P. Deuffhard, W. Huisinga, A. Fischer, and Ch. Schuette. Identification of almost invariant aggregates in reversible nearly uncoupled Markov chains. *Linear Algebra and its Applications*, 315:39–59, 2000.
- [8] P. Deuffhard and M. Weber. Robust perron cluster analysis in conformation dynamics. *Linear Algebra and its Applications*, 398 Special issue on matrices and mathematical biology:161–184, 2005.
- [9] Tijun Li, Weinan E, and Eric Vanden Eijnden. Optimal partition and effective dynamics of complex networks. *Proc. Nat. Acad. Sci.*, 105, 2008.
- [10] Ch. Schuette. *Conformational Dynamics: Modelling, Theory, Algorithm, and Applications to Biomolecules*. Habilitation thesis, Fachbereich Mathematik und Informatik, FU Berlin, 1998.
- [11] Ch. Schuette, A. Fischer, W. Huisinga, and P. Deuffhard. A direct approach to conformational dynamics based on hybrid Monte Carlo. *J. Comp. Physics Special Issue on Computational Biophysics*, 151:146–168, 1999.
- [12] Ch. Schuette and W. Huisinga. Biomolecular conformations can be identified as metastable sets of molecular dynamics. In *Handbook of Numerical Analysis*, pages 699–744. Elsevier, 2003.
- [13] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability and low lying spectra in reversible markov chains. *Comm. Math. Phys.*, 228:219–255, 2002.
- [14] F. Noe, Ch. Schütte, E. Vanden-Eijnden, L. Reich, and T. Weikl. Constructing the full ensemble of folding pathways from short off-equilibrium trajectories. *PNAS*, 106(45):19011–19016, 2009.
- [15] F. Noé, I. Horenko, Ch. Schuette, and J. Smith. Hierarchical analysis of conformational dynamics in biomolecules: Transition networks of metastable states. *J. Chem. Phys.*, 126:155102, 2007.
- [16] J. Chodera, N. Singhal, V. S. Pande, K. Dill, and W. Swope. Automatic discovery of metastable states for the construction of Markov models of macromolecular conformational dynamics. *Journal of Chemical Physics*, 126, 2007.
- [17] Nicaolae V. Buchete and Gerhard Hummer. Coarse master equations for peptide folding dynamics. *Journal of Physical Chemistry B*, 112:6057–6069, 2008.
- [18] A. C. Pan and B. Roux. Building Markov state models along pathways to determine free energies and rates of transitions. *Journal of Chemical Physics*, 129, 2008.
- [19] M. Sarich, F. Noé, and Ch. Schuette. On the approximation quality of markov state models. *to appear in Multiscale Modeling and Simulation*, 2010.
- [20] A. Voter. Introduction to the kinetic Monte Carlo method. In *Radiation Effects in Solids*. Springer, NATO Publishing Unit, Dordrecht, The Netherlands, 2005.
- [21] M. Freidlin and A. D. Wentzell. *Random perturbations of dynamical systems*. Springer, New York, 1998.

- [22] Weinan E and E. Vanden Eijnden. Metastability, conformation dynamics, and transition pathways in complex systems. In *Multiscale Modelling and Simulation*, pages 38–65. Springer, 2004.
- [23] A. Bovier, M. Eckhoff, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes. I. sharp asymptotics for capacities and exit times. *J. Eur. Math. Soc. (JEMS)*, 6:399–424, 2004.
- [24] A. Bovier, V. Gayrard, and M. Klein. Metastability in reversible diffusion processes. II. precise asymptotics for small eigenvalues. *J. Eur. Math. Soc. (JEMS)*, 7:69–99, 2005.
- [25] W. Huisinga, S. Meyn, and Ch. Schuette. Phase transitions and metastability for Markovian and molecular systems. *Ann. Appl. Probab.*, 14:419–58, 2004.
- [26] E. B. Davies. Spectral properties of metastable markov semigroups. *J. Funct. Anal.*, 52:315–329, 1983.
- [27] M. Weber, S. Kube, L. Walter, and P. Deuffhard. Stable computation of probability densities for metastable dynamical systems. *Mult. Mod. Sim.*, 6(2):396–416, 2007.
- [28] Anton K. Faradjian and Ron Elber. Computing time scales from reaction coordinates by milestoning. *J. Chem. Phys.*, 120:10880–10889, 2004.
- [29] Weinan E and E. Vanden-Eijnden. Towards a theory of transition paths. *Journal of statistical physics*, 123:503–523, 2006.
- [30] P. Metzner, Ch. Schuette, and E. Vanden-Eijnden. Transition path theory for markov jump processes. *Multiscale Modeling and Simulation*, 7(3):1192–1219, 2009.
- [31] P. Metzner, Ch. Schuette, and E. Vanden-Eijnden. Illustration of transition path theory on a collection of simple examples. *J. Chem. Phys.*, 125, 2006.
- [32] Ch. Schütte, F. Noe, J. Lu, M. Sarich, and E. Vanden-Eijnden. Markov state model building using milestoning. *submitted to J. Chem. Phys.*, 2010. Preprint download via <http://www.math.fu-berlin.de/groups/biocomputing/publications/index.html>.
- [33] A.V. Knyazev and M. E. Argentati. Rayleigh-ritz majorization error bounds with applications to fem. *SIAM Journal on Matrix Analysis and Applications*, 31:1521, 2010.
- [34] N. Djurdjevac, M. Sarich, and Ch. Schütte. Estimating the eigenvalue error of Markov state models. *submitted to Mult. Mod. Sim.*, 2010. Preprint download via <http://www.math.fu-berlin.de/groups/biocomputing/publications/index.html>.
- [35] Ph. Metzner, F. Noé, and Ch. Schütte. Estimating the sampling error: Distribution of transition matrices and functions of transition matrices for given trajectory data. *Phys. Rev. E*, 2008.
- [36] F. Noé. Probability distributions of molecular observables computed from Markov models. *J. Chem. Phys.*, 128:244103, 2008.

Second Order Backward SDEs, Fully Nonlinear PDEs, and Applications in Finance

Nizar Touzi*

Abstract

The martingale representation theorem in a Brownian filtration represents any square integrable r.v. ξ as a stochastic integral with respect to the Brownian motion. This is the simplest Backward SDE with nul generator and final data ξ , which can be seen as the non-Markov counterpart of the Cauchy problem in second order parabolic PDEs. Similarly, the notion of Second order BSDEs is the non-Markov counterpart of the fully-nonlinear Cauchy problem, and is motivated by applications in finance and probabilistic numerical methods for PDEs.

Mathematics Subject Classification (2010). Primary 60H10; Secondary 60H30.

Keywords. Backward stochastic differential equations, stochastic analysis, non-dominated mutually singular measures, viscosity solutions of second order PDEs.

1. Introduction

The theory of backward stochastic differential equations (BSDE hereafter) received a considerable attention in the recent literature. The ongoing developments are motivated by financial mathematics, stochastic control, stochastic differential games, and probabilistic numerical methods for partial differential equations (PDEs hereafter). We refer to [12] for a review.

*Based on a long collaboration with Mete Soner, Jianfeng Zhang, Patrick Cheridito, and Bruno Bouchard. Research supported by the Chair *Financial Risks* of the *Risk Foundation* sponsored by Société Générale, the Chair *Derivatives of the Future* sponsored by the Fédération Bancaire Française, and the Chair *Finance and Sustainable Development* sponsored by EDF and Calyon.

Ecole Polytechnique Paris, CMAP, Route de Saclay, 91128 Palaiseau Cedex, France.
E-mail: nizar.touzi@polytechnique.edu.

These notes provide an overview on the recent extension to the second order which correspond to second order PDEs. Our objective is to define second order BSDEs in the general non Markov case, which can be viewed as the natural counterpart of PDEs in the non Markovian framework. We put a special emphasis on the examples, mainly from financial mathematics, which acted as a driving line for the progress which was achieved.

Section 2 provides a quick review of the basics of standard BSDEs and their connection to semilinear PDEs. We also provide a non-expert exposition of the main applications in financial mathematics.

In Section 3, we report our main example of hedging under gamma constraints, which show the main difficulties that one has to solve. The main result of this section is the uniqueness result of [8] obtained within a restricted class of integrands.

Section 4 provides a new definition of solutions of 2BSDE motivated by the quasi-sure stochastic analysis developed by Denis and Martini [10] in the context of their analysis of the uncertain volatility model.

Section 5 collects the mains results of these notes, mainly the wellposedness of the quasi-sure formulation of the 2BSDE. We state a representation result which implies uniqueness. With the representation result, comparison becomes trivial. Then, we provide the appropriate a priori estimates. Finally, existence is obtained as follows. First for bounded uniformly continuous final data, the representation suggest a natural candidate for the solution of the 2BSDE, that can be defined by means of the notion of regular conditional probability density. Then, using the a priori estimates, we prove the existence of a solution in an appropriate closure of the space of bounded uniformly continuous random variables. Finally in the Markovian case, under natural condition, the solution of the 2BSDE is a viscosity solution of the corresponding fully nonlinear PDE.

Notations: Scalar products will be denotes by dots, and transposition of matrices by an exponent \top . For a σ -algebra \mathcal{F} , a filtration \mathbb{F} , and a probability measure \mathbb{P} , we will denote

- $\mathbb{L}^2(\mathcal{F}, \mathbb{P})$, the set of \mathcal{F} -measurable r.v. with finite second moment under \mathbb{P} ,
- $\mathbb{H}^2(\mathbb{F}, \mathbb{P})$, the set of all \mathbb{F} -progressively measurable processes H with $\mathbb{E}[\int |H_t|^2 dt] < \infty$,
- $\mathbb{S}^2(\mathbb{F}, \mathbb{P})$, the subset of $\mathbb{H}^2(\mathbb{F}, \mathbb{P})$ with \mathbb{P} -a.s. càdlàg sample paths.

2. Review of Standard Backward SDEs

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space supporting a Brownian motion W on \mathbb{R}^d , and denote by $\mathbb{F} = \{\mathcal{F}_t, t \geq 0\}$ the corresponding \mathbb{P} -augmented canonical filtration.

Consider the two ingredients:

- the generator $F : \mathbb{R}_+ \times \Omega \times \mathbb{R} \times \mathbb{R}^d \rightarrow \mathbb{R}$ is such that $(t, \omega) \mapsto F_t(\omega, y, z)$ is \mathbb{F} -progressively measurable for all $(y, z) \in \mathbb{R} \times \mathbb{R}^d$,
- the final data $\xi \in \mathbb{L}^2(\mathbb{P}, \mathcal{F}_T)$ for some time horizon $T > 0$.

Given a time horizon $T > 0$, a (scalar) backward stochastic differential equation (BSDE in short) is defined by:

$$Y_t = \xi + \int_t^T F_s(Y_s, Z_s) ds - \int_t^T Z_s \cdot dW_s, \quad t \leq T. \tag{1}$$

Equations of this type appeared naturally in the work of Bismut [5] on the stochastic maximum Pontryagin principle for stochastic control problems. A systematic study was started by Pardoux and Peng [18], where an existence and uniqueness theory of an \mathbb{F} -progressively measurable solution (Y, Z) was introduced. This seminal work generated an extensive literature in stochastic analysis, with natural motivations from financial mathematics.

In this section, we provide a quick review of this theory under the condition

$$F \text{ Lipschitz-continuous in } (y, z) \text{ uniformly in } (t, \omega) \tag{2}$$

2.1. The linear case. Consider first the case $F \equiv 0$:

$$Y_t = \xi - \int_t^T Z_s \cdot dW_s, \quad t \leq T. \tag{3}$$

Then, for any $\xi \in \mathbb{L}^2(\mathbb{P}, \mathcal{F}_T)$, there is a unique \mathbb{F} -progressively measurable square integrable process Y satisfying (1), given by $Y_t := \mathbb{E}[\xi | \mathcal{F}_t]$, $t \leq T$. Moreover, by the martingale representation theorem in the present Brownian filtration, the process Y can be considered in its continuous version, and there exists a unique \mathbb{F} -progressively measurable square integrable process Z satisfying (1). By the Doob’s maximal inequality, this construction provides a unique solution (Y, Z) of (1) in the space $\mathbb{S}^2(\mathbb{P}, \mathbb{F}) \times \mathbb{H}^2(\mathbb{P}, \mathbb{F})$, i.e.

$$\mathbb{E}[\sup_{t \leq T} |Y_t|^2] + \mathbb{E}[\int_0^T |Z_t|^2 dt] < \infty. \tag{4}$$

We next consider the linear case

$$F_t(y, z) = -k_t y + \lambda_t \cdot z + \alpha_t, \tag{5}$$

for some \mathbb{F} -progressively measurable processes k, λ, α , that we assume to be bounded, for simplicity. Defining

$$\tilde{Y}_t := Y_t e^{-\int_0^t k_s ds}, \quad t \in [0, T], \quad \text{and} \quad \tilde{\xi} := \xi e^{-\int_0^t k_s ds} + \int_0^T \alpha_s e^{-\int_0^s k_u du} ds, \tag{6}$$

we can convert the BSDE (1) into a BSDE with nul generator under the equivalent probability measure

$$\left. \frac{d\mathbb{Q}}{d\mathbb{P}} \right|_{\mathcal{F}_T} := e^{\int_0^T \lambda_t \cdot dW_t - \frac{1}{2} \int_0^T |\lambda_t|^2 dt}. \tag{7}$$

Example Hedging contingent claims in frictionless financial markets. Consider a financial market consisting of d risky assets with price processes:

$$dS_t = \text{diag}[S_t] (b_t dt + \sigma_t dW_t), \tag{8}$$

where $\text{diag}[S_t]$ denotes the diagonal matrix with diagonal entries S_t^i , and b, σ, σ^{-1} are \mathbb{F} -progressively measurable bounded processes.

- A portfolio strategy is an \mathbb{F} -progressively measurable process $\{\theta_t, t \in [0, T]\}$ with values in \mathbb{R}^d . Here each component θ_t^i indicates the amount invested in asset S^i at time t . The self-financing condition defines the dynamics of the liquidation value of the portfolio:

$$dV_t = \sum_{i=1}^d \theta_t^i \frac{dS_t^i}{S_t^i} + \left(V_t - \sum_{i=1}^d \theta_t^i \right) r_t dt, \tag{9}$$

where the instantaneous interest rate r is \mathbb{F} -progressively measurable and bounded. The latter equation is the budget constraint which says that the change in the liquidation value of the portfolio has two components. First, for each asset i the change of value of the holding in asset S^i is given by the change of the corresponding price times the number of shares of this asset held in portfolio at time t . The difference $V_t - \sum_{i=1}^d \theta_t^i$ represents the holding in cash on the bank account. Then the second component of the above budget constraint simply says that this investment in the bank has an instantaneous riskless return defined by the instantaneous interest rate.

- A portfolio strategy θ is admissible if $\sigma^T \theta \in \mathbb{H}^2(\mathbb{P}, \mathbb{F})$, so that the process V is well-defined in $\mathbb{H}^2(\mathbb{P}, \mathbb{F})$. We denote by V^θ the corresponding liquidation value process.

- A European contingent claim is a r.v. $\xi \in \mathbb{L}^2(\mathbb{P}, \mathcal{F}_T)$ which indicates the random payoff of a contract between two parties. The seller of such a contract bears the risk of the random payment, and wishes to hedge his position against the bad states of the world. A natural problem is then to

$$\text{Find an admissible portfolio } \theta \text{ so that } V_T^\theta = \xi, \mathbb{P} - \text{a.s.} \tag{10}$$

This is a BSDE problem with final data ξ , and affine generator $F_t(y, z) = -r_t y - (\sigma^T)^{-1}(b_t - r_t \mathbf{1}) \cdot z$, where $\mathbf{1}$ is the vector of ones in \mathbb{R}^d .

2.2. Wellposedness of Backward SDEs. Next, let F be a generator satisfying (2) and denote $F_t^0 := F_t(0, 0)$. Then assuming $\xi \in \mathbb{L}^2(\mathbb{P}, \mathcal{F}_T)$ and

$F^0 \in \mathbb{H}^2(\mathbb{P}, \mathbb{F})$, it follows from a fixed point argument that the BSDE (1) has a unique solution in $\mathbb{S}^2(\mathbb{P}, \mathbb{F}) \times \mathbb{H}^2(\mathbb{P}, \mathbb{F})$.

When the generator is either convex or concave, the solution of the BSDE corresponds to a stochastic control problem in standard form but without diffusion control.

Various extensions of this result have been obtained in the previous literature by weakening the Lipschitz condition (2). The most challenging is probably the case where F has quadratic growth in z , see Kobylanski [16] and Tevzadze [24].

A comparison result is easily obtained, and reads as follows. Suppose that (F, ξ) and (F', ξ') satisfy the above conditions for the existence and uniqueness of solutions (Y, Z) and (Y', Z') of the corresponding BSDEs. Assume that $\xi \leq \xi'$ and $f_t(Y_t, Z_t) \leq f'_t(Y_t, Z_t)$. Then $Y \leq Y'$ on $[0, T]$, \mathbb{P} -a.s.

Such a comparison result plays a central role in the theory. For instance, it allows to define the notion of reflected BSDEs (a misleading denomination, to which I prefer the name of *obstacle BSDE*) which are connected to optimal stopping problems and Dynking games.

Example: *Hedging under different borrowing and lending rates.* Let us turn to the example of the previous subsection. The holding in cash $V_t - \theta_t \cdot \mathbf{1}$ can be either positive, meaning a positive amount on the bank account, or negative, meaning a loan from the bank. In the real life, borrowing and lending rates are different and are given respectively by $\bar{r}_t \geq \underline{r}_t$. Then, the dynamics of the liquidation value of the portfolio (9) is replaced by:

$$dV_t = \sum_{i=1}^d \theta_t^i \frac{dS_t^i}{S_t^i} + ((V_t - \theta_t \cdot \mathbf{1})^+ \underline{r}_t - (V_t - \theta_t \cdot \mathbf{1})^- \bar{r}_t) dt, \tag{11}$$

which is our simplest example of nonlinear BSDE.

2.3. Markov BSDEs. The Markov case corresponds to the particular specification

$$F_t(\omega, y, z) = f(t, X_t(\omega), y, z) \quad \text{and} \quad \xi = g(X_T(\omega)) \tag{12}$$

where X is the solution of some (well-posed) stochastic differential equation

$$X_t = X_0 + \int_0^t b(s, X_s) ds + \int_0^t \sigma(s, X_s) dW_s, \quad t \leq T. \tag{13}$$

Moving the time origin to an arbitrary $t \in [t, T]$, we denote by $\{X_s^{t,x}, s \in [t, T]\}$ the solution of the above SDE with initial data $X_t^{t,x} = x$, and by $\{(Y_s^{t,x}, Z_s^{t,x}), s \in [t, T]\}$ the solution of the corresponding BSDE. Then, since the Brownian motion has independent increments and is translation invariant, we easily see that

$$u(t, x) := Y_t^{t,x}, \quad t \in [0, T], \quad x \in \mathbb{R}^d, \tag{14}$$

defines a deterministic function satisfying the semigroup property (or the dynamic programming principle, in the language of stochastic control):

$$u(s, X_s^{t,x}) = Y_s^{t,x} = u(t, x) + \int_t^s f(r, X_r^{t,x}, Y_r^{t,x}, Z_r^{t,x})dr - \int_t^s Z_r^{t,x}dW_r \quad (15)$$

Then, if u is $C^{1,2}$, it follows that $Z_t^{t,x} = \sigma^T(t, x)Du(t, x)$, and u is a classical solution of the semilinear Cauchy problem:

$$-\partial_t u - \frac{1}{2}\text{Tr} [\sigma\sigma^T D^2 u] - f(t, x, u, Du) = 0, u(T, \cdot) = g. \quad (16)$$

Of course, this equation can be derived in the sense of viscosity solutions when no regularity of u is available.

2.4. Numerical implications. From the latter connection with the Cauchy problem, one can formulate an extension of the so-called Feynman-Kac representation formula to the semilinear case, which states that whenever the Cauchy problem (16) has a classical solution u , then it has a representation (14) in terms of a corresponding BSDE. Among the various applications of this representation, I would like to highlight its numerical implications.

1. The case of a null generator $f \equiv 0$ is well known to open the door to probabilistic numerical methods for the approximation of the solution of (16). Indeed, in this case, the BSDE representation reduces to $u(t, x) = Y_t^{t,x} = \mathbb{E}[g(X_T^{t,x})]$ which suggests an approximation based on the law of large numbers. For instance, one can generate independent copies of the r.v. $g(X_T^{t,x})$ (or an appropriate approximation), and define the crude Monte Carlo approximation by simple averaging. A remarkable feature of this approximation is that the rate of convergence, as provided by the central limit theorem, is independent of the dimension d of the state x . This represents a clear advantage of probabilistic schemes.
2. For a general nonlinearity f , let $\pi : t = t_0 < \dots < t_n = T$ be a partition of the interval $[t, T]$ with time steps $\delta t_k := t_k - t_{k-1}$, and corresponding increments of the Brownian motion $\delta W_{t_k} := W_{t_k} - W_{t_{k-1}}$. Denote by X^π the Euler discretization of X along the partition π . The following discretization of (1) was suggested by Bally and Pagès [1] when f does not depend on z , and independently by Bouchard Touzi [6] and Zhang [25] for a general nonlinearity:

$$Y_{t_n}^\pi = g(X_{t_n}^\pi), \quad (17)$$

and

$$Y_{t_{k-1}}^\pi = \mathbb{E}[Y_{t_k}^\pi | X_{t_{k-1}}^\pi] + \delta t_k f(t_{k-1}, X_{t_{k-1}}^\pi, Y_{t_{k-1}}^\pi, Z_{t_{k-1}}^\pi), \quad (18)$$

$$Z_{t_{k-1}}^\pi = \mathbb{E}[Y_{t_k}^\pi (\delta t_k \sigma(t_k, X_{t_k}^\pi))^{-1} \delta W_{t_k} | X_{t_{k-1}}^\pi]. \quad (19)$$

For a feasible scheme, one further needs to introduce an implementable approximation of the regression operator $\mathbb{E}[\cdot | X_{t_{k-1}}^\pi = x]$. Convergence results of the discrete-time process (Y^π, Z^π) towards the solution (Y, Z) of the Markov BSDE, together with bounds on the rate of convergence are available in the literature, see [6, 14, 9]. Notice however that the asymptotic results in the present nonlinear case depend on the dimension of the state d .

3. Second Order BSDEs: Difficulties and Intuitions

Backward stochastic differential equation are naturally connected to semilinear PDEs of the form (16), i.e. linear dependence of the equation in terms of the hessian matrix. The first objective of the notion of second order BSDEs is to enlarge the notion of BSDEs so as to obtain a connection with fully nonlinear PDEs. This allows to capture more interesting examples. In this section, we provide a simple example which is beyond the scope of standard BSDEs. moreover, this example reveals the difficulty we are facing for our extension.

3.1. Hedging under Gamma constraints. Let us specialize the example of Subsection 2.1 to the one-dimensional case $d = 1$. Denote $\pi_t := \frac{\theta_t}{S_t}$ the number of shares of S held in portfolio at time t , and $V^\pi := V^\theta$. The practice of the optimal hedging strategy induced by this model leads to a portfolio adjustment at each time t from π_t to π_{t+dt} , i.e. the investor has to buy or sell (depending on the sign) $\pi_{t+dt} - \pi_t$ shares of the asset S . Although our model assumes that the price process is exogeneous, practioners are fully aware of the nonlinear dependence of the price in terms of the transaction volume, and the impact of their strategies on the price process. This is the so-called illiquidity effect.

To avoid (or at least minimize) such illiquidity costs, we assume that π_t is a continuous semimartingale with

$$d\langle \pi, S \rangle_t = \Gamma_t \langle S \rangle_t, \mathbb{P} - \text{a.s.} \quad (20)$$

and we impose some constraints on the process Γ . In fact, the interpretation of Γ , as viewed by practitioners, is the portfolio adjustment consequent to an immediate jump of the underlying price process. Although jumps are not allowed by the model, this is a conservative behavior aiming at building strategies which are robust to such a specification error of the model.

Given a contingent claim $\xi \in \mathbb{L}^2(\mathbb{P}, \mathcal{F}_T)$, our new hedging problem is now:

$$\text{Find an admissible portfolio } \pi \text{ so that } \Gamma \in [\underline{\Gamma}, \bar{\Gamma}] \text{ and } V_T^\pi = \xi, \mathbb{P} - \text{a.s.} \quad (21)$$

where $\underline{\Gamma} < 0 < \bar{\Gamma}$ are given.

We also observe that in the Markov framework, “we expect” that Γ_t should identify the Hessian matrix of the function u defined in (14). Then, this problem is expected to be connected to a fully nonlinear PDE.

However, there is a fundamental difficulty related to the following result due to Bank and Baum [2].

Lemma 3.1. *Let ϕ be a progressively measurable process with $\int_0^T |\phi_t|^2 dt < \infty$, \mathbb{P} -a.s. Then, for every $\varepsilon > 0$, there exists a progressively measurable process ϕ^ε , absolutely continuous with respect to the Lebesgue measure, with $\int_0^T |\phi_t^\varepsilon|^2 dt < \infty$, and*

$$\sup_{0 \leq t \leq T} \left\| \int_0^t \phi_t \cdot dW_t - \int_0^t \phi_t^\varepsilon \cdot dW_t \right\|_\infty \leq \varepsilon. \quad (22)$$

This result shows a high instability of the problem: by accepting to miss the target ξ within a small range of ε , we may approximate the optimal hedging strategy of the frictionless financial market (Subsection 2.1) so that the Gamma process of the approximation is zero !

3.2. Non-uniqueness in \mathbb{L}^2 . The latter difficulty which appears naturally in the context of the financial application is not exceptional. Let us consider the simplest backward SDE problem involving the Gamma process, similar to the above example:

$$Y_t = c \int_t^T \Gamma_s ds - \int_t^T Z_s dW_s \quad \text{where} \quad d\langle Z, W \rangle_t = \Gamma_t dt, \quad t \in [0, T], \quad \mathbb{P} - \text{a.s.} \quad (23)$$

Obviously, $Y = Z = \Gamma = 0$ is a solution. However, if we admit any square integrable semimartingale Z with square integrable corresponding Γ process, it is shown in Example 6.1 of [23] that, except for the case $c = 0$, the above problem has a non-zero solution.

Consequently, introducing a second order term in the BSDE can not be performed within the classical framework, and one has to face the difficulties due to the instability highlighted in Lemma 3.1. This is the main object of these notes which was dealt with by to approaches

- the first approach, developed in the subsequent subsection 3.3, is to restrict the process Z to an appropriate space, so as to obtain uniqueness. This approach was successful for uniqueness in the Markov framework, but we were not able to have a satisfactory existence theory.
- the second approach is motivated by the example of Subsection 3.4 below, and consists in reinforcing the constraint by requiring the BSDE to be satisfied on a bigger support... This is the content of Section 4 below which contains our main wellposedness results of second order BSDEs.

3.3. A first uniqueness result. In order to involve the process Γ in the problem formulation, we need that the process Z be a semimartingale. Then, we have the following correspondence between the Itô and the Fisk-Stratonovich integrals

$$a \int_0^t Z_t \cdot dW_t = \frac{1}{2} \Gamma_t dt + \int_0^t Z_t \circ dW_t. \tag{24}$$

a We prefer to write the problem using the Fisk-Stratonovich stochastic integral rather than the Itô one. In the present subsection, this is just cosmetic, but it will play a crucial role in Section 4.

Consider the Markov 2BSDE:

$$Y_t = g(X_T) + \int_t^T h(s, X_s, Y_s, Z_s, \Gamma_s) ds - \int_t^T Z_s \circ dX_s, \quad \mathbb{P} - \text{a.s.} \tag{25}$$

where X is defined by the stochastic differential equation

$$dX_t = b(t, X_t)dt + \sigma(t, X_t)dW_t, \tag{26}$$

that we assume wellposed with support in the whole space \mathbb{R}^d .

An appropriate class \mathcal{Z} of processes Z is introduced in [8]. Since we will be mainly concentrating on the alternative approach, we do not report the precise description of this class in these notes. To prove the uniqueness result, we introduce the stochastic target problems

$$\mathcal{V}(0, X_0) := \inf \{Y_0 : Y_T \geq g(X_T), \mathbb{P} - \text{a.s. for some } Z \in \mathcal{Z}\}, \tag{27}$$

$$\mathcal{U}(0, X_0) := \sup \{Y_0 : Y_T \leq g(X_T), \mathbb{P} - \text{a.s. for some } Z \in \mathcal{Z}\}. \tag{28}$$

By moving the time origin to an arbitrary $t \in [0, T]$, we also define the value functions $\mathcal{V}(t, s)$ and $\mathcal{U}(t, x)$ for all $(t, x) \in [0, T] \times \mathbb{R}^d$. The following result is obtained in [8] by proving that \mathcal{V} and \mathcal{U} are respectively viscosity supersolution and subsolution of the (fully nonlinear) dynamic programming equation:

$$-\partial_t v - h(t, x, v, Dv, D^2v) = 0 \text{ on } [0, T] \times \mathbb{R}^d, \quad \text{and} \quad v(T, \cdot) = g. \tag{29}$$

Theorem 3.2. *Let h be continuous, locally Lipschitz in y , uniformly in all other variables, non-increasing in γ , and has polynomial growth in (x, y, z, γ) . Let g be continuous with polynomial growth. Assume further that the nonlinear PDE (29) satisfies a comparison result in the sense of viscosity solutions, within the class of polynomially growing functions. Then there is at most one solution to the backward SDE (25) with $Z \in \mathcal{Z}$.*

3.4. Intuition from uncertain volatility models. The objective of this example is to introduce uncertainty about the volatility process σ in our first example of Subsection 2.1. To do this, we reformulate the problem in the setting of the canonical space $\Omega = \{\omega \in C([0, T]) : \omega(0) = 0\}$ as suggested

by Denis and Martini [10]. We denote by B be the coordinate process, \mathbb{F} the corresponding canonical filtration, and \mathbb{P}_0 the Wiener measure, so that B is a Brownian motion under \mathbb{P}_0 .

By obvious discounting, we may reduce to the zero interest rate case. Moreover, after an equivalent change of measure, we may also assume without loss of generality that $b = 0$. The liquidation value process (9) is then given by:

$$V_t := V_0 + \int_0^t \theta_s \cdot dB_s, \quad (30)$$

where the volatility coefficient can be viewed to be absorbed into the canonical process by a time change argument. To model the uncertainty on the volatility, we consider two given constants $0 < \underline{a} \leq \bar{a}$, and we introduce the set $\mathcal{P} = \mathcal{P}_{\underline{a}, \bar{a}}$ of all probability measures on Ω such that B is a martingale under \mathbb{P} with quadratic variation absolutely continuous with respect to Lebesgue, and

$$\underline{a} \leq \frac{d\langle B \rangle_t}{dt} \leq \bar{a}, \quad t \in [0, T]. \quad (31)$$

Notice that the family \mathcal{P} has no dominating measure, and all measures contained therein are mutually singular. Since the stochastic integral is defined \mathbb{P} -a.s. for all $\mathbb{P} \in \mathcal{P}$, it is not clear how to define the liquidation value V in (30) simultaneously under every $\mathbb{P} \in \mathcal{P}$. This is achieved in [10] by revisiting the stochastic integration theory, replacing the reference probability measure by the capacity

$$c(A) := \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}[A] \quad \text{for all } A \in \mathcal{F}_T. \quad (32)$$

An event A is said to be *polar* if $c(A) = 0$, and a property is said to hold *quasi-surely* (q.s. hereafter) if it holds on the complement of a polar set. The first main contribution of [10] is to isolate a set of integrands \mathcal{H} , such that the stochastic integral (30) with $\theta \in \mathcal{H}$ is defined quasi-surely, i.e. \mathbb{P} -almost surely for all $\mathbb{P} \in \mathcal{P}$.

The superhedging problem can now be formulate rigorously:

$$\mathcal{V}(\xi) := \inf \{V_0 : V_T \geq \xi, \text{ q.s. for some } \theta \in \mathcal{H}\}. \quad (33)$$

This is weaker than the BSDE problem as existence is not required in the formulation (33). The main result of [10] is the following dual formulation of this problem:

$$\mathcal{V}(\xi) = \sup_{\mathbb{P} \in \mathcal{P}} \mathbb{E}^{\mathbb{P}}[\xi], \quad (34)$$

for random variables ξ in a suitable class.

The interesting feature of this result is that, in the Markov framework $\xi = g(B_T)$, the dynamic programming equation corresponding to the dual problem

(34) is fully nonlinear:

$$-\partial_t v - G(D^2 v) = 0, \text{ where } G(\gamma) := \sup_{\underline{a} \leq a \leq \bar{a}} \frac{1}{2} a D^2 v = \frac{1}{2} (\bar{a}(D^2 v)^+ - \underline{a}(D^2 v)^-). \tag{35}$$

In other words, this observation suggests that the fully nonlinear PDE corresponds to a BSDE defined quasi-surely, similar to the super-hedging problem (33). This is the starting point of our alternative formulation of second order BSDE in the subsequent Section 4, which will turn out to allow for a complete existence and uniqueness theory.

Finally, we observe that the above quasi-sure stochastic analysis is closely related to the G -stochastic integral which was recently introduced by Peng [19, 11].

4. A Quasi-sure Formulation of Second Order BSDEs

This section introduces the new framework motivated from [10] and [19].

4.1. A nondominated family of singular measures. As in Subsection 3.4, we work on the canonical space Ω . For the purpose of our second order BSDEs, we need to extend the set of non-dominated mutually singular measure \mathcal{P} to the collection of all \mathbb{P} which turn the canonical process B into a local martingale. It follows from Karandikar [15] that there exists an \mathbb{F} -progressively measurable process, denoted as $\int_0^t B_s dB_s^T$, which coincides with the Itô's integral, \mathbb{P} -a.s. for all local martingale measure \mathbb{P} . In particular, this provides a pathwise definition of

$$\langle B \rangle_t := B_t B_t^T - 2 \int_0^t B_s dB_s^T \quad \text{and} \quad \hat{a}_t := \overline{\lim}_{\varepsilon \downarrow 0} \frac{1}{\varepsilon} (\langle B \rangle_t - \langle B \rangle_{t-\varepsilon}), \tag{36}$$

where the $\overline{\lim}$ is componentwise. Clearly, $\langle B \rangle$ coincides with the \mathbb{P} -quadrature variation of B , \mathbb{P} -a.s. for all local martingale measure \mathbb{P} .

For all \mathbb{F} -progressively measurable process α taking values in the set $\mathbb{S}_d^{>0}$ of positive definite symmetric matrices and satisfying $\int_0^T |\alpha_t| dt < \infty$, \mathbb{P}_0 -a.s. we introduce the measure

$$\mathbb{P}^\alpha := \mathbb{P}_0 \circ (X^\alpha)^{-1} \quad \text{where} \quad X_t^\alpha := \int_0^t \alpha_s^{1/2} dB_s, t \in [0, T], \mathbb{P}_0 - \text{a.s.} \tag{37}$$

We denote by $\overline{\mathcal{P}}_S$ the collection of all such measures. It can be shown that

$$\begin{aligned} \text{every } \mathbb{P} \in \overline{\mathcal{P}}_S \text{ satisfies the Blumenthal zero-one law} \\ \text{and the martingale representation property.} \end{aligned} \tag{38}$$

4.2. The nonlinear generator. Consider the map $H_t(\omega, y, z, \gamma) : [0, T] \times \Omega \times \mathbb{R} \times \mathbb{R}^d \times D_H \rightarrow \mathbb{R}$, where $D_H \subset \mathbb{R}^{d \times d}$ is a given subset containing 0. We start with the following natural condition.

Assumption 4.1. For fixed (y, z, γ) , H is \mathbb{F} -progressively measurable; H is uniformly Lipschitz continuous in (y, z) , uniformly continuous in ω under the $\|\cdot\|_\infty$ -norm, and lower semi-continuous in γ .

An important role is played by the conjugate of H with respect to γ :

$$F_t(y, z, a) := \sup_{\gamma \in D_H} \left\{ \frac{1}{2} \text{Tr}[a\gamma] - H_t(y, z, \gamma) \right\}, \quad a \in \mathbb{S}_d^{>0}. \tag{39}$$

and we denote

$$\hat{F}_t(y, z) := F_t(y, z, \hat{a}_t), \quad \hat{F}_t^0 := \hat{F}_t(0, 0). \tag{40}$$

Then F is a $\mathbb{R} \cup \{\infty\}$ -valued measurable map. By the above conditions on H , the domain D_{F_t} of F as a function of a is independent of (ω, y, z) , and

$$F(\cdot, a) \text{ is uniformly Lipschitz continuous in } (y, z) \text{ and uniformly continuous in } \omega, \text{ uniformly on } (t, a), \text{ for every } a \in D_{F_t}. \tag{41}$$

For every constant $\kappa \in (1, 2]$, we denote by \mathcal{P}_H^κ the collection of all those $\mathbb{P} \in \overline{\mathcal{P}}_S$ such that

$$\begin{aligned} \underline{a}_\mathbb{P} \leq \hat{a} \leq \overline{a}_\mathbb{P}, \quad dt \times d\mathbb{P} - \text{a.s. for some } \underline{a}_\mathbb{P}, \overline{a}_\mathbb{P} \in \mathbb{S}_d^{>0}, \\ \text{and } \mathbb{E}^\mathbb{P} \left[\left(\int_0^T |\hat{F}_t^0|^\kappa dt \right)^{2/\kappa} \right] < \infty. \end{aligned} \tag{42}$$

In particular, $\hat{a}_t \in D_{F_t}$, $dt \times d\mathbb{P}$ -a.s. for all $\mathbb{P} \in \mathcal{P}_H^\kappa$.

By slightly abusing the terminology of Denis and Martini [10], we say a property holds \mathcal{P}_H^κ -quasi-surely (\mathcal{P}_H^κ -q.s. for short) if it holds \mathbb{P} -a.s. for all $\mathbb{P} \in \mathcal{P}_H^\kappa$.

Our main results requires the following conditions on \hat{F} .

Assumption 4.2. (i) \mathcal{P}_H^κ is not empty.

(ii) The process \hat{F}^0 satisfies:

$$\|\hat{F}^0\|_{\mathbb{H}_H^{2,\kappa}}^2 := \sup_{\mathbb{P} \in \mathcal{P}_H^\kappa} \mathbb{E}^\mathbb{P} \left[\text{ess sup}_{0 \leq t \leq 1}^\mathbb{P} \left(\mathbb{E}_t^{H,\mathbb{P}} \left[\int_0^1 |\hat{F}_s^0|^\kappa ds \right] \right)^{2/\kappa} \right] < \infty. \tag{43}$$

(iii) There exists a constant C such that for all $(y, z_1, z_2) \in \mathbb{R} \times \mathbb{R}^d \times \mathbb{R}^d$ and $\mathbb{P} \in \mathcal{P}_H^\kappa$:

$$|\hat{F}_t(y, z_1) - \hat{F}_t(y, z_2)| \leq C |\hat{a}_t^{1/2}(z_1 - z_2)|, \quad dt \times d\mathbb{P} - \text{a.s.} \tag{44}$$

Here we abuse the notation $\mathbb{H}_H^{p,\kappa}$ slightly by noting that, unlike the elements in \mathbb{H}_H^p , \hat{F}^0 is 1-dimensional and the norm in (43) does not contain the factor $\hat{a}^{1/2}$.

4.3. The spaces and norms. This subsection collects all norms needed for our results.

- $L_H^{p,\kappa}$: space of all \mathcal{F}_T -measurable \mathbb{R} -valued random variables ξ with

$$\|\xi\|_{L_H^{2,\kappa}}^p := \sup_{\mathbb{P} \in \mathcal{P}_H^\kappa} \mathbb{E}^\mathbb{P}[|\xi|^p] < \infty. \tag{45}$$

- $\mathbb{H}_H^{2,\kappa}$: space of all \mathbb{F}^+ -progressively measurable \mathbb{R}^d -valued processes Z with

$$\|Z\|_{\mathbb{H}_H^{2,\kappa}}^2 := \sup_{\mathbb{P} \in \mathcal{P}_H^\kappa} \mathbb{E}^\mathbb{P} \left[\int_0^T |\hat{a}_t^{1/2} Z_t|^2 dt \right] < \infty. \tag{46}$$

- $\mathbb{D}_H^{2,\kappa}$ the space of all \mathbb{F}^+ -progressively measurable \mathbb{R} -valued processes Y with \mathcal{P}_H^κ -q.s. càdlàg paths and

$$\|Y\|_{\mathbb{D}_H^{2,\kappa}}^2 := \sup_{\mathbb{P} \in \mathcal{P}_H^\kappa} \mathbb{E}^\mathbb{P} \left[\sup_{0 \leq t \leq T} |Y_t|^2 \right]. \tag{47}$$

- For $\xi \in L_H^{1,\kappa}$, $\mathbb{P} \in \mathcal{P}_H^\kappa$, and $t \in [0, T]$:

$$\mathbb{E}_t^{H,\mathbb{P}}[\xi] := \operatorname{ess\,sup}_{\mathbb{P}' \in \mathcal{P}_H^\kappa(t,\mathbb{P})} \mathbb{E}^{\mathbb{P}'}[\xi | \mathcal{F}_t] \quad \text{where} \quad \mathcal{P}_H^\kappa(t,\mathbb{P}) := \{\mathbb{P}' \in \mathcal{P}_H^\kappa : \mathbb{P}' = \mathbb{P} \text{ on } \mathcal{F}_t\}. \tag{48}$$

- $\mathbb{L}_H^{2,\kappa}$: subspace of all $\xi \in L_H^2$ such that

$$\|\xi\|_{\mathbb{L}_H^{2,\kappa}}^2 := \sup_{\mathbb{P} \in \mathcal{P}_H^\kappa} \mathbb{E}^\mathbb{P} \left[\operatorname{ess\,sup}_{0 \leq t \leq 1} \left(\mathbb{E}_t^{H,\mathbb{P}}[|\xi|^\kappa] \right)^{2/\kappa} \right] < \infty. \tag{49}$$

- $\operatorname{UC}_b(\Omega)$: space of all bounded and uniformly continuous maps $\xi : \Omega \rightarrow \mathbb{R}$ with respect to the $\|\cdot\|_\infty$ -norm.

- $\mathcal{L}_H^{2,\kappa}$: closure of $\operatorname{UC}_b(\Omega)$ under the norm $\|\cdot\|_{\mathbb{L}_H^{2,\kappa}}$.

We observe that when \mathcal{P}_H^κ is reduced to a singleton:

$$\mathcal{P}_H^\kappa = \{\mathbb{P}\} \implies \mathcal{L}_H^{2,\kappa} = \mathbb{L}_H^{2,\kappa} = L_H^{2,\kappa} = L^2(\mathbb{P}) \quad \text{for } 1 \leq \kappa < p. \tag{50}$$

4.4. Definition. We shall obtain a complete existence and uniqueness theory for the second order BSDE (25) by considering instead the quasi-sure formulation:

$$Y_t = \xi - \int_t^T \hat{F}_s(Y_s, Z_s) ds - \int_t^T Z_s \cdot dB_s + K_1 - K_t, \quad 0 \leq t \leq T, \quad \mathcal{P}_H^\kappa\text{-q.s.} \tag{51}$$

A solution to the 2BSDE (51) is a pair $(Y, Z) \in \mathbb{D}_H^{2,\kappa} \times \mathbb{H}_H^{2,\kappa}$ such that:

- $Y_T = \xi, \mathcal{P}_H^\kappa$ -q.s.
- For all $\mathbb{P} \in \mathcal{P}_H^\kappa$, the process

$$K_t^\mathbb{P} := Y_0 - Y_t + \int_0^t \hat{F}_s(Y_s, Z_s) ds + \int_0^t Z_s dB_s, \quad 0 \leq t \leq T, \quad \mathbb{P} - \text{a.s.} \quad (52)$$

has non-decreasing paths, \mathbb{P} -a.s.

- The family $\{K^\mathbb{P}, \mathbb{P} \in \mathcal{P}_H^\kappa\}$ satisfies the minimality condition:

$$K_t^\mathbb{P} = \operatorname{ess\,inf}_{\mathbb{P}' \in \mathcal{P}_H^\kappa(t, \mathbb{P})} \mathbb{E}_t^{\mathbb{P}'} [K_T^{\mathbb{P}'}], \quad \mathbb{P} - \text{a.s. for all } \mathbb{P} \in \mathcal{P}_H^\kappa \text{ and } t \in [0, 1]. \quad (53)$$

The above definition is motivated in [22, 23] by the corresponding stochastic target problem. Let us just verify that it reduces to the standard notion of BSDE when the generator H is linear in γ :

$$H_t(y, z, \gamma) = \frac{1}{2} \operatorname{Tr}[a_t^0 \gamma] - f_t(y, z), \quad (54)$$

where $a^0 : [0, T] \times \Omega \rightarrow \mathbb{S}_d^{>0}$ is \mathbb{F} -progressively measurable and has uniform lower and upper bounds. We remark that in this case we do not need to assume that a^0 and f are uniformly continuous in ω . Then, under obvious extension of notations, we have

$$D_{F_t(\omega)} = \{a_t^0(\omega)\} \quad \text{and} \quad \hat{F}_t(y, z) = f_t(y, z). \quad (55)$$

Assume further that there exists $\mathbb{P} \in \overline{\mathcal{P}}_S$ such that

$$\hat{a} = a^0, \quad \mathbb{P} - \text{a.s. and} \quad \mathbb{E}^\mathbb{P} \left[\int_0^T (|f_t(0, 0)|^2 dt) \right] < \infty, \quad (56)$$

then $\mathcal{P}_H^\kappa = \mathcal{P}_H^2 = \{\mathbb{P}\}$. In this case, the minimum condition (53) implies

$$0 = K_0 = \mathbb{E}^\mathbb{P}[K_T] \quad \text{and thus} \quad K = 0, \quad \mathbb{P} - \text{a.s.} \quad (57)$$

Hence, the 2BSDE (51) is equivalent to the following standard BSDE:

$$Y_t = \xi - \int_t^T f_s(Y_s, Z_s) ds - \int_t^1 Z_s dB_s, \quad 0 \leq t \leq T, \quad \mathbb{P} - \text{a.s.} \quad (58)$$

Finally, we recall from the previous subsection that in the present case, we have $\mathcal{L}_H^{2, \kappa} = \mathbb{L}_H^{2, \kappa} = L_H^2 = \mathbb{L}^2(\mathbb{P})$ for all $\kappa \in [1, 2)$.

5. Wellposedness of Second Order BSDEs

This section contains the main results of the papers [20, 21, 22, 23].

For any $\mathbb{P} \in \mathcal{P}_H^\kappa$, \mathbb{F} -stopping time τ , and \mathcal{F}_τ -measurable random variable $\xi \in \mathbb{L}^2(\mathbb{P})$, we denote by $(\mathcal{Y}^\mathbb{P}, \mathcal{Z}^\mathbb{P}) := (\mathcal{Y}^\mathbb{P}(\tau, \xi), \mathcal{Z}^\mathbb{P}(\tau, \xi))$ the solution to the following standard BSDE:

$$\mathcal{Y}_t^\mathbb{P} = \xi - \int_t^\tau \hat{F}_s(\mathcal{Y}_s^\mathbb{P}, \mathcal{Z}_s^\mathbb{P}) ds - \int_t^\tau \mathcal{Z}_s^\mathbb{P} dB_s, \quad 0 \leq t \leq \tau, \quad \mathbb{P} - \text{a.s.} \tag{59}$$

Our first result provides a representation of any solution of the 2BSDE (51).

Theorem 5.1. *Let Assumptions 4.1 and 4.2 hold. Assume that $\xi \in \mathbb{L}_H^{2,\kappa}$ and that $(Y, Z) \in \mathbb{D}_H^{2,\kappa} \times \mathbb{H}_H^{2,\kappa}$ is a solution to 2BSDE (51). Then, for any $\mathbb{P} \in \mathcal{P}_H^\kappa$ and $0 \leq t \leq T$,*

$$Y_t = \operatorname{ess\,sup}_{\mathbb{P}' \in \mathcal{P}_H^\kappa(t, \mathbb{P})} \mathbb{P} \mathcal{Y}_t^{\mathbb{P}'}(T, \xi), \quad \mathbb{P} - \text{a.s.} \tag{60}$$

Consequently, the 2BSDE (51) has at most one solution in $\mathbb{D}_H^{2,\kappa} \times \mathbb{H}_H^{2,\kappa}$.

The above representation, together with the comparison principle for standard BSDEs, implies the following comparison principle for 2BSDEs.

Corollary. *Let Assumptions 4.1 and 4.2 hold. Assume $\xi^i \in \mathbb{L}_H^{2,\kappa}$ and $(Y^i, Z^i) \in \mathbb{D}_H^{2,\kappa} \times \mathbb{H}_H^{2,\kappa}$ is a corresponding solution of the 2BSDE (51), $i = 1, 2$. If $\xi^1 \leq \xi^2$, \mathcal{P}_H^κ -q.s. then $Y^1 \leq Y^2$, \mathcal{P}_H^κ -q.s.*

We next state the a priori estimates which will be used in the subsequent existence result.

Theorem 5.2. *Let Assumptions 4.1 and 4.2 hold.*

(i) *Assume that $\xi \in \mathbb{L}_H^{2,\kappa}$ and that $(Y, Z) \in \mathbb{D}_H^{2,\kappa} \times \mathbb{H}_H^{2,\kappa}$ is a solution to 2BSDE (51). Then there exist a constant C_κ such that*

$$\|Y\|_{\mathbb{D}_H^{2,\kappa}}^2 + \|Z\|_{\mathbb{H}_H^{2,\kappa}}^2 + \sup_{\mathbb{P} \in \mathcal{P}_H^\kappa} \mathbb{E}^\mathbb{P} [|K_1^\mathbb{P}|^2] \leq C_\kappa (\|\xi\|_{\mathbb{L}_H^{2,\kappa}}^2 + \|\hat{F}^0\|_{\mathbb{H}_{\mathcal{P}_H^\kappa}^{2,\kappa}}^2). \tag{61}$$

(ii) *Assume that $\xi^i \in \mathbb{L}_H^{2,\kappa}$ and that $(Y^i, Z^i) \in \mathbb{D}_H^{2,\kappa} \times \mathbb{H}_H^{2,\kappa}$ is a corresponding solution to 2BSDE (51), $i = 1, 2$. Denote $\delta\xi := \xi^1 - \xi^2$, $\delta Y := Y^1 - Y^2$, $\delta Z := Z^1 - Z^2$, and $\delta K^\mathbb{P} := K^{1,\mathbb{P}} - K^{2,\mathbb{P}}$. Then there exists a constant C_κ such that*

$$\begin{aligned} \|\delta Y\|_{\mathbb{D}_H^{2,\kappa}} &\leq C_\kappa \|\delta\xi\|_{\mathbb{L}_H^{2,\kappa}}, \|\delta Z\|_{\mathbb{H}_H^{2,\kappa}}^2 + \sup_{\mathbb{P} \in \mathcal{P}_H^\kappa} \mathbb{E}^\mathbb{P} \left[\sup_{0 \leq t \leq 1} |\delta K_t^\mathbb{P}|^2 \right] \\ &\leq C_\kappa \|\delta\xi\|_{\mathbb{L}_H^{2,\kappa}} \left(\|\xi^1\|_{\mathbb{L}_H^{2,\kappa}} + \|\xi^2\|_{\mathbb{L}_H^{2,\kappa}} + \|\hat{F}^0\|_{\mathbb{H}_{\mathcal{P}_H^\kappa}^{2,\kappa}} \right). \end{aligned} \tag{62}$$

The main result of this paper is:

Theorem 5.3. *Let Assumptions 4.1 and 4.2 hold. Then for any $\xi \in \mathcal{L}_H^{2,\kappa}$, the 2BSDE (51) has a unique solution $(Y, Z) \in \mathbb{D}_H^{2,\kappa} \times \mathbb{H}_H^{2,\kappa}$.*

Our final result concern the connection between the 2BSDE (51) and the corresponding fully nonlinear PDE in the Markov case:

$$H_t(\omega, y, z, \gamma) = h(t, B_t(\omega), y, z, \gamma) \quad \text{and} \quad \xi = g(\omega). \tag{63}$$

Observe that h may not be nondecreasing in γ , but the following \hat{h} is:

$$\hat{h}(t, x, y, z, \gamma) = \sup_{a \in \mathbb{S}_d^{>0}} \left\{ \frac{1}{2} \text{Tr}[a\gamma] - f(t, x, y, z, a) \right\}, \quad \gamma \in \mathbb{R}^{d \times d}. \tag{64}$$

To obtain the connection with the corresponding fully nonlinear PDE, we need more assumptions which are detailed in [23]. Let us just mention that under those assumptions, we have

$$Y_t = u(t, B_t), \quad t \in [0, T], \tag{65}$$

where

(i) u is a viscosity subsolution of

$$-\partial_t u^* - \hat{h}^*(\cdot, u^*, Du^*, D^2 u^*) \leq 0 \quad \text{on} \quad [0, 1) \times \mathbb{R}^d. \tag{66}$$

(ii) u is a viscosity supersolution of

$$-\partial_t u_* - \hat{h}_*(\cdot, u_*, Du_*, D^2 u_*) \geq 0 \quad \text{on} \quad [0, 1) \times \mathbb{R}^d. \tag{67}$$

Here, we used the classical notation in the theory of viscosity solutions:

$$u_*(\theta) := \liminf_{\theta' \rightarrow \theta} u(\theta') \quad \text{and} \quad u^*(\theta) := \overline{\lim}_{\theta' \rightarrow \theta} u(\theta'), \quad \text{for } \theta = (t, x), \tag{68}$$

$$\hat{h}_*(\theta) := \liminf_{\theta' \rightarrow \theta} \hat{h}(\theta') \quad \text{and} \quad \hat{h}^*(\theta) := \overline{\lim}_{\theta' \rightarrow \theta} \hat{h}(\theta'), \quad \text{for } \theta = (t, x, y, z, \gamma). \tag{69}$$

Example Hedging under Gamma constraints. Consider the quasi-sure reformulation of the problem of Subsection 3.1. The generator is given by

$$h(t, x, y, z, \gamma) := \frac{1}{2} \gamma \text{ if } \gamma \in [\underline{\Gamma}, \bar{\Gamma}], \quad \text{and} \quad \infty \text{ otherwise,} \tag{70}$$

where $\underline{\Gamma} < 0 < \bar{\Gamma}$ are given constants. By direct calculation, we see that

$$f(a) = \frac{1}{2} (\bar{\Gamma}(a - 1)^+ - \underline{\Gamma}(a - 1)^-), \quad a \geq 0, \tag{71}$$

and

$$\hat{h}(\gamma) = \frac{1}{2} (\gamma \vee \underline{\Gamma}) \text{ if } \gamma \leq \bar{\Gamma}, \quad \text{and} \quad \infty \text{ otherwise.} \tag{72}$$

Then,

$$\hat{h}_* = \hat{h} \quad \text{and} \quad \hat{h}^*(\gamma) = \frac{1}{2}(\gamma \vee \underline{\Gamma})\mathbf{1}_{\{\gamma < \bar{\Gamma}\}} + \infty\mathbf{1}_{\{\gamma \geq \bar{\Gamma}\}}. \tag{73}$$

In view of this, the above viscosity properties (66)-(67) are equivalent to

$$\min \left\{ -\partial_t u^* - \frac{1}{2}(D^2 u^* \vee \underline{\Gamma}), \bar{\Gamma} - D^2 u^* \right\} \leq 0, \tag{74}$$

$$\min \left\{ -\partial_t u_* - \frac{1}{2}(D^2 u_* \vee \underline{\Gamma}), \bar{\Gamma} - D^2 u_* \right\} \geq 0. \tag{75}$$

6. A Probabilistic Scheme for Fully Nonlinear PDEs

Consider the fully nonlinear Cauchy problem:

$$-\mathcal{L}^X v - h_0(\cdot, v, Dv, D^2 v) = 0, \quad \text{on } [0, T] \times \mathbb{R}^d, \tag{76}$$

$$v(T, \cdot) = g, \quad \text{on } \mathbb{R}^d. \tag{77}$$

where

$$\mathcal{L}^X \varphi := \partial_t \varphi + \mu \cdot D\varphi + \frac{1}{2} \text{Tr}[aD^2 \varphi] \tag{78}$$

is the Dynkin operator of some Markov diffusion process. Similar to Subsection 2.4, a probabilistic numerical scheme for fully nonlinear PDEs was suggested in [8] and analyzed later in [13].

To simplify the notation, we consider the case $\sigma = I_d$. let $\pi : t = t_0 < \dots < t_n = T$ be a partition of the interval $[t, T]$ with time steps $\delta t_k := t_k - t_{k-1}$, and corresponding increments of the Brownian motion $\delta W_{t_k} := W_{t_k} - W_{t_{k-1}}$. Denote by X^π the Euler discretization of X along the partition π . Then the probabilistic numerical scheme for the fully nonlinear PDE is defined by:

$$Y_{t_n}^\pi = g(X_{t_n}^\pi), \tag{79}$$

and

$$Y_{t_{k-1}}^\pi = \mathbb{E}[Y_{t_k}^\pi | X_{t_{k-1}}^\pi] + \delta t_k h_0(t_{k-1}, X_{t_{k-1}}^\pi, Y_{t_{k-1}}^\pi, Z_{t_{k-1}}^\pi, \Gamma_{t_{k-1}}^\pi), \tag{80}$$

$$Z_{t_{k-1}}^\pi = \mathbb{E}\left[Y_{t_k}^\pi \frac{\delta W_{t_k}}{\delta t_k} | X_{t_{k-1}}^\pi \right], \tag{81}$$

$$\Gamma_{t_{k-1}}^\pi = \mathbb{E}\left[Y_{t_k}^\pi \frac{\delta W_{t_k} \delta W_{t_k}^\top - \delta t_k}{(\delta t_k)^2} | X_{t_{k-1}}^\pi \right]. \tag{82}$$

The convergence of this probabilistic numerical scheme is analyzed in [13] by the method of monotonic schemes introduced by Barles and Souganidis [4] and further developed by Krylov [17], Barles and Jakobsen [3].

Moreover, a numerical implementation is reported in [13] for the 3-dimensional mean curvature flow, and a five dimensional stochastic control problem.

References

- [1] V. Bally and G. Pagès, Error analysis of the quantization algorithm for obstacle problems, *Stochastic Processes and Their Applications*, 106(1), 1–40, (2003).
- [2] P. Bank and D. Baum, Hedging and portfolio optimization in financial markets with a large trader, *Mathematical Finance*, 14, 1–18 (2004).
- [3] G. Barles, E. R. Jakobsen, On the convergence rate of approximation schemes for Hamilton-Jacobi-Bellman equations, *Mathematical Modelling and Numerical Analysis*, ESAIM, M2AM, Vol. 36 (2002), No. 1, 33–54.
- [4] G. Barles, P. E. Souganidis, Convergence of Approximation Schemes for Fully Non-linear Second Order Equation, *Asymptotic Analysis: Theory, Methods, and Applications*, 4, pp. 271–283, 1991.
- [5] J.M. Bismut, Conjugate convex functions in optimal stochastic control, *J. Math. Anal. Appl.* 44, 384–404 (1973).
- [6] B. Bouchard and N. Touzi, Discrete-time approximation and Monte Carlo simulation of backward stochastic differential equations, *Stochastic Processes and their Applications*, 111, 175–206 (2004).
- [7] P. Cheridito, H.M. Soner, and N. Touzi, The multi-dimensional super-replication problem under Gamma constraints, *Annales de l'Institut Henri Poincaré, Série C: Analyse Non-Linéaire* 22, 633–666 (2005).
- [8] P. Cheridito, H.M. Soner, N. Touzi, and N. Victoir, Second order BSDE's and fully nonlinear PDE's, *Communications in Pure and Applied Mathematics*, 60 (7): 1081–1110 (2007).
- [9] D. Crisan, K. Manolarakis et N. Touzi, On the Monte Carlo simulation of Backward SDES: an improvement on the Malliavin weights, *Stochastic Processes and Their Applications*, to appear.
- [10] L. Denis and C. Martini, A Theoretical Framework for the Pricing of Contingent Claims in the Presence of Model Uncertainty, *Annals of Applied Probability* 16, 2, 827–852 (2006).
- [11] L. Denis, M. Hu, and S. Peng, Function Spaces and Capacity Related to a Sub-linear Expectation: Application to G-Brownian Motion Paths, arXiv:0802.1240 (February 2008).
- [12] N. El Karoui, S. Peng and M.-C. Quenez, Backward stochastic differential equations in finance. *Mathematical Finance* 7, 1–71.
- [13] Fahim A, N. Touzi. and X. Warin, A Probabilistic Numerical Scheme for Fully Nonlinear PDEs. Preprint.
- [14] E. Gobet, J.P. Lemor, and X. Warin, Rate of convergence of an empirical regression method for solving generalized backward stochastic differential equations, *Bernoulli*, Volume 12 (5), 889–916 (2006).

-
- [15] R. Karandikar, On pathwise stochastic integration, *Stochastic Processes and Their Applications*, 57 (1995), 11–18.
 - [16] M. Kobylanski, Backward stochastic differential equations and partial differential equations with quadratic growth, *Annals of Probability* 28, 558–602.
 - [17] N. V. Krylov, On The Rate Of Convergence Of Finite-Difference Approximations For Bellman's Equations, *St. Petersburg Math. J.*, Vol. 9 (1997), No. 3, 245–256.
 - [18] E. Pardoux and S. Peng, Adapted solution of a backward stochastic differential equation, *Systems Control Lett.*, 14, 55–61 (1990).
 - [19] S. Peng, G-Brownian motion and dynamic risk measure under volatility uncertainty, arXiv:0711.2834v1 (2007).
 - [20] M. Soner, J. Zhang, and N. Touzi, Quasi-sure stochastic analysis through aggregation. Preprint.
 - [21] M. Soner, J. Zhang, and N. Touzi, Martingale representation theorem for the G-Expectation. Preprint.
 - [22] M. Soner, J. Zhang, and N. Touzi, Dual formulation of second order target problems. Preprint.
 - [23] M. Soner, J. Zhang, and N. Touzi, Wellposedness of second order backward SDEs. Preprint.
 - [24] R. Tevzadze, Solvability of backward stochastic differential equations with quadratic growth, *Stochastic Processes and Their Applications* 118(3), 503–515 (2008).
 - [25] J. Zhang, A numerical scheme for backward stochastic differential equations, *Annals of Applied Probability*, 14 (1), 459–488 (2004).

Data Modeling: Visual Psychology Approach and $L_{1/2}$ Regularization Theory

Zongben Xu*

Abstract

Data modeling provides data analysis with models and methodologies. Its fundamental tasks are to find structures, rules and tendencies from a data set. The data modeling problems can be treated as cognition problems. Therefore, simulating cognition mechanism and principles can provide new subtle paradigm and can solve some basic problems in data modeling.

In pattern recognition, human eyes possess a singular aptitude to group objects and find important structure in an efficient way. I propose to solve a clustering and classification problem through capturing the structure (from micro to macro) of a data set from a dynamic process observed in adequate scale spaces. Three types of scale spaces are introduced, respectively based on the neural coding, the blurring effect of lateral retinal interconnections, the hierarchical feature extraction mechanism dominated by receptive field functions and the feature integration principle characterized by Gestalt law in psychology.

The use of L_1 regularization has now been widespread for latent variable analysis (particularly for sparsity problems). I suggest an alternative of such commonly used methodology by developing a new, more powerful approach – $L_{1/2}$ regularization theory. Some related open questions are raised in the end of the talk.

Mathematics Subject Classification (2010). 6IH30, 68T10, 62-07, 94A12.

Keywords. Data modeling, sparse signal recovery, visual psychology approach, L_1 regularization, $L_{1/2}$ regularization.

*The research was supported by National 973 Program (2007CB311002) and NSF Projects (60975036,60905003) of China.

Department of Mathematics & Institute for Information and System Sciences, Xi'an Jiaotong University, XI'an, 710049, P.R. China. E-mail: zbxu@mail.xjtu.edu.cn.

1. Introduction

We are in the era of knowledge economy. One of the main features is the rapid growing of the information technology which has become the most lucrative segment of the world economy, with much of the growth occurring in the development, management, and application of prodigious streams of data for scientific, medical, engineering, and commercial purposes. Responding to the rapid advances in information technology, data analysis has been developed at break-neck pace in the last few decades. It has now been a very significant, or even main part of science and engineering, as predicted by John Tukey [1] forty years ago.

The main purpose of data analysis is to help people to understand the meaning and value of the data. Initiated from statistics, data analysis has, however, strong connections with many other disciplines such as computer science, information processing and pattern recognition. It is inarguably accepted as a part of information technology today.

Data Modeling provides data analysis with models and methodologies. In other words, data modeling yields the data analysis techniques that have solid mathematical basis. Different from traditional mathematical modeling that aims to formulate a phenomenon, a principle or a system, data modeling models a data set. This is perhaps a basic form of applications of mathematics nowadays.

The fundamental tasks of data modeling are to find patterns, structures, rules, relations or tendencies from a data set, which serves then to explain which measurement(s) or attribute(s) is relevant to the phenomenon of interest, or what kind of structures or rules existed in a collection of data. The aims are provision of computational models which makes it possible that data can be automatically perceived and understood for decision. The basic problems of data modeling include clustering, classification, regression and latent variable analysis [2].

Clustering is a problem of partitioning a data set into subgroups based on similarity among data. It seeks to arrange an unordered collection of objects in a fashion so that nearby objects are similar. Very basic to knowledge discovery, the clustering is capable of finding new concepts, new phenomenon or new patterns of data. *Classification* is a problem of seeking a general discriminative rule (normally, a function) to categorize the data by their attributes. The sought discriminative function is then used in discriminative analysis, and therefore, laid the basis of any pattern recognition application. *Regression* aims to determine a quantitative cause/result relationship between variables in data, where M variables in the data are quantitative response variables, and the other N variables are used to predict it. This quantitative relationship is generally modeled as a continuous function (say, a polynomial or a neural net), and mainly used for prediction/forecasting application. *Latent variable analysis* attempts to identify the intrinsic variables from the observation, fundamental to visu-

alization, feature extraction and motion modeling. In such a problem, we are given

$$\mathbf{y} = A\mathbf{x}, \quad \mathbf{x} \in R^N, \quad \mathbf{y} \in R^M$$

\mathbf{y} is a observation, \mathbf{x} is a unobserved latent variables, and A is a linear transformation converting one into the other. The hope is that a few underlying latent variables are responsible for essentially the structure we see in the observation, and by uncovering those variables, we can achieve important insights. We easily see that the latent variable analysis problem can be reexpressed as a *sparsity problem* [3], as will be explained latter in section 3 of this talk.

All the above problems can be tackled within the frameworks of statistics and information science. A great number of useful and effective tools and techniques, for instance, have been developed from those methodologies. The k-means, Graph-based Clustering, Fisher Discriminant Analysis, Support Vector Machine, Neural Networks, Fuzzy Systems, Boosting, PCA, Manifold Learning are just a few of the popularly used techniques. Nevertheless, all those techniques face challenges when applied to real data sets we are meeting today and in future.

The challenges come mainly from several striking features of real data sets: (i) *massiveness*, say, think of the huge volumes of data automatically generated by a satellite; (ii) *high dimensionality*, say, think of the DNA microarrays for patients, where genes are huge, but relatively few patients with a given genetic disease; (iii) *inhomogeneity*, say, think of a multi-medium data set which contains images, texts, media, and video in the same time; and (iv) *uncertainty*, say, think of hyperspectral imagery, internet portals, and financial tick-by-tick data, in which noise and inaccuracy are inevitably involved in gathering or measurement. All these features may make the existing techniques either infeasible or ineffective.

To be further, for example, the massiveness of a data set may cause ineffectiveness for any algorithms related to inversion of a matrix, which takes $\mathcal{O}(N^3)$ operations and for large N (say in the millions) is prohibitively expensive. The high dimensionality may lead to infeasibility and ineffectiveness of most techniques based on traditional statistical methodology. This is because, in traditional statistical methodology, we assumed many observations and a few, well-chosen variables (namely, $M \gg N$, the *large sample problem*). The data set today is, however, towards more observations but even more so, to radically larger number of variables. We are seeing examples where the observations gathered on individual instances are curves, or spectrums, or images, or even movies, so that a single observation has dimensions in the thousands or billions, while there are only tens or hundreds of instances available for study (thus, $M \ll N$, the *small sample problem*). Such high-dimension/small sample problems cannot be solved effectively by the large sample algorithms.

The challenges will get more serious if we take it into account that our purpose of data modeling, hopefully, is to provide computational models for automatical understanding of data (such type of models are referred to as

Machine Cognition Models). In other words, a machine cognition model provides a technique that can perform an automatical data analysis without any other assistance. From this sense, most of existing techniques are still far from the end.

It is unlikely to have all the problems being solved simultaneously. For some special and separate cases, however, some significant progresses can be made. In this talk, I review some of these progresses.

As the terminology “Pattern Recognition” implies, pattern recognition (essentially, a classification problem) could be accomplished by repeating the human cognition rules (that is, Re-cognition is the way to solve the problem). Through viewing a data modeling problem as a cognition problem, clustering, classification and regression problems can be tackled by mimicking visual psychology. Such visual psychology approach brings many benefit, defines machine cognition models of the problems, and provides satisfactory solutions to several long-standing problems in data analysis. We summarize the related works in the next section.

The way how our visual system encodes observation naturally motivates the methodology for solving latent variable analysis problem. Such an approach could be considered in a more general framework, sparsity problems — to find sparse solution(s) of a representation or an underdetermined equation. A common practice for solution of sparsity problems is L_1 regularization, formalized by Tibshirani [4] and Chen, Donoho, and Saunders [5]. The use of L_1 regularization has become so widespread that it could arguably be considered the “modern least squares” [6]. However, for many applications, the solutions of the L_1 regularization are often less sparse than that expected. As an alternative, $L_{1/2}$ regularization then has been developed in recent years by my group. I introduce such new methodology in section 3.

In section 4 I propose problems open to be answered along the line of research topics talked here.

2. Visual Psychology Approach

We begin with an observation that for most of the data modeling problems in low dimensions (say, $N = 1, 2$), the solutions of problems can always be promptly captured with our eyes. Why it is so is due to the unrivaled cognition ability of human being! The approach I will introduce in this section just follows this modus of human being to solve a data modeling problem.

Thus, my basic point of view is: *A data modeling problem is a cognition problem*. Although this is supported only with the low dimensional problems, we can solve the problem through modeling it in the way of human beings in low dimensions, and then generalizing it to the high dimensions through formalization plus mathematical justification.

Let us first explain how a data set can be transformed into an object that can be observed by our eyes. Naturally, such an object should be somewhat

an image, and we call it the *Data Image*. The data image is a real one only in low dimensions, but imaginary in high dimensional cases. Given a data set $D = \{z_i = (x_i, y_i)\}_{i=1}^M$ with $x_i \in R^N, y_i \in R^1$, the data image of data set D can be defined with its empirical distribution respective to the problems we are tackling. For example, for clustering problem, the data image can be defined as

$$g_D(z) = \frac{1}{M} \sum_{i=1}^M \delta(z - z_i) \quad (1)$$

For classification problem, it is then defined by

$$g_D(x) = \frac{1}{M_+ + M_-} \left(\sum_{i=1}^{M_+} \delta(x - x_i^+) - \sum_{i=1}^{M_-} \delta(x - x_i^-) \right) \quad (2)$$

where the classification problem is assumed to be canonical, that is, a two-class problem, and the data set is correspondingly splitted into two parts:

$$D = \{(x_i^+, +1)\}_{i=1}^{M_+} \cup \{(x_i^-, -1)\}_{i=1}^{M_-}.$$

Data images are very special images without color and continuous texture information. A data image, however, contains various macro-information like cluster structure, separation structure, tendency, dependence, all of those interested us. According to physics, any macro-structure must consist of micro-structures. The macro-structure of a data set thus can be observed only when various micro-structures of the data have been perceived. What types of micro-structures have been captured then when we observe a data image? The psychology experiments conducted by Santos and Marqures [7] suggested the following ingredients:

- **Density Feature** It is the distribution difference feature of data, which can be measured with the number of data in a certain volume of data space; A data set with uniform distribution is normally accepted as no feature because no visual difference is perceived.
- **Connectedness Feature** It is the feature of a data set in which some data look like the samplings on a curve or a manifold. When they are observed from appropriately far away, those data appear as continuous curves or manifolds.
- **Orientation Feature** A datum together with its surrounding data defines a subregion of data space. If the subregion has a distinct principle direction, the datum is said to have local orientation; If the local orientation of some data are almost same, those data are said to have a structure direction. Whenever there exists structure direction in the set, the data set is said to have orientation feature.

Those structures are reexpressed in [8] with computational models. We remark that the micro-structures of a data set is by no means accountable, and it actually depends on the *visual attention* and what type of *observation purpose* is taking. For example, when a discrimination task is taking, the separation extent (margin) and boundary may be also perceived, besides the features mentioned above.

The crucial problems are: How those structure features have been organized into a macro-structure, and how the macro-structures have in turn been captured by our human eyes? This is the key to any attempt of solving the data modeling problems in the same or similar ways as our human beings do. The complete solutions are clearly in brain science, cognition science, and perhaps whole sciences, are in future but still unknown today. However, in recent years, physiological discoveries and researches in computer-aided neuroanatomy, neurobiology, and psychology have advanced several quite accurate computational models of primary visual system, each modeling some parts of the human visual system at a particular level of details. By simulating those known facts and discoveries, it is possible to form data modeling techniques more or less like the human eyes. Taking clustering problem as an example, I introduce those progresses below.

2.1. Scale Space Based Approach. One of our common visual experiences is that how clearly we observe an object depends on the distance of our observation. This is the principle of blurring effect of lateral retinal interconnections in primary visual system. The scale space theory, which models the blurring effect by applying Gaussian filtering to a digital image, was introduced by Witkin [9] in 1983. Suppose $P(x)$ is the intensity distribution of one object in nature and $P(x, \sigma)$ is the intensity distribution of the projected image of the object on the retina, where σ is a scale, understood either as the distance between the object and eyes or as the curvature of crystalline lens. Then $P(x, \sigma)$ can be mathematically described by

$$\begin{cases} \frac{\partial P(x, \sigma)}{\partial \sigma} = \Delta_x P(x, \sigma) \\ P(x, 0) = P(x) \end{cases}, \quad (3)$$

the solution of which is explicitly given by

$$P(x, \sigma) = P(x) * g(x, \sigma) = \int g(x - y)P(y)dy \quad (4)$$

where ‘*’ denotes the convolution operation and $g(x, \sigma)$ is the Gaussian function

$$g(x, \sigma) = \frac{1}{(\sqrt{2\pi}\sigma)^n} e^{-\|x\|^2/2\sigma^2}. \quad (5)$$

In this way, $P(x)$ has been embedded into a continuous family $P(x, \sigma)$ of gradually smoother versions of $P(x)$. The original image corresponds to the scale

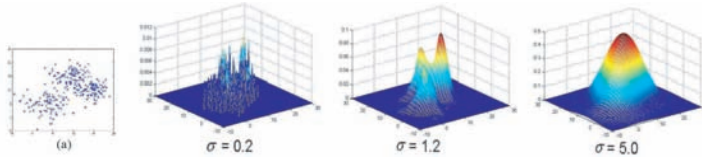


Figure 1. How evolves the data set (a) in scale space.

$\sigma = 0$, and, as the scale σ increases, $P(x, \sigma)$ gives a more and more blurring while simplified representation of $P(x)$ without creating spurious structure. Due to this, $P(x, \sigma)$ is referred to as a *multi-scale representation* of the image $P(x)$, and $\{P(x, \sigma)\}_{\sigma \geq 0}$ is a scale space. For any σ , $P(x, \sigma)$ is called a *scale space image*.

Interestingly, it can be shown that the above representation is unique if the retina property is assumed to be isotropic and spatially invariant. Without those assumptions, nevertheless, several other complicated PDE models, say, Anisotropic Diffusion Models, can be built. These models can not be directly applied to the approach introduced here.

Now, applying the scale space theory to the data image (1), we have the following multi-scale representation of data set D

$$P(x, \sigma) = \frac{1}{N} \sum_{i=1}^N \frac{1}{(\sigma\sqrt{2\pi})^2} e^{-\frac{\|x-x_i\|^2}{2\sigma^2}} \quad (6)$$

which coincides with the Parzen distribution estimations of D with Gaussian window function. Figure 1 illustrates how a data set evolves in the scale space, i.e., what a multi-scale representation of a data set looks like.

As demonstrated in Figure 1, the data set appears as a data image with each datum being a light point attached with a uniform luminous flux. As we blur this image, each datum first becomes a light blob. Throughout the blurring process, smaller blobs merge into larger ones until the whole image contains only one light blob at a low enough level of resolution. In the process, small blobs always merge into large ones and new ones are never created. If we equate each blob with a cluster, the above blurring process seems providing a natural hierarchical clustering with resolution being the height of a dendrogram.

This is the point of our approach. That is, our idea is to capture the structure (from micro to macro) of a data set from the dynamic process observed in the scale space. This is a natural way to structure-finding, as inspired by the function of a lens in the visual system and our everyday visual experience.

However, to formalize this idea into a standard procedure of data clustering, three questions must be answered. (i) What means a cluster and how it can be formalized? (ii) How the continuous scale σ can be discretized so as not to affect our observation (say, not cause the loss of important structures)? and (iii) Does the blobs (clusters) evolve in an somewhat regular way? We answer those questions one by one below.

First, each blob can be defined as a cluster. So, for each fixed scale σ , we define a cluster (a light blob) as being the region in data set (corresponding to scale $\sigma = 0$) that satisfies

$$C_{y_\sigma} = \left\{ x_0 \in R^N : \lim_{t \rightarrow \infty} x(t, x_0) = y_\sigma \right\},$$

where $x(t, x_0)$ is the solution of gradient flow

$$\begin{cases} \frac{dx}{dt} = \nabla_x P(x, \sigma) \\ x(0) = x_0 \end{cases} \quad (7)$$

Here y_σ is a maxima of scale space image $P(x, \sigma)$, and referred as the blob center or cluster center of C_{y_σ} . Thus, at each scale σ , all blobs in $P(x, \sigma)$ produce a partition of data set D with each point belonging to a unique blob (cluster) except the boundary point. Each blob has its own survival range of scale, and larger blobs are made up of smaller blobs through the evolution. In consequence, a higher scale partition of D can be deduced from its lower scale partition, as long as the evolution of clusters is regular, leading to the third question in turn.

Second, we discretize the continuous scale σ according to the way of our human being. In psychophysics, Weber's law says that the minimal size of the difference ΔI in stimulus intensity which can be sensed is related to the magnitude of standard stimulus intensity I by $\Delta I = kI$, where k is a constant called Weber fraction. Coren [10] experimentally showed that $k = 0.029$ in one-dimensional observation. Consequently, we suggest the following discretization scheme for our observation:

$$\sigma_i - \sigma_{i-1} = k\sigma_{i-1}$$

where k is any constant not larger than Weber fraction. According to psychology, such a discretization scheme provides us a guarantee with which we cannot sense the difference between any two scale space images $P(x, \sigma_i)$ and $P(x, \sigma_{i-1})$.

The third question is essentially concerned with whether the cluster number, $\pi(\sigma)$, can be monotonically decreasing in the scale space. Define the cluster center curve $\Gamma = \{y_\sigma : \sigma \geq 0\}$. The following Theorem 2.1 justifies that Γ exactly consists of N simple curves, like Figure 2. So the monotonically decreasing of $\pi(\sigma)$ follows.

Theorem 2.1 ([11]). *For almost all data sets, we have: 1) zero is a regular value of $\nabla_x P(x, \sigma)$; 2) as $\sigma \rightarrow 0$, the clustering obtained for $P(x, \sigma)$ with $\sigma > 0$ induces a clustering at $\sigma = 0$ in which each datum is a cluster and the corresponding partition is a Voronoi tessellation, i.e., each point in the scale space belongs to its nearest-neighbor datum, and 3) as σ increases from $\sigma = 0$, there are N maximal curves in the scale space with each of them starting from a datum of the data set.*

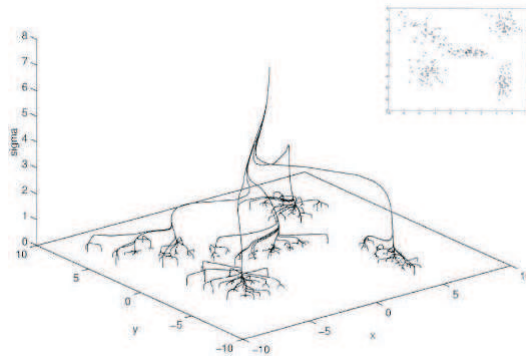


Figure 2. The cluster center curves defined by maxima of scale space data images.

Theorem 2.1 not only shows the simplicity of the cluster center curves that contains no forking, but also implies that for sufficiently small scale, the cluster center curves consist exactly of N branches with each datum being a cluster center. This shows that the deduced approach is independent of initialization. In addition, “zero is a regular value of $\nabla_x P(x, \sigma)$ ” implies the local uniqueness of stationary state of system (7), thus underlies the convergence of the gradient flow.

Based on the expositions above, a complete procedure, called Clustering by Scale Space Filtering (CSSF), for data clustering is developed. See [11] for the details.

The clustering approach made here has many exclusive advantages: Some readily observed advantages, for example, are: (i) The patterns of clustering are highly consistent with the perception of human eyes; (ii) The algorithms thus derived are computationally stable and insensitive to initialization; (iii) They are totally free from solving difficult global optimization problems; (iv) It allows cluster in a partition to be obtained at different scales, and more subtle clustering, such as the classification of land covers, can be obtained; and (v) The algorithms work equally well in small and large data sets with low and high dimensions.

The most promising advantage of the approach, however, is the provision of a cognitive answer to the long-standing problem of *cluster validity*. Cluster validity is a vexing but very important problem in cluster analysis because each clustering algorithm always finds clusters even if the data set is entirely random. While many cluster algorithms can be applied to a given problem, there is in general no guarantee that any two algorithms will produce consistent answers (so, it is why clustering has been regarded as a problem with a part art form and part scientific undertaking [2]).

What is a meaningful (real) cluster? The basis of human visual experience that the real cluster should be perceivable over a wide range of scales leads us to adopt the notion of “lifetime” of a cluster as its validity criterion: A cluster

with longer lifetime is preferred to a cluster with shorter lifetime; The cluster with longest lifetime in the scale space is the most meaningful or real cluster of a data set. We define the lifetime of a cluster and the lifetime of a clustering respectively as follows:

Definition 1. Lifetime of a cluster is the range of logarithmic scales over which the cluster survives, i.e., the logarithmic difference between the point when the cluster is formed and the point when the cluster is absorbed into or merged with other clusters.

Definition 2. Let $\pi(\sigma)$ be the number of clusters in a clustering achieved at a given scale σ . Suppose C_σ is a clustering obtained at σ with $\pi(\sigma) = m$. The σ -lifetime of C_σ is defined as the supremum of the logarithmic difference between two scales within which $\pi(\sigma) = m$.

The reasons why logarithmic scale is used was proven in [11] based on the experimental tests reported in [12], which experimentally justified that $\pi(\sigma)$ decays with scale σ according to $\pi(\sigma) = ce^{-\beta\sigma}$, where c is a constant and β is an unknown parameter.

See Figure 2, by Definitions 1 and 2, the data set D thus contains 5 real clusters, and the partitions of multi-scale representation of D at $\sigma = 1.5 \sim 2.5$ result in the most valid clustering, precisely consistent with the perception of the human eyes.

With the lifetime criterion for cluster validity, we can also answer some questions like whether or not there is a valid structure in a data set. The answer for example is: If $\pi(\sigma)$ takes a constant over a wide range of the scale, a valid structure exists, otherwise, no structure in the data. We can also apply the lifetime criterion to do outlier check. The deduced criterion, say, is that if C_i contains a small number of data and survives a long time, then C_i is an outlier, otherwise, it is a normal cluster.

The scale space based approach thus can provide us an automatic validity check and result in the final most valid clustering. It is also robust to noise in the data.

The scale space approach has provided a unified framework for scale-related clustering techniques derived recently from many other fields such as estimation theory, recurrent signal processing, statistical mechanics, and artificial neural networks. The approach has been extensively applied nowadays as a useful clustering analysis tool in science and engineering. Examples, e.g., see the series of works conducted in Laurence's lab on protein structure identification [13].

2.2. Receptive Field Function Based Approach. This is also a scale space approach, but, different from the last subsection where a continuous scale space is used. I introduce a discrete scale space approach in this subsection.

The continuous scale space approach provides a promising paradigm for clustering. However the high expense is obvious: The scale needs to be discretized and generation of partition at each fixed scale requires an iteration,

too. As a result, two theoretically infinite processes have to be executed in order that a clustering analysis task is accomplished. Moreover, the CSSF can be essentially understood as the Gaussian kernel density based clustering. It works perfectly for the data sets with Gaussian distribution, but not necessarily good (actually very bad sometimes) for non-Gaussian data sets. We hope to generalize the approach to cope with any complex data set, while within a discrete scale space framework.

Some more intrinsic visual mechanism and principles are thus needed. I summarize those preliminary knowledge ([14] [15] [16]) on Visual Information Processing (VIP) and Receptive Field Mechanism first in the following.

2.2.1. VIP and Receptive Field Mechanism. Visual system is a highly complex biological system, which is mainly composed of the retina, primary visual cortex and extra-striate visual cortex. As justified in physiology and anatomy, visual information is transmitted through a certain pathway layer by layer in visual system. Visual information are firstly captured by photoreceptor cells, and then received by ganglion cells. After this retina level, visual information will be transmitted through optic nerves to cross the lateral geniculate and finally reach the primary visual cortex. At the retina and primary visual cortex level, the main function of information processing is *Feature Extraction*. Then the visual information is transmitted into advanced visual cortex for *Feature Integration* or Concept Recognition.

VIP with large connected neurons is very complex, however, it can be easily described and simulated with electrophysiology. Many tests show that each neuron of a certain level corresponds to a spatial region of front layer, where neurons transform visual information to the neuron, and the region is called *Receptive Field* of the neuron (RF) [17] [18]. Each neuron has a certain response pattern (prototype) on the corresponding RF which is called *Receptive Field Function* (RFF). Physiological and biological tests reveal that the shapes of the RF are spatially variant in visual cortex. The RFs of ganglion cells are mainly concentric circle, while the RFs of neurons in visual cortex are more complex.

Given a stimulus $I(x)$, the response of a neuron in primary visual system can be measured by

$$\begin{aligned} f(x; \Theta) &= I(x) * R(x; \Theta) \\ &= \int I(y - x)R(y; \Theta)dy \end{aligned} \quad (8)$$

where $R(x; \Theta)$ is RFF of the neuron, and Θ is a set of parameters. In Eq.(8), $f(x; \Theta)$ is the response of the neuron with stimulus $I(x)$, which is the filtering response and called as a *feature* of $I(x)$.

Different features of a visual input can be extracted by different neurons at different layer. In terms of Eq.(8), this can be equivalently made by different RFFs. Some of the well recognized RFFs in visual system are Gaussian function [11], Gaussian derivative function [19], Gabor function [20], DoG (different

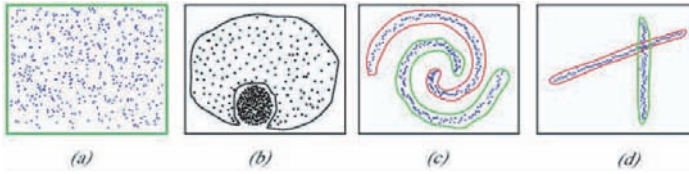


Figure 3. Structure features of data image: (a) No structure (uniform distributed); (b) Density feature; (c) Connectedness feature; (d) Orientation feature.

of Gaussian) function and 3-Gaussian functions. With these different RFFs, various features of visual input can be extracted. These extracted features can then be integrated to form a more complicated feature until a concept is identified.

There are various investigations into feature integration mechanism. However no general solution is resolved up to now. Gestalt principle in psychology, nevertheless, summarizes some very fundamental perception rules of human being, which provides us useful guidance on how features can be organized. Gestalt principle summarizes the perception laws of how the objects (features) are perceived as a whole [21]. It says that human being tend to order our experience in a manner that is regular, orderly, symmetric, and simple. These are formalized respectively as the Law of Proximity, the Law of Continuity, the Law of Similarity, the Law of Closure and the Law of Symmetry. According to these laws, the objects with spatial or temporal proximity, with similar properties such as density, color, shape and texture, with connectedness and orientation features, with symmetric structure, are prone to be perceived as a whole. Our human being tends to group objects to an entity or a closure even it is actually not.

In this view, we can regard the VIP as a procedure of the hierarchical feature extraction dominated by RFFs and the feature integration characterized with Gestalt laws.

2.2.2. Receptive Field Function when Data Image Is Perceived. As the first step towards formalization of a more generic approach for scale space clustering, according to the VIP mechanism, we must first answer what type of RFFs should be adopted in the feature extraction process.

When a data set is observed, the receptive fields of neurons are adaptively formed. In other words, the RFFs are adaptive to the structure features of data image, particularly those of *Density Feature*, *Connectedness Feature* and *Orientation Feature*, as shown in Figure 3. Let χ be the data space. In [8], the following RFF was then suggested:

$$R(x; y, \Theta) = \min_{x \in \Gamma(y), y \in \Gamma(x)} \left\{ \hat{R}(x; y, \Theta), \hat{R}(y; x, \Theta) \right\} \quad (9)$$

where $x, y \in \chi$ is any element,

$$\hat{R}(x; y, \Theta) = \exp \left(-\frac{1}{2} V(x, y; k) A(y; m) V^T(x, y; k) \right) \quad (10)$$

and $\Theta = \{m, k : m, k \text{ are integers}\}$ is a parameter set that is used to confine the neighborhoods of a data set on which the data features are extracted.

In (10), $V(x, y; k)$ is a vector, called *manifold vector*, designed to model the connectedness features of the data image, defined by

$$V(x, x_j; k) = \begin{cases} \frac{x_j - x}{\|x - x_j\|} d_g(x, x_j; k) & x \neq x_j \\ 0 & x = x_j \end{cases}$$

where $d_g(x, y; k)$ is the geodesic metric between x and y , k is a neighborhood size parameter in computation of geodesic distance. It is clear that with such a definition, the manifold vector $V(x, y; k)$ is a vector from x to y with its norm being geodesic metric between x and y . The matrix $A(y; m)$ in (10), called *anisotropy matrix*, is designed to describe the orientation feature of the data set. Assume $\Gamma(x)$ is a chosen m -neighborhood of x , and $A(y; m)$ is then defined as $A(x; m) = B^{-1}(x; m)$ with $B(x; m)$ being the covariance matrix

$$B(x; m) = \frac{\sum_{x_i \in \Gamma(x)} (x - x_i)(x - x_i)^T}{|\Gamma(x)|}$$

where $|\Gamma(x)|$ denotes the number of data contained in $\Gamma(x)$.

It is immediate to see from (9) that the RFF so defined is a symmetric function. The symmetrization procedure in (9) was invented to characterize the density feature of the data set.

As suggested in real visual system, the RFF defined here is spatially localized, anisotropic and orientation selective. When $A(x; m) = I$ and $V(x, x_j; k) = x_j - x$, the RFF defined by (9)–(10) coincides with exactly the Gaussian function (5) used in CSSF.

2.2.3. Discrete Scale Space. With the RFF specified as in (9)–(10), according to VIP mechanism, a set of features of data image can then be extracted by formula (8). In effect, viewed as a data image, each datum of the data is a light point, which projects into χ at a certain location on retina. Suppose that each light point corresponds to a neuron (a photoreceptor cell) on retina photoreceptor level, and, for any $x \in \chi$, it most activates the neuron x' at the t -th layer of VIP. Then the receptive field, $\Gamma(x')$, of x' is a region of pattern space (or photoreceptor cell) which contains x , and RFF of x' is a function $R(x; x', \Theta)$ such that

(i) The nontrivial domain of R coincides with $\Gamma(x')$, i.e.,

$$\Gamma(x') = \{x \in \chi : R(x, x'; \Theta) \neq 0\}$$

(ii) The response of x' is given by

$$f(x; \Sigma) = X * R(x', x; \Theta) = \sum_{x_k \in \Gamma(x)} R(x_k - x; 0, \Theta) x_k \quad (11)$$

Let $X(t)$ be the feature of data set D extracted by VIP at t -th layer, and $X(0)$ simply corresponds to D . Then, $X(t)$ can be expressed as

$$X(t+1) = U(D)X(t), \quad X(0) = D \quad (12)$$

with

$$U(D) = [u_{ij}]_{1 \leq i, j \leq N} = \left[\frac{R(x_i; x_j, k)}{\sum_{s=1}^N R(x_i; x_s, k)} \right]_{1 \leq i, j \leq N} \quad (13)$$

Here $X(t)$ and $X(0)$ are understood as $M \times N$ matrices.

The representation (12) defines a discrete scale space $\{X(t) = U(D)^t D : t \geq 0\}$. We call it the *discrete scale space* of data set D deduced from its feature. Correspondingly, it defines a multi-scale representation of data set D based on its features.

2.2.4. A Visual Clustering Framework (VClust). As in the continuous scale space case, a generic clustering procedure, called VClust in [8], can now be defined as follows:

$$\begin{cases} X(t+1) = U(D)X(t), \quad X(0) = D; \quad t = 1, 2, \dots \\ P_t(X) = G_1(\{X(t)\}_{t=0,1,\dots,t}). \\ P(t) = G_2(\{P_t(X)\}) \end{cases}$$

where operator G_1 is the operation to get partition (clustering) of D at scale t , and G_2 is the operation to read the final most valid clustering of D . Both G_1 and G_2 can be defined completely similar to the case in CSSF.

It can be justified that VClust maintains all the promising properties of CSSF, while dismissing the two crucial drawbacks of CSSF: the high complexity and infeasibility to non-Gaussian data sets. Table 1 provides a direct support for this assertion. It further demonstrates the feasibility, effectiveness and robustness of VClust, as compared with some other competitive clustering techniques.

The data sets in Figure 4 are all with complicated structures (particularly, non-Gaussian). The algorithms used for comparison are all well developed, representatives of respective approaches. Besides CSSF, the Chameleon [22] is derived from the graph-based approach, the spectral-Ng [23] and spectral-shi [24] are spectrum-based, the DBSCAN [25] and the Gaussian Blurring Mean Shift (GBMS) are density-based. The latest LEGClust algorithm [26] based on the information entropy is also tested. In comparison, NMI, the normalized mutual information, was taken as the criterion for measuring the performance of each algorithm.

Table 1. Performance comparison of different clustering algorithms when applied to data sets in Figure 4, measured with NMI.

Methods \ Data sets	(a)	(b)	(c)	(d)	(e)	(f)
VClust	1.0	1.0	1.0	1.0	1.0	1.0
CSSF	0.4357	0.7682	0.4732	0.3269	0.2718	0.4966
Chameleon	1.0	0.9379	0.6824	1.0	0.5991	0.6425
Spectral-Ng	1.0	0.8326	1.0	0.4157	0.4921	0.7103
Spectral-Shi	0.8726	0.9721	1.0	0.7892	0.5283	0.6947
LEGClust	1.0	0.9846	0.4919	1.0	0.3721	1.0
DBScan	0.4115	0.7286	0.4351	0.3924	0.2362	0.4529

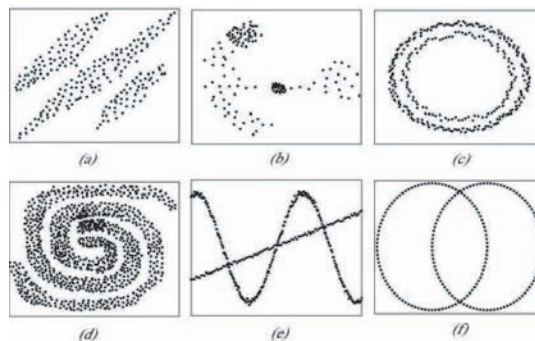


Figure 4. Some data sets with complicated structures used for comparison of different clustering techniques.

2.3. Neural Coding Based Approach. The scale space approach for clustering has been extended to classification problems. A similar idea was also used to do model selection for Gaussian Support Vector Machine, and in particular, a very useful data-driven formulae for Gaussian width parameter σ was discovered [27] (cf. Figure 5). Nevertheless, a much more significant extension of the scale space approach is the development of a new methodology: A neural coding based approach for data modeling.

In our brain, a neuron receives information from other neurons and processes/ responses through integrating information from other neurons, then sends the integrated information to others. We can generally classify the neurons into two types: the *stimulative neurons* (understood as the photoreceptor cells in visual system), which stimulate other neurons, and the *active neurons*, which receive information from stimulative neurons and produce response. Let $X = \{X_i\}_{i=1}^M$ be stimulative neurons and $Y = \{Y_j\}_{j=1}^N$ the active neurons, where X_i is a canonical stimulus, and Y_j is the receptive field function of neuron j that characterizes its response property. Let $e_j(X_i)$ denote the activation extent of active neuron j when the stimulative neuron i is stimulated, and let

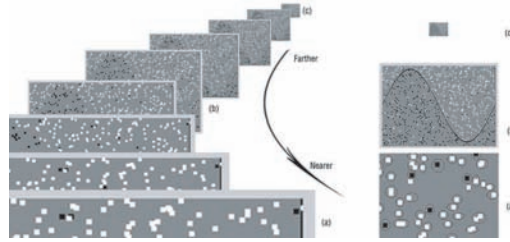


Figure 5. When a data set is observed from different distances, different structures are perceived.

$S(X_i, Y_j)$ denote the matching degree, or say, the similarity between the stimulus X_i and RFF Y_j . Then there holds a very fundamental coding principle: For any stimulative input, we response always maximally. That is to say, the neural coding in brain system is always such that for every input X , it maximizes the following response function

$$E(Y) = \left\{ \sum_{i,j} e_j(X_i)S(X_i, Y_j) \right\} \tag{14}$$

In preliminary visual system, neural coding is basically linear. Thus, let $f(X) = (f_1(X_1), f_2(X_2), \dots, f_M(X_M))^T$ be a stimulation mode, and $R(Y_j, X; \Theta)$ be the RFF of neuron j . Then we have [16] [28]

$$S(X_i, Y_j) = |f_j(X; \Theta)|$$

and

$$e_j(X_i) = \begin{cases} \frac{f_j(X; \Theta)}{|f_j(X; \Theta)|}; & \text{if } X_i \in \Gamma(Y_j) \\ 0; & \text{otherwise} \end{cases} \tag{15}$$

where $\Gamma(Y_j)$ is the receptive field of neuron j and $f_j(X; \Theta)$ is the response of Y_j given by

$$f_j(X; \Theta) = f(X) * R(Y_j, X; \Theta) = \sum_{x_k \in \Gamma(Y_j)} R(Y_j - X_k; 0, \Theta) f_k(X_k).$$

In this case, the response function (14) becomes

$$E_\Theta(Y) = \sum_{i,j} e_j(X_i)S(X_i, X_j) = \sum_j f(X) * R(Y_j, X; \Theta).$$

If one takes the parameter Θ be σ , then $\{E_\sigma(Y) : \sigma \geq 0\}$ gives the continuous scale space, and maximization of the response function directly leads to CSSF.

We naturally consider the nonlinear coding case. Different from linear case, nonlinear neural coding theory [29] [30] views the relationship between stimulative neurons and active neurons nonlinear. The theory says that the response

of a neuron is accomplished in two stages. In the first stage, as linear case, it integrates all stimuli from input cells, according to linear coding

$$U_{ij}^{(1)} = f(X) * R(Y_j, X; \Theta) \tag{16}$$

and in the second stage, it goes to two successive independent nonlinear procedures: within-pathway-nonlinearity and the divisive gain control nonlinearity,

$$e_j(X_i) = \frac{[U_{ij}^{(1)}]^p}{[C_2^p + \sum_k U_{ik}^{(1)}]^p} \times \frac{[U_{ij}^{(1)}]^r}{[C_1^r + U_{ij}^{(1)}]^r} \tag{17}$$

where C_1 and C_2 are semi-saturation constants; r, p are the normalization parameters, controlling the degree of increasing response to the most sensitive stimulus, and decreasing the effect of insensitive stimulus.

With a neural coding scheme, a data modeling problem can be tackled in the subsequent way: Let $X = \{X_i\}_{i=1}^N$ be the data set, and $Y = \{Y_j\}_{j=1}^M$ be the solution we would like to find. We model the data modeling problem as an optimization problem

$$\max_Y \left\{ E(Y) = \sum_{i,j} e_j(X_i) S(X_i, Y_j) \right\} \tag{18}$$

through defining an appropriate similarity measure $S(X_i, Y_j)$, where $e_j(X_i)$ is any specified neural coding.

Examples are as follows:

Let $X = \{X_i\}_{i=1}^N$ be a data set with M clusters. Y_j is centroid of j -th cluster; d_{kj} is distance between X_k and the centroid Y_j of the j -th cluster; $g(\frac{1}{d_{kj}})$ is similarity between X_k and the centroid Y_j of the j -th cluster, and $g(\cdot)$ is any an increasing function. Then, (18) degenerates to CSSF when $e_j(X_i)$ is taken as the linear neural coding.

The Improved Probabilistic C-Means [31] provides an example with the nonlinear coding, where $S(X_i, Y_j) = 1/d_{kj}$. The technique improves substantially on Fuzzy C-means, noise clustering, and possible C-means. A comparison between PCM and its neural coding based counterpart is shown in Figure 6.

I suggest a methodology for solving a generic regression problem in section 4.

3. $L_{1/2}$ Regularization Theory

Latent variable analysis aims to identity the intrinsic variables from observation, while *Neural Coding* in neurobiology is concerned with how sensory and other information is represented in the brain by neurons. The aims of these two seemingly irrelevant subjects coincide with each other. So, borrowing the

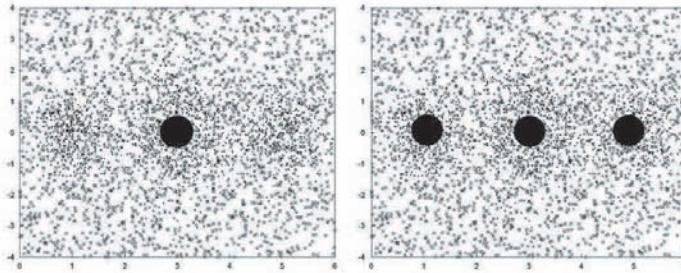


Figure 6. Comparison of clustering results with PCM and its neural coding based revision, where \cdot is data point, \bullet is cluster center and \times denotes noisy data point.

methodology from neural coding can shed light on the way we solve a latent variable analysis problem.

The most striking feature of neural coding is its sparsity, which means that only a relatively small set of neurons in brain have strong response when a stimulus is received. Substantial biological evidence for such property occurs at earlier stages of processing across a variety of organisms, for example, auditory system of rats, visual system of primates and layer 6 in the motor cortex of rabbits [32]. Olshausen and Field [33] developed a mathematical model of sparse coding of natural image in visual system. Validated by neurobiological experiments, the receptive fields of simple cells in mammalian primary visual cortex are characterized as being spatially localized, oriented and bandpass. They demonstrated that such receptive fields emerge in their model when only the two global objectives are placed on a linear coding of natural images. In this case, the information of natural image is preserved, and the representation is sparse. Their model reads as

$$\min \{ \|I - Bx\|_2^2 + \lambda p(x) \} \quad (19)$$

where I denotes the grey scale value of an image patch, B denotes the basis matrix consisted of the simple-cell receptive fields that are learned from samples, x is the sparse representation of natural image, and $p(x)$ is the sparse-promoting function which could be chosen as $-e^{-x^2}$, $\log(1+x^2)$ or $|x|_1$. The research conducted by Olshausen and Field is important. It shows not only that the neural coding in primary visual processing (mainly with simple cells) does be sparse and can be linear, but also that the visual sparse coding can be simulated and found via a mathematical model. Such study has been generalized to complex cells in [34]. We observe that the model (19) is nothing else but a regularization scheme for solution of general sparsity problems.

Mathematically, a sparsity problem can be described as a problem of finding sparse solution(s) of an representation or a underdetermined equation. Besides the neural coding problem introduced above, variable selection, graphical modeling, error correction, matrix completion and compressed sensing (particularly,

signal recovery and image reconstruction) are all the typical examples. All these problems can be described as the following:

Given a $M \times N$ matrix A and a procedure of generating observation y such that $y = Ax$, we are asked to recover x from observation y such that x is of the sparsest structure (that is, x has the fewest nonzero components).

The problem then can be modeled as the following L_0 optimization problem

$$\min \|x\|_0 \text{ subject to } y = Ax \quad (20)$$

where (and henceforth) $\|x\|_0$, formally called L_0 norm, is the number of nonzero components of x . Obviously, when $M \ll N$ (namely, the high dimension/small sample case), the sparsity problems are seriously ill-posed and may have multiple solutions. A common practice is then to apply regularization technique for the solution(s). Thus, the sparsity problems can be frequently transformed into the following so called L_0 regularization problem

$$\min_{x \in R^N} \{ \|y - Ax\|_2^2 + \lambda \|x\|_0 \} \quad (21)$$

where $x = (x_1, \dots, x_N)^T \in R^N$ and $\lambda > 0$ is a regularization parameter.

The L_0 regularization can be understood as a penalized least squares with penalty $\|x\|_0$, in which parameter λ functions as balancing the two objective terms. The complexity of the model is proportional with the number of variables, and solving the model generally is intractable, particularly when N is large (It is a NP-hard problem, see, e.g., [35]). In order to overcome such difficulty, many researchers have suggested to relax L_0 regularization and instead, to consider the following L_1 regularization

$$\min_{x \in R^N} \{ \|y - Ax\|_2^2 + \lambda \|x\|_1 \} \quad (22)$$

where $\|x\|_1$ is the L_1 norm of R^N .

The use of the L_1 norm as a sparsity-promoting function appeared early in 1970's. Taylor, Banks and McCoy [36] proposed the use of L_1 norm to deconvolve seismic traces by improving on earlier ideas of Claerbout and Muir [37]. This idea was latter refined to better handle observation noise [38], and the sparsity-promoting nature of L_1 regularization was empirically confirmed. Rigorous uses of (22) began to appear in the late-1980's, with Donoho and Stark [39] and Donoho and Logan [40] quantifying the ability to recover sparse reflectivity functions. The application areas of L_1 regularization began to broaden in the mid-1990's, as the LASSO algorithm [4] was proposed as a method in statistics for sparse model selection, Basis Pursuit [5] was proposed in computational harmonic analysis for extracting a sparse signal representation from highly overcomplete dictionaries, and a technique known as total variation minimization was proposed in image processing [41, 42].

The L_1 regularization has now become so widespread that it could arguably be considered the “modern least squares” [6]. This is promoted not only by the sparsity-promoting nature of L_1 norm and the existence of very fast algorithms for solution of the problem, but also by the fact that there are conditions guaranteeing a formal equivalence between the combinatorial problem (21) and its relaxation (22)[43].

The L_1 regularization is, however, still far from satisfaction. For many applications, the solutions of the L_1 regularization are less sparse than those of the L_0 regularization. It can not handle the collinearity problem, and may yield inconsistent selections [44] when applied to variable selection; It often introduces extra bias in estimation [45], and can not recover a signal or image with the least measurements when applied to compressed sensing. Thus, a mandatory and crucial question arises: Can the sparsity problems be solved by some other means? As shown below, I suggest the use of following alternative: the $L_{1/2}$ regularization

$$\min_{x \in R^N} \left\{ \|y - Ax\|_2^2 + \lambda \|x\|_{1/2}^{1/2} \right\}. \quad (23)$$

3.1. Why $L_{1/2}$ Regularization? We may seek other sparsity-promoting functions $p(x)$ to replace $\|x\|_1$ in (22). The generality of polynomial functions then naturally leads us to try $p(x) = \|x\|_q^q$ with $q \geq 0$. The geometry of Banach space implies, as suggested also by the classical least squares, $q > 1$ may not lead to the sparsity-promoting property of functions $p(x)$. So $q \in (0, 1]$ are only candidates. In consequence, the L_q regularizations have been suggested [46], that is, instead of L_1 regularization (22), using

$$\min_{x \in R^N} \left\{ \|y - Ax\|_2^2 + \lambda \|x\|_q^q \right\} \quad (24)$$

where $\|x\|_q$ is the L_q quasi-norm of R^N , defined by $\|x\|_q = \left(\sum_{i=1}^N |x_i|^q \right)^{1/q}$.

The problem is which q is the best? By using the phase diagram tool introduced by Donoho and his collaborators [47, 48], Wang, Guo and Xu [49] provided an affirmative answer to the question. Through applying the L_q regularizations to the typical sparsity problems of variable selection, error correction and compressed sensing with the reweighted L_1 technique suggested in [46], they experimentally showed that the L_q regularizations can assuredly generate more sparse solutions than L_1 regularization does for any $q \in (0, 1)$, and, while so, the index $1/2$ somehow plays a representative role: Whenever $q \in [1/2, 1)$, the smaller q , the sparser the solutions yielded by L_q regularizations, and, whenever $q \in (0, 1/2]$, the performances of L_q regularizations have no significant difference (cf. Figures 7 and 8). From this study, the special importance of $L_{1/2}$ regularization is highlighted.

Figure 7 shows how sparsity (k/M , k is the number of nonzeros in x , and M is number of rows in A) and indeterminacy (M/N) affect the success of L_q regularizations. The contours indicate the success rates for each combination of $\{k, M, N\}$, where red means the 0% success, blue means 100% success, the

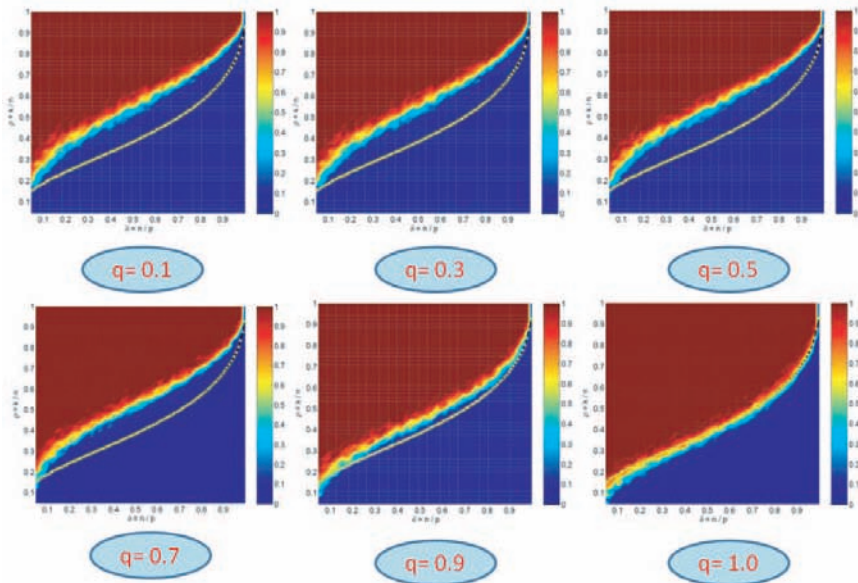


Figure 7. Phase diagrams of L_q ($q = 0.1, 0.3, 0.5, 0.7, 1.0$) when applied to a sparsity problem (signal recovery).

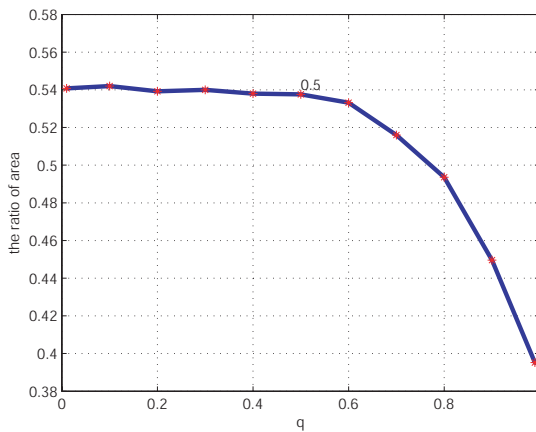


Figure 8. The interpolated success percentage curve of L_q regularizations, when applied to signal recovery.

belt area means others.¹ In the figure, the commonly occurred yellow curves are *Theoretical L_1/L_0 Equivalence Threshold Curve* found by Donoho [47, 48],

¹In the print version Figure 7 appears as a gray-scale picture and does not show colors. The coloured figure can be referred in the corresponding electronic version.

which consists of the values at which equivalence of the solutions to the L_1 and L_0 regularizations breaks down. The curve delineates a phase transition from the lower region where the equivalence holds, to the upper region, where the equivalence does not hold. Along the x -axis the level of underdeterminedness decreases, and along the y -axis the level of sparsity of the underlying model increases. The belt area in each case roughly defines a curve, which can be referred to as $A L_q/L_0$ Equivalence Threshold Curve. Then, Figure 7 exhibits that the L_q/L_0 equivalence threshold curves are always above of the theoretical L_1/L_0 equivalence threshold curve, showing a preferable sparsity-promoting nature of L_q regularizations.

Figure 8 shows the interpolated success percentage curve of L_q regularizations. Here the success percentage for a regularization is defined as the ratio of the blue region in the whole region of the phase plane. It is very clearly demonstrated that the $L_{1/2}$ regularization is nearly best, and therefore, can be taken as a representative of L_q regularizations with all q in $(0, 1]$.

Another reason why $L_{1/2}$ is selected is due to its privilege of permitting fast solution, as that for L_1 regularization.

3.2. How $L_{1/2}$ Fast Solved? The increasing popularity of L_1 regularization comes mainly from the fact that the problem is convex and can be very fast solved. The $L_{1/2}$ regularization, however, is a nonconvex, non-smooth and non-Lipschitz optimization problem. There is no directly available fast algorithm for the solution. Fortunately, I and my PhD students recently found such a fast algorithm for $L_{1/2}$ regularization problem [50].

The found fast algorithm is an iterative method, called the iterative *half* thresholding algorithm or simply *half* algorithm, which reads as

$$x_{n+1} = H_{\lambda_n \mu_n, \frac{1}{2}}(x_n + \mu_n A^T(y - Ax_n)) \quad (25)$$

Here $H_{\lambda \mu, \frac{1}{2}}$ is a diagonally nonlinear, thresholding operator specified as in Theorem 3.1, μ_n are parameters to control convergence and λ_n are adaptive regularization parameters. The derivation of the algorithm is based on a fundamental property of $L_{1/2}$ regularization problem, the thresholding representation property, as defined and proved in [50].

Theorem 3.1 ([50]). *The $L_{1/2}$ regularization permits a thresholding representation, i.e., there is a thresholding function h such that any of its solution, x , can be represented as*

$$x = H(Bx) \quad (26)$$

where H is a thresholding operator deduced from h and B is a linear operator from R^N to R^N . More specifically, one can take in (26) that for any fixed λ , $\mu > 0$,

$$B(x) = B_\mu(x) = x + \mu A^T(y - Ax) \quad (27)$$

$$H(x) = H_{\lambda \mu, 1/2} = (h_{\lambda \mu/2}(x_1), h_{\lambda \mu/2}(x_2), \dots, h_{\lambda \mu/2}(x_N))^T \quad (28)$$

where $h_{\lambda\mu/2}(t)$ is defined by

$$h_{\lambda\mu/2}(t) = \begin{cases} \frac{2}{3}t \left(1 + \cos\left(\frac{2\pi}{3} - \frac{2\varphi_\lambda}{3}\right)\right), & |t| > \frac{3}{4}(\lambda\mu)^{\frac{2}{3}} \\ 0, & \text{otherwise} \end{cases}. \quad (29)$$

with

$$\varphi_{\lambda\mu} = \arccos\left(\frac{\lambda\mu}{8} \left(\frac{|t|}{3}\right)^{-\frac{3}{2}}\right). \quad (30)$$

With the thresholding representation (27)-(30), the iterative algorithm (25) then can be seen as the successive approximation for common fixed point of operators H and B . The diagonal nonlinearity of the thresholding operator $H_{\lambda\mu,1/2}$ makes it possible to implement the iteration (25) component-wisely. The high efficiency and fastness of the *half* algorithm thus follows. The thresholding representation (27)-(30) also has other meaningful consequences, say, it can be applied to justify the finiteness of local minimizers of $L_{1/2}$ regularization problem. This is an unusual, very useful property of a nonconvex problem, which distinguishes the $L_{1/2}$ regularization strikingly from other optimization problems.

Theorem 3.1 can also be used to derive an alternative theorem on solutions of $L_{1/2}$ regularization problem. From the theorem, some almost optimal parameter setting strategies can then be suggested. For example, the following parameter-setting strategy in (25) has been recommended in [50]:

$$\mu_n = \frac{(1 - \varepsilon)}{\|A\|^2}, \lambda_n = \frac{4}{3} \|A\|^2 |[B_{\mu_n}(x_n)]_k|^{3/2}$$

where ε is any small fixed positive constant.

The half algorithm has been applied to a wide rang of applications associated with signal recovery, image reconstruction, variable selection and matrix completion in [50]. The applications consistently support that the algorithm is a fast solver of $L_{1/2}$ regularization, comparable with and corresponding to the well known iterative *soft* thresholding algorithm for L_1 regularization.

It is interesting to ask a question here: *Is there other index q in $(0, 1)$, except $1/2$, which permits a thresholding representation for L_q regularization?* In [50], an observation was made to guess that only with $q = 1, 2/3, 1/2$, L_q regularization admits a (27)-(28) like representation. A general answer is still unknown.

3.3. What Theory Says? The following theorem justify the convergence of the iterative half thresholding algorithm.

Theorem 3.2 ([51]). *Assume $\mu_n \in (0, \|A\|^{-2})$ and λ_n is monotonically decreasing to a fixed $\lambda \geq 0$. Then the half thresholding algorithm converges to a local minimum of $L_{1/2}$ regularization problem (23). Furthermore, if any k*

columns of A (denoted by A_k) are linear independent, and μ_n, λ_n satisfies

$$\mu_n < 1/s_{\min}(A_k^T A_k); \lambda_n = \frac{4}{3} \|A\|^2 |[B_{\mu_0}(x_n)]_k|^{3/2}$$

where $s_{\min}(A_k^T A_k)$ is the smallest eigenvalue of matrix $A_k^T A_k$, then the algorithm converges to a k -sparsity solution of the $L_{1/2}$ regularization.

For the proof of Theorem 3.2, we refer to [51]. The proof depends upon a very careful analysis on the thresholding operator H defined as in (28). In the considered case, H is deduced intrinsically from the resolvent of gradient of $\|\cdot\|_{1/2}^{1/2}$. Unlike the L_1 regularization case, where $\|x\|$ is a convex function, so that $\partial(\|x\|)$ is maximal monotone and the resolvent operator $(I + \partial(\|x\|))^{-1}$ is nonexpansive. In the $L_{1/2}$ regularization case, however, $\|x\|_{1/2}^{1/2}$ is non-convex and non-Lipschitz, so that the resolvent operator $(I + \lambda\partial(\|\cdot\|_{1/2}^{1/2}))^{-1}$ is only restrainedly defined and is not nonexpansive.

When applying a nonconvex sparsity-promoting function as a penalty, a problem we commonly worry about is the local minimum problem: The algorithm might only converge to a local minimum. Sometimes, this becomes the reason why a nonconvex regularization scheme would not be adopted in practice. However, due to the finiteness of local minima of $L_{1/2}$ regularization problem, Theorem 3.2 provides a promise that it can find the global optimal solution provided we run the algorithm many times with uniformly distributed random initial values.

With application to latent variable analysis or compressed sensing, the independence condition in Theorem 3.2 can be very intuitively explained. In the later case, for example, we have $A = \Psi\Phi$ with Ψ being $M \times N$ sampling matrix and Φ a basis matrix. Theorem 3.2 then says that a k -sparsity signal x can be recovered from M measurements with $M \ll N$ only if the sampling Ψ is such that every k columns of $\Psi\Phi$ are independent. This is obviously reasonable, and in fact constitutes the basis how the sampling should be taken.

The next theorem shows the condition when $L_{1/2}/L_0$ equivalence occurs. Recall that a matrix is said to possess *Restricted Isometry Property* (RIP) if

$$(1 - \delta_k) \|x\|_2^2 \leq \|Ax\|_2^2 \leq (1 + \delta_k) \|x\|_2^2 \text{ whenever } \|x\|_0 \leq k$$

The restricted isometry constant $\delta_k(A)$ is the smallest constant for which the RIP holds for all k -sparsity vector x .

Theorem 3.3 ([52]). *Any k -sparsity vector x can be exactly recovered via $L_{1/2}$ regularization if $\delta_{2k}(A) < 1/2$.*

Note that Candès and Tao showed the L_1/L_0 equivalence when $\delta_{3k}(A) + \delta_{4k}(A) < 2$ [53], and later Candès relaxed to $\delta_{2k}(A) < \sqrt{2} - 1 \approx 0.414$ [54], Foucart and Lai [55] verified the L_q/L_0 equivalence under the condition

$\delta_{2k}(A) < 2(3 - \sqrt{2})/7 \approx 0.4531$. Theorem 3.3 provided a looser $L_{1/2}/L_0$ condition $\delta_{2k}(A) < 0.5$.

It is interesting to compare the convergence condition in Theorem 3.2 with the $L_{1/2}/L_0$ equivalence condition $\delta_{2k}(A) < 1/2$ in Theorem 3.3. In effect, the condition “any k columns of A (denoted by A_k) are linear independent” in Theorem 3.2 can be reformulated as $\delta_k(A) < 1$, which is much looser than $\delta_{2k}(A) < 1/2$. This leads to a natural question: Whether Theorem 3.3 is still true when the condition $\delta_{2k}(A) < 1/2$ is relaxed to $\delta_k(A) < 1$. I guess this is the case. However the real answer is open.

Theorems 3.4 and 3.5 below summarize two important statistical properties of $L_{1/2}$ regularization. Consider the linear model

$$y = X^\top \beta + \varepsilon, E\varepsilon = 0, Cov(\varepsilon) = \sigma^2 I \tag{31}$$

where $y = (y_1, y_2, \dots, y_M)^\top$ is an $M \times 1$ response vector, $X = (X_1, X_2, \dots, X_M)$ ($X_i \in R^N$) and $\beta = (\beta_1, \beta_2, \dots, \beta_N)^\top$ is unknown target vector, ε is a random error and σ is a constant. For any $1 \leq k \leq N$, let β_k denote the k -sparsity vector of β , that is, the vector whose k components coincide with those of β whenever the corresponding components β_i are among the k largest ones in magnitude, and other $N - k$ components are zeros. Note that when $L_{1/2}$ regularization is applied to problem (31), its solution is given by

$$\hat{\beta} = \arg \min_{\beta \in R^N} \left\{ \sum_{i=1}^M (\beta^\top X_i - y_i)^2 + \lambda \|\beta\|_{1/2} \right\}. \tag{32}$$

Theorem 3.4 ([56]). *Let β^* be any solution of (31) and $\hat{\beta}$ any solution of (32). Then for any $a > 0$ and under some mild conditions, for any $\delta \in (0, 1)$ with probability larger than $1 - \delta$, there holds the following estimation*

$$\left\| \hat{\beta} - \beta^* \right\|_2 \leq O(\lambda\sqrt{k} + \|\beta^* - \beta_k^*\|_2 + \|\beta^* - \beta_k^*\|/\sqrt{l}) \tag{33}$$

where l is any constant satisfying $k \leq l \leq (N - k)/2$, t is constant satisfying $0 < t \leq C(k, l)$, $\lambda \geq \frac{8(2-t)}{t} \max\{\sqrt{C_0}, 1\} \left(a\sigma \sqrt{\frac{2}{M} \ln \frac{2N}{\delta}} \right)$, and β_k^* is the k -sparsity vector of β^* .

The estimation (33), which measures how well the solution yielded by $L_{1/2}$ regularization approximates the target solution, can be shown to be optimal in the sense of achieving an ideal bound. It reveals that even though the number of samples is much smaller than that of the dimension of parameters, the solutions of $L_{1/2}$ regularization can achieve a loss within logarithmic factor of the ideal mean squared error one would achieve with an oracle. This shows that $L_{1/2}$ regularization is good at tackling the high-dimension/small sample problems.

One of direct applications of model (31) is variable selection. Fan [57] has ever suggested a standard of measuring how well an algorithm performs variable

selection via the model (31). That is the so called oracle property: An ideal variable selection algorithm should automatically set the irrelevant variables to zero. The following Theorem 3.5 shows that $L_{1/2}$ regularization has the oracle property. Without loss of generality, we assume that the target vector $\beta^* = (\beta_1^{*\top}, \beta_2^{*\top})^\top$ with β_1^* having no zero component and $\beta_2^* = 0$.

Theorem 3.5 ([56]). *If $\lambda = o(M^{1/4})$, then the $L_{1/2}$ regularization possesses the following properties:*

- (i) *Consistency in variable selection: $\lim_{M \rightarrow \infty} P(\widehat{\beta}_2 = 0) = 1$;*
- (ii) *Asymptotic normality: $\sqrt{M}((\widehat{\beta}_1 - \beta_1^*) \rightarrow_d N(0, \sigma^2 C)$.*

Theorem 3.5 shows that $L_{1/2}$ regularization is an idea variable selection method.

3.4. How Useful? The $L_{1/2}$ regularization has been applied to solve various sparsity problems, and among them compressed sensing is a very typical example. The compressed sensing (CS) has been one of the hottest research topics in recent years. Different from the traditional Shannon/Nyquist theory, CS is a novel sampling paradigm that goes against the common wisdom in data acquisition. Given a sparse signal in a high dimensional space, one wishes to reconstruct the signal accurately from a number of linear measurements much less than its actual dimensionality. The problem can be modeled as the sparsity problem (20) with

$$A = \Psi\Phi \tag{34}$$

where Ψ is a $M \times N$ sampling matrix, Φ is a $N \times N$ basis matrix and A is called a sensing matrix. A very fundamental requirement here is $M \ll N$. Given fewer measurements $y = Ax = \Psi\Phi x$ on a signal, we then are asked to reconstruct the signal x from y .

Let us take MRI as a concrete example. In MRI, the scanner takes slices from two dimensional Fourier domain of an image [58]. In order to reduce scan time and the exposure of patients to electromagnetic radiation, it is desirable to take fewer measurements. In this case, we hope to exploit the sparsity of the image in the Fourier or wavelet domain for reconstructing the image from fewer measurements. In application, the measurements are normally accomplished via sampling the image in its Fourier spectre domain. According to [59], when sampling in this way on L rays in the domain and taken a Gaussian sampling on each ray, the resultant sensing matrix is Gaussian random, satisfying the so called RIP condition ([60]) so that the image can be exactly reconstructed.

We experimented with the standard Shepp-Logan phantom, a 256×256 MRI image shown as in Figure 9(a). The *half* thresholding algorithm (*Half*) in [50] and the Reweighted L_1 method (*RL1*) in [61] for $L_{1/2}$ regularization were applied in comparison with L_1 regularization. In implementation of L_1 regularization, the well known L_1 magic algorithm (*L1magic*) [62] and the soft thresholding algorithm (*soft*) [63] were applied, while the hard thresholding

algorithm (*hard*) [64] was adopted to perform L_0 regularization. We ran the simulations by varying the measurements from $L = 70$ to 40. The simulations reveal that before $L = 60$, all the algorithms succeeded in exactly recovery. Nevertheless, when L reduced to under 55, the L_1 regularization algorithms failed, but $L_{1/2}$ algorithm still succeeded, as listed in Table 2. It is seen that when sampling are taken on 52 rays, the *half* and *hard* algorithms both can recover the image, with *half* having the highest precision. When we reduce the sampling rays to $L = 40$, the algorithms L_1 magic, *RL1*, *soft* and *hard* are all perform very poor, while the *half* algorithm reconstructed the image with a very high precision, which distinguishes the *half* from other competitive algorithms very obviously. See Figure 9 and Table 2 for the reconstructed images.

Table 2. The image reconstruction results

L	Method	MSE	Time	L	Method	MSE	Time
40	LImagic	fail	∞	52	LImagic	9.3458 (fail)	1008.72
	RL1	fail	∞		RL1	4.6881 (fail)	3650.24
	soft	8.2469	882.0637		soft	0.9812 (fail)	1795.5
	hard	15.3978	1038.1		hard	7.98e-6	105.0087
	half	5.30e-7	2738.8		half	3.15e-7	181.6311

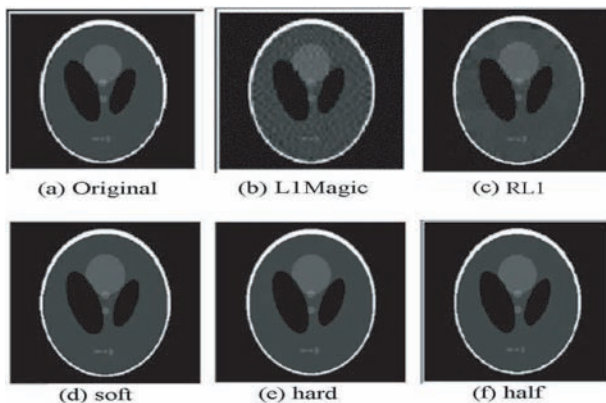


Figure 9. The reconstructed images by different regularization algorithms when $L = 52$.

This application demonstrates the outperformance of $L_{1/2}$ regularization over L_1 regularization. Such outperformance of $L_{1/2}$ regularization is also consistently supported by other experiments and applications.

Before ending this section, I would like to make an observation on overall features of L_p regularizations when p takes over entire real axis. The L_1 regularization is well known, that has the sparsity-promoting property and leads to a convex problem easy to be solved; When $p > 1$, the L_p regularizations

have not maintained the sparsity-promoting property any more, but possess a stronger convex property (uniformly convex property) and the problems get more and more easily solved; While when $p < 1$, the L_p regularizations have a stronger and stronger sparsity-promoting property, but have not maintained the convex property any more, and the problems get more and more difficult to be solved. This demonstrates a threshold or center position of $p = 1$ over which the sparsity-promoting property, the convex property and the easiness of solution all break down. In this sense, we can see that L_1 regularization just is the scheme with the weakest sparsity-promoting property and the weakest convex property (so, the weakest scheme), but more positively, it provides the best convex approximation to L_0 regularization and the best sparsity-promoting approximation to L_2 regularization. It is well known that all p with $1 \leq p \leq \infty$ constitute a complete system within which $p = 2$ plays a very special role. I therefore guess that $p = 1/2$ might somehow plays also a special role in another system $\{p : 0 \leq p \leq 1\}$. The study on $L_{1/2}$ regularization is providing a direct support to this view.

4. Concluding Remarks

Data modeling is emerging as a cross-disciplinary, fast developing discipline. New ideas and new methodologies have been called for. In this talk I have introduced two new methodologies which seems meaningful and potentially important. Along the line of research in this talk, however, there are many problems open. As final remarks, I list some of those problems for further study.

Problem 1. Towards $L_{1/2}$ regularization theory

I first summarize the open questions I have raised in exposition of the last section. Firstly, *Does any other L_q regularizations rather than $q = 1/2$ permit a thresholding representation?* Following the idea in [50], it is not difficult to say “yes” for $q = 2/3$, but how about for other q in $(0, 1)$? The answer for this question is meaningful to development of other more effective sparsity-promoting algorithms. Secondly, we have shown the superiority and representative of $L_{1/2}$ regularization among L_q regularizations with $q \in (0, 1)$ based on a phase diagram study. This is certainly an experiment based approach. So, *Does the representative role of $L_{1/2}$ regularization can be justified in a somewhat theoretical way?* An tightly relevant question arises from an observation of phase diagrams in Figure 7. The belt area in each diagram roughly defines an empirical L_q/L_0 equivalence threshold curve, which fundamentally characterizes the sparsity-promoting capability of each corresponding regularization scheme. *Does there exist theoretical L_q/L_0 equivalence threshold curves for any L_q regularization? Are those L_q/L_0 equivalence threshold curves in Figure 3 the theoretical ones?* Thirdly, we have proved the convergence of the $L_{1/2}$ regularization algorithm (half thresholding algorithm) under the condition $\delta_k(A) < 1$, while justified the

$L_{1/2}/L_0$ equivalence under the much tighter condition $\delta_{2k}(A) < 1/2$. A natural question is: *Whether $\delta_k(A) < 1$ is also a sufficient condition for $L_{1/2}/L_0$ equivalence?*

Problem 2. Towards geometry of $L_{1/2}$ space

Let $\Gamma = \{p : 1 \leq p \leq \infty\}$. It is well known that with any $p \in \Gamma$, L_p space (understood either as function spaces or as sequence spaces) is a Banach space, and, within the duality framework $\frac{1}{p} + \frac{1}{q} = 1$, L_2 is self-dual and can be characterized with Parallelogram Law or equivalently Binomial Formula

$$\|x + y\|_2^2 = \|x\|_2^2 + 2\langle x, y \rangle + \|y\|_2^2, \forall x, y \in L_2$$

It is such characteristic identity law that makes many mathematical tools available, say, Fourier analysis and wavelet analysis. The Hilbert characteristic identity law was extended by Xu and Roach [65] into Banach space setting, which states that a Banach space X is uniformly convex if and only if there is a positive function σ_p such that

$$\|x + y\|^p \geq \|x\|^p + p\langle J_p x, y \rangle + \sigma_p(x, y) \|y\|^p, \forall x, y \in X \quad (35)$$

and it is uniformly smooth if and only if there is a positive function δ_p such that

$$\|x + y\|^p \leq \|x\|^p + p\langle J_p x, y \rangle + \delta_p(x, y) \|y\|^p, \forall x, y \in X \quad (36)$$

where J_p is the duality mapping with the gauge t^p/p , σ_p is uniquely determined by the convexity modulus of X and δ_p uniquely determined by the smoothness modulus of X . These Banach characteristic inequality laws admit two sets of explicit homogenous forms in L_p spaces with $1 < p < \infty$, since in this case, the spaces are both uniformal convex and uniformly smooth. A space with two or one of the two inequalities of the form (35) and (36) is very fundamental. In the case, many quantitative analysis and mathematical deductions then can be made in the space.

Let $\Sigma = \{p : 0 \leq p \leq 1\}$. It is then known that for any $p \in \Sigma$, L_p is not a Banach space, but is a quasi-normed space. Promoted by studying L_q regularization, I would like to know the geometry of quasi-normed spaces L_p with $p \in \Sigma$. More particularly, due to the speciality of $L_{1/2}$ regularization, I want to ask: *Does there exist a some kind of duality framework (say, $p + q = 1$) such that within the framework $L_{1/2}$ space is self-dual?* Also, for studying $L_{1/2}$ regularization purpose, I would like to know: *Does there hold some kinds of characteristic laws like (35) and (36)?* If so, the convergence of $L_{1/2}$ algorithm and $L_{1/2}/L_0$ equivalence can be done in a straightforward way.

Problem 3. Towards a neural coding based machine learning theory

The neural coding based data modeling suggests also a new paradigm for solution of generic learning problem. In effect, assume X is a feature space,

Y is a response space and $Z = X \times Y$ is the data space. For a given set of training examples $D = \{z_i = (x_i, y_i)\}_{i=1}^M$ which are drawn *i.i.d* from an unknown distribution P on Z , and a preset family of functions $F = \{f : X \rightarrow Y\}$, a learning problem is asked to seek a function f^* in F such that the expected risk $E(f)$ is minimized, that is,

$$f^* = \arg \min_{f \in F} E(f) = \int l(f, z) dP$$

The ERM principle suggests to use the empirical error $E_{emp}(f)$ to replace $E(f)$ and find f^* through

$$f^* = \arg \min_{f \in F} \left\{ \frac{1}{M} \sum_{i=1}^M l(z_i, f) \right\} \quad (37)$$

while regularization principle is to solve the problem through

$$f^* = \arg \min_{f \in F} \left\{ \frac{1}{M} \sum_{i=1}^M l(z_i, f) \right\} + \lambda \|f\|_p^p$$

where $l(\cdot, f)$ is a loss measure when f is taken as a solution, and $p \geq 0$ is a parameter.

The above learning principles are tightly connected with the neural coding methodology introduced in section 2.3. Actually, for any $f \in F$, if we let $z = (f(x), x)$ be a candidate solution, then the loss $l(z_i, f)$ measures the dissimilarity between z_i and z , so $1/l(z_i, f)$ describes the similarity. Consequently, (37) can be recast as $f^* = \arg \max_f \sum_i S(z_i, z)$ with $S(z_i, z) = 1/l(z_i, f)$.

Based on the neural coding methodology, we thus propose to solve the learning problem by the revised ERM principle

$$f^* = \arg \min_{f \in F} \left\{ \frac{1}{M} \sum_{i=1}^M w(z_i) l(z_i, f) \right\}$$

and the revised regularization principle

$$f^* = \arg \min_{f \in F} \left\{ \frac{1}{M} \sum_{i=1}^M w(z_i) l(z_i, f) \right\} + \lambda \|f\|_p^p \quad (38)$$

where $w(z_i)$ is any fixed neural coding or something like. This then provides a more reasonable learning paradigm. The problems are: *Can we develop a similar statistical learning theory for such neural coding based paradigm? Can we develop a corresponding $L_{1/2}$ or L_1 theory for (38)?*

References

- [1] J.W. Tukey, The future of data analysis, *Ann. Math. Statist.* 33(1962) 1–67.
- [2] D.L. Donoho, High-dimensional data analysis: the curses and blessings of dimensionality, *American Math. Society Lecture—Match Challenges of the 21st Century*, 2000.
- [3] D.J. Bartholomew and M. Knott, *Latent variable methods and factor analysis*, (2nd ed.), London: Arnold, 1999.
- [4] R. Tibshirani, Regression shrinkage and selection via the lasso, *J. Royal. Statist. Soc B.*, 58(1996) 267–288.
- [5] S.S. Chen, D.L. Donoho and M. A. Saundera, Atomic decomposition by basis pursuit, *SIAM Journal of Scientific Computing*, 20(1998) 33–61.
- [6] E. Candés, M. Wakin and S. Boyd, Enhancing sparsity by reweighted L_1 minimization. *J.Fourier A*, 14(2008) 877–905.
- [7] J.M. Santos and J. Marques, Human clustering on bi-dimensional data: an assessment, *Technical Report 1*, INEB-Instituto de Engenharia Biomedica, 2005.
- [8] Z.B. Xu, C.Z. Li and J. Sun, Visual clustering: an approach inspired from visual psychology, Submitted to *IEEE Trans. Pattern Analysis and Machine Intelligence*, 2010.
- [9] A.P. Witkin, Scale space filtering, *Proc. Int’l Joint Conf. Artificial Intelligence*, (1983) 1,019–1,022.
- [10] S. Coren, L.M. Ward, and J.T. Enns, *Sensation and Perception*, Harcourt Brace College Publishers, 1994.
- [11] Y. Leung, J.S. Zhang and Z.B. Xu, Clustering by scale-space filtering, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(2000) 1396–1410.
- [12] S.J. Roberts, Parametric and nonparametric unsupervised clustering analysis, *Pattern Recognition*, 30(1997) 261–272.
- [13] L. Laurence, L. Dury and D. P. Vercauteren, Structural identification of local maxima in low-resolution promolecular electron density distributions, *J.Phys. Chem. A*, 107(2003), 9857–9886.
- [14] S. Grossberg and E. Mingolla, Neural dynamics of form perception: boundary completion, illusory figures, and neon color spreading, *Psychological Review*, 92(1985) 173–211.
- [15] S. E. Palmer, Model theories of gestalt perception, In: Humphreys G W, ed. *Understanding Vision*. CA: Blackwell, 1992.
- [16] S. W. Kuffler, Discharge pattern and functional organization of the mammalian retina, *Journal of Neurophysiology*, 16(1953) 37–68.
- [17] H. B. Barlow, Summation and inhibition in the frog’s retina, *Journal of Neurophysiology*, 19(1953) 69–88.
- [18] D. H. Huber and T. N. Wiesel, Receptive fields of cells in striate cortex of very young, visually inexperienced kittens, *Journal of Neurophysiology*, 26(1971) 994–1002.

-
- [19] R. A. Young, R. M. Lesperance and W. W. Meyer, The Gaussian derivative model for spatial-temporal vision: I. Cortical model, *Spatial Vision*, 14(2001) 261–319.
- [20] D. J. Field and D. J. Tolhurst, The structure and symmetry of simple-cell receptive-field profiles in the cat's visual cortex, *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 228(1986) 379–400.
- [21] S.E. Palmer, Model theories of Gestalt perception, in: Humphreys G.W., ed. *Understanding Vision*, CA:Blackwell, 1992.
- [22] G. Karypis, E.-H.S. Han and V. Kumar, Chameleon: Hierarchical clustering using dynamic modeling, *Computer*, 32(1999) 68–75.
- [23] A. Y. Ng, M. I. Jordan and Y. Weiss, On spectral clustering analysis and algorithm, *Advances in Neural Information Processing Systems*, 14(2001) 849–856.
- [24] J.B. Shi and J. Malik, Normalized cuts and image segmentation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(2000) 888–905.
- [25] M. T. Ester, H. P. Kriegel, J. Scander and X. W. Wu, DBScan: a density-based algorithm for discovering clusters in large spatial databases with noise, *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, (1996) 226–231.
- [26] J. M. Santos, J. Marques and L. A. Alexandre, LEGClust – a clustering algorithm based on layered entropic subgraphs, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(2008) 62–75.
- [27] Z.B. Xu, M.W. Dai and D.Y. Meng. A fast heuristic strategy for model selection of support vector machines, *IEEE Trans. Systems, Man and Cybernetics, Part B*. 39(2009) 1292–1307.
- [28] D.H. Hubel and T.N. Wiesel, Receptive fields and functional architecture of monkey striate cortex. *J. Physiol. London*, (1968) 215–243.
- [29] K. Naka and W. Rushton, S-potentials from luminosity units in the retina of fish, *J. Physiology*, 185(1996) 587–599.
- [30] Olzak & Thomas, L.A. Olzak and J.P. Thomas, Neural recoding in human pattern vision: model and mechanisms, *Vision Research*, 39(1999) 231–256.
- [31] J.S. Zhang and Y.W. Leung, Improved possibilistic C-Means clustering algorithm, *IEEE Trans. Fuzzy System*, 12(2004) 209–217.
- [32] W.E. Vinje and J.L. Gallant, Sparse coding and decorrelation in primary visual cortex during natural vision, *Science*, 287(2000) 1273–1276.
- [33] B.A. Olshausen and D.J. Field, Emergence of simple cell receptive field properties by learning a sparse code for natural images, *Nature*, 381(1996) 607–609.
- [34] Y. Karklin and M.S. Lewicki. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature*, 457(2009) 83–86.
- [35] B.K. Natarajan, Sparse approximate solutions to linear systems, *SIAM. J. Comput.* 24(1995) 227–234.
- [36] H.L. Taylor, S.C. Banks and J.F. McCoy, Deconvolution with the l_1 norm, *Geophysics*, 44(1979) 39–52.
- [37] J.F. Claerbout and F. Muir, Robust modeling with erratic data, *Geophysics*, 38(1973) 826–844.

-
- [38] F. Santosa and W. W. Symes, Linear inversion of band-limited reflection seismograms, *SIAM J. Sci. Stat. Comput.*, 7(1986) 1307–1330.
- [39] D. L. Donoho and P. B. Stark, Uncertainty principles and signal recovery, *SIAM J. Appl. Math.*, 49(1992) 906–931.
- [40] D. L. Donoho and B. F. Logan, Signal recovery and the large sieve, *SIAM J. Appl. Math.*, 52(1992) 577–591.
- [41] L. I. Rudin, S. Osher and E. Fatemi, Nonlinear total variation based noise removal algorithms, *Phys. D*, 60(1992) 259–268.
- [42] P. Blomgren and T. F. Chan, Color TV: total variation methods for restoration of vector-valued images, *IEEE Trans. Image Processing*, 7(1998) 304–309.
- [43] D.L. Donoho and X. Huo, Uncertainty principles and ideal atomic decomposition, *IEEE Trans. Information Theory*, 47(2001) 2845–2862.
- [44] P. Zhao and B. Yu, On model selection consistency of Lasso, *Journal of Machine Learning Research*, 7(2006) 2541–2563.
- [45] N. Meinshausen and B. Yu, Lasso-type recovery of sparse representations for high-dimensional data, *Annals of Statistics*, 720(2009) 246–270.
- [46] Z.B. Xu, H. Zhang, Y. Wang and X.Y. Chang, $L_{1/2}$ regularization, *Science in China Series F-Information Sciences*, 40(2010) 1–11.
- [47] D.L. Donoho. Neighborly polytopes and the sparse solution of underdetermined systems of linear equations. Technical Report, Statistics Department, Stanford University, 2005.
- [48] D.L. Donoho. High-dimensional centrosymmetric polytopes with neighborliness proportional to dimension. *Discrete and Computational Geometry*, 35(2006) 617–652.
- [49] Y. Wang, H.L. Guo, Z.B. Xu and H. Zhang, The representative of $L_{1/2}$ regularization among L_q ($0 < q < 1$) regularizations: an experimental study based on a phase diagram, submitted.
- [50] Z.B. Xu, F.M. Xu, X.Y. Chang and H. Zhang, $L_{1/2}$ regularization: an iterative half thresholding algorithm, submitted.
- [51] Z.B.Xu, J.J.Wang and Z.S.Zhang, Convergence of iterative half thresholding algorithm for $L_{1/2}$ regularization, submitted.
- [52] J.J. Wang, Z.B. Xu, Y. Wang and H. Zhang, Sparse signal recovery based on L_q ($0 < q \leq 1$) regularization, to appear in *Science in China*.
- [53] E.J. Candés, J. Romberg and T. Tao, Signal recovery from incomplete and inaccurate measurements. *Comm. Pure Appl. Math.*, 59(2005) 1207–1223.
- [54] E.J. Candés, The restricted isometry property and its implications for compressed sensing, *Comptas Rendus de l'Académie des Scidence, Serie I*, 346(2008) 589–592.
- [55] S. Foucart and M.J. Lai. Sparsest solutions of underdetermined linear systems via L_q -minimization for $0 < q \leq 1$, *Applied and Computational Harmonic Analysis*, 26(2009) 395–407.
- [56] H.Zhang, Z.B. Xu, X.Y.Chang and Y. Liang, Variable selection and sparse reconstruction via $L_{1/2}$ regularization, submitted.

-
- [57] J. Fan and R. Li, Variable selection via nonconcave penalized likelihood and its oracle properties, *Journal of the American Statistical Association*, 96(2001) 1348–1360.
 - [58] Z.P. Liang and P.C. Lauterbur, *Principles of magnetic resonance imaging: a signal processing perspective*, Wiley Blackwell, 1999.
 - [59] E. J. Candés, J. Romberg and T. Tao. Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inform. Theory*, 52(2004) 489–509.
 - [60] E.J. Candés and T. Tao. Decoding by linear programming, *IEEE Trans. Information Theory*, 51(2005) 4023–4215.
 - [61] E.J. Candés and J. Romberg, L1-maGIC: recovery of sparse signals via convex programming, Technical Report, Caltech, 2005.
 - [62] B. Efron, T. Haistie, I. Johnstone and R. Tibshirani. Least angle regression. *Ann Statist*, 32(2004) 407–499.
 - [63] I. M. Defrise and C.D. Mol, An iterative thresholding algorithm for linear inverse problems with a sparsity constraint, *Communication on Pure and Applied Mathematics*, 11(2004) 1413–1457.
 - [64] T. Blumensath and M.E. Davies, Iterative hard thresholding for compressed sensing, to appear in *Applied and Computational Harmonic Analysis*, 2009.
 - [65] Z.B. Xu and G.F. Roach, Characteristic inequalities of unformaly convex and uniformly smooth Banach spaces, *J.Math.Anal. Appl.* 157(1991) 189–210.

Mathematicalising Behavioural Finance

Xun Yu Zhou*

Abstract

This article presents an overview of the recent development on mathematical treatment of behavioural finance, primarily in the setting of continuous-time portfolio choice under the cumulative prospect theory. Financial motivations and mathematical challenges of the problem are highlighted. It is demonstrated that the solutions to the problem have in turn led to new financial and mathematical problems and machineries.

Mathematics Subject Classification (2010). Primary 91G10; Secondary 91C99.

Keywords. Behavioural finance, cumulative prospect theory, Yaari's criterion, SP/A theory, portfolio selection, continuous time, reference point, S -shaped function, probability distortion, Choquet integral, quantile formulation

1. Introduction

Finance ultimately deals with the interplay between market risks and human judgement. The history of finance theory over the last 60 years has been

*I have been engaged in the mathematical behavioural finance research since 2005. I feel extremely lucky to have worked along the way with some of the highly talented students of mine, including Hanqing Jin, Xuedong He, Zuoquan Xu, Hualei Chang, and Yifei Zhong. Most of the results reported here are based on joint works with them, to whom I am very grateful. These works have been communicated with many colleagues and presented at numerous conferences and seminars around the globe, and I am indebted to my colleagues and the participants at these events for discussions and comments. I gratefully acknowledge financial supports (over the years) from the Hong Kong RGC Earmarked Grants, the Croucher Senior Research Fellowship, a start-up fund of the University of Oxford, the Nomura Centre for Mathematical Finance, and the Oxford–Man Institute of Quantitative Finance. Finally, I thank the Editor of this proceedings for extending twice(!) the deadline of my submission, which has made this article possible. The lesson is behavioural: people always work in the last minute to finish a solicited article regardless how far the deadline is set; hence a) there must be a deadline, and b) there ought to be certain flexibility on the deadline – in order for people to deliver.

Mathematical Institute, University of Oxford, 24-29 St Giles', Oxford, OX1 3LB, UK, and Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, Hong Kong. E-mail: zhoxu@maths.ox.ac.uk.

characterised by two revolutions. The first is neoclassical or maximising finance starting in the 1950s, including mean–variance portfolio selection and expected utility maximisation, the capital asset pricing model (CAPM), efficient market theory, and option pricing. The foundation of neoclassical finance is that the world and its participants are rational “wealth maximisers”; hence finance and economics, albeit primarily about human activities, can be made as logical and predictable as natural sciences. The other revolution is behavioural finance, starting in the 1980s. Its key components are (cumulative) prospect theory, security–potential/aspiration (SP/A) theory, regret and self-control, heuristics and biases. The behavioural theories posit that emotion and psychology do influence our decisions when faced with uncertainties, causing us to behave in unpredictable, inconsistent, incompetent, and most of all, irrational ways. Behavioural finance attempts to explain how and why emotions and cognitive errors influence investors and create stock market anomalies such as bubbles and crashes. It seeks to explore the consistency and predictability in human flaws so that such flaws can be understood, avoided or beneficially exploited.

Mathematical and quantitative approaches have played a pivotal rule in the development of neoclassical finance, and they have led to several groundbreaking, Nobel-prize-winning works. For instance, Markowitz’s mean–variance portfolio selection model (Markowitz 1952), which uses probabilistic terms to quantify the risks as well as quadratic programming to derive the solutions, is widely regarded as the cornerstone of quantitative finance. Black–Scholes–Merton’s option pricing theory (Black and Scholes 1973, Merton 1973), which employs the Itô calculus and partial differential equations as the underlying mathematical tools, is a fine example of “mathematising finance”. On the other hand, while Daniel Kahneman won a Nobel prize in 2002 for his work on the prospect theory, behavioural finance is still a relatively new field in which research has so far been largely limited to be descriptive, experimental, and empirical. Rigorous mathematical treatment of behavioural finance, especially that for the continuous-time setting, is very rare in the literature. An important reason for this is that behavioural problems bring in highly unconventional and challenging features for which the known mathematical techniques and machineries almost all fall apart. Therefore, new mathematical theories and approaches, instead of mere extensions of the existing ones, are called for in formulating and solving behavioural models.

This article is to give an account of the recent development on mathematical behavioural finance theory, primarily in the realm of continuous-time behavioural portfolio selection. Study on continuous-time portfolio choice has so far predominantly centred around expected utility maximisation since the seminal papers of Merton (1969, 1971). Expected utility theory (EUT), developed by von Neumann and Morgenstern (1944) based on an axiomatic system, is premised upon the assumptions that decision makers are rational and risk averse when facing uncertainty. In the context of financial portfolio choice, its basic tenets are: investors evaluate wealth according to final asset positions; they are

uniformly risk averse; and they are able to evaluate probabilities objectively. These, however, have long been challenged by many observed and repeatable empirical patterns as well as a number of famous paradoxes and puzzles such as Allais paradox (Allais 1953), Ellsberg paradox (Ellsberg 1961), Friedman and Savage puzzle (Friedman and Savage 1948), and the equity premium puzzle (Mehra and Prescott 1985).

Hence, many alternative preference measures to expected utility have been proposed, notably Yaari's *dual theory of choice* (Yaari 1987) which attempts to resolve a number of puzzles and paradoxes associated with the expected utility theory. To illustrate Yaari's theory, consider first the following expected utility

$$Eu(X) = \int_{-\infty}^{+\infty} u(x)dF_X(x) \tag{1}$$

where X is a random payoff with $F_X(\cdot)$ as its cumulative distribution function (CDF), and $u(\cdot)$ is a utility function. This expression shows that $u(\cdot)$ can be regarded as a *nonlinear* "distortion" on payment when evaluating the mean of X (if $u(x) \equiv x$, then the expression reduces to the mean). Yaari (1987) introduces the following criterion

$$V(X) = \int_{-\infty}^{+\infty} w(P(X > x))dx, \tag{2}$$

where $w(\cdot)$, called the *probability distortion* (or *weighting*) *function*, maps from $[0, 1]$ onto $[0, 1]$, with $w(0) = 0, w(1) = 1$. Mathematically, (2) involves the so-called *Choquet integral* with respect to the *capacity* $w \circ P$ (see Denneberg 1994 for a comprehensive account on Choquet integrals). This criterion can be rewritten, assuming $w(\cdot)$ is suitably differentiable, as

$$V(X) = \int_{-\infty}^{+\infty} xd[-w(1 - F_X(x))] = \int_{-\infty}^{+\infty} xw'(1 - F_X(x))dF_X(x). \tag{3}$$

The first identity in (3) suggests that the criterion involves a distortion on the CDF, in contrast to (1). The second identity reveals the role $w(\cdot)$ plays in this new risk preference measure. The term $w'(1 - F_X(x))$ puts a weight on the payment x . If $w(\cdot)$ is convex, the value of $w'(p)$ is greater around $p = 1$ than around $p = 0$; so $V(X)$ overweights payoffs close to the low end and underweights those close to the high end. In other words, the agent is risk averse. By the same token, the agent is risk seeking when $w(\cdot)$ is concave. Thus, in Yaari's theory risk attitude is captured by the nonlinear distortion of decumulative distribution rather than the utility of payoff.

Probability distortion has been observed in many experiments. Here we present two (rather simplified) examples. We write a random variable (*prospect*) $X = (x_i, p_i; i = 1, 2, \dots, m)$ if $X = x_i$ with probability p_i . We write $X \succ Y$ if prospect X is preferred than prospect Y . Then it has been observed that $(\pounds 5000, 0.1; \pounds 0, 0.9) \succ (\pounds 5, 1)$ although the two prospects have the same mean.

One of the explanations is that people usually exaggerate the small probability associated with a big payoff (so people buy lotteries). On the other hand, it is usual that $(-\pounds 5, 1) \succ (-\pounds 5000, 0.1; \pounds 0, 0.9)$, indicating an inflation of the small probability in respect of a big loss (so people buy insurances).

Other theories developed along this line of involving probability distortions include Lopes' SP/A model (Lopes 1987) and, most significantly, Kahneman and Tversky's (cumulative) prospect theory (Kahneman and Tversky 1979, Tversky and Kahneman 1992), both in the paradigm of modern behavioural decision-making. Cumulative prospect theory (CPT) uses cognitive psychological techniques to incorporate anomalies in human behaviour into economic decision-making. In the context of financial asset allocation, the key elements of CPT are:

- People evaluate assets on *gains* and *losses* (which are defined with respect to a *reference point*), instead of on final wealth positions;
- People behave differently on gains and on losses; they are not uniformly risk averse, and are distinctively more sensitive to losses than to gains (the latter is a behaviour called *loss aversion*);
- People overweight small probabilities and underweight large probabilities.

The significance of the reference point and the presence of non-uniform risk preferences can be demonstrated by the following two experiments.

Experiment 1 You have been given $\pounds 1000$. Now choose between 1A) Win $\pounds 1000$ with 50% chance and $\pounds 0$ with 50% chance, and 1B) Win $\pounds 500$ with 100% chance.

Experiment 2 You have been given $\pounds 2000$. Now choose between 2A) Lose $\pounds 1000$ with 50% chance, and $\pounds 0$ with 50% chance, and 2B) Lose $\pounds 500$ with 100% chance.

It turns out that 1B) and 2A) were more popular in Experiments 1 and 2 respectively¹. However, if one takes the initial amounts ($\pounds 1000$ and $\pounds 2000$ respectively) into consideration then it is easy to see that 1A) and 2A) are exactly the same *as random variables*, and so are 1B) and 2B). The different choices of reference points ($\pounds 1000$ and $\pounds 2000$ in these experiments) have led to completely opposite decisions. On the other hand, the choice of 2A) in Experiment 2 indicates that in a *loss* situation, people favours risky prospects (namely they become risk-seeking), in sharp contrast to a *gain* situation in Experiment 1.

The loss aversion can be defined as $(x, 0.5; -x, 0.5) \succ (y, 0.5; -y, 0.5)$ when $y > x > 0$ are gains with respect to some reference point. So the marginal utility of gaining an additional $\pounds 1$ is lower than the marginal disutility of losing an additional $\pounds 1$.

¹The outcomes of these experiments – or their variants – are well documented in the literature. I have myself conducted them in a good number of conference and seminar presentations, and the results have been very consistent.

The aforementioned CPT elements translate respectively into the following technical features when formulating a CPT portfolio choice model:

- A reference point in wealth that defines gains and losses;
- A value function or utility function, *concave for gains* and *convex for losses* (such a function is called *S-shaped*), and steeper for losses than for gains;
- A probability distortion (or weighting) that is a *nonlinear* transformation of the probability measure, which inflates a small probability and deflates a large probability.

There have been burgeoning research interests in incorporating behavioural theories into portfolio choice; nonetheless these have been hitherto overwhelmingly limited to the single-period setting; see for example Benartzi and Thaler (1995), Lopes and Oden (1999), Shefrin and Statman (2000), Bassett *et al.* (2004), Gomes (2005), and De Giorgi and Post (2008). Most of these works focus on empirical and numerical studies, and some of them solve the underlying optimisation problems simply by heuristics. Recently, Bernard and Ghossoub (2009) and He and Zhou (2009) have carried out analytical treatments on single-period CPT portfolio choice models and obtained closed-form solutions for a number of interesting cases.

There has been, however, little analytical treatment on *dynamic*, especially continuous-time, asset allocation featuring behavioural criteria. Such a lack of study on continuous-time behavioural portfolio choice is certainly not because the problem is uninteresting or unimportant; rather, we believe, it is because all the main mathematical approaches dealing with the conventional expected utility maximisation model fail completely. To elaborate, despite the existence of thousands of papers on the expected utility model, there are essentially only two approaches involved. One is the stochastic control or dynamic programming approach, initially proposed by Merton (1969), which transforms the problem into solving a partial differential equation, the Hamilton-Jacobi-Bellman (HJB) equation. The other one is the martingale approach. This approach, developed by Harrison and Kreps (1979) and Harrison and Pliska (1981), employs a martingale characterisation to turn the dynamic wealth equation into a static budget constraint and then identifies the optimal terminal wealth via solving a static optimisation problem. If the market is complete, an optimal strategy is then derived by replicating the established optimal terminal wealth, in the spirit of perfectly hedging a contingent claim. In an incomplete market with possible portfolio constraints, the martingale approach is further developed to include the so-called convex duality machinery; see, e.g., Cvitanić and Karatzas (1992).

Now, nonlinear probability distortions in behavioural finance abolish virtually all the nice properties associated with the standard additive probability and linear expectation. In particular, the time-consistency of the conditional expectation with respect to a filtration, which is the core of the dynamic programming principle, is absent due to the distorted probability. Moreover, in the

CPT framework, the utility function is non-convex and non-concave, while the global convexity/concavity is a necessity in traditional optimization. Worse still, the coupling of these two ill-behaved features greatly amplifies the difficulty of the problem.

Berkelaar, Kouwenberg and Post (2004) study a CPT mode with a specific two-piece power utility function. They employ a convexification technique to tackle the non-convexity of the problem. However, the probability distortion, which is one of the major ingredients of all the behavioural theories and which causes a main technical difficulty, is absent in that paper. Jin and Zhou (2008) develop a new theory in solving systematically continuous-time CPT models, featuring both S -shaped utility functions and probability distortions. The whole machinery is very involved; however its essential ideas are clear and intuitive. It constitutes several steps. First, to handle the S -shaped utility function one decomposes the problem, by parameterising some key variables, into a gain part problem and a loss part problem. The gain part problem is a Choquet maximisation problem involving a concave utility function and a probability distortion. The difficulty arising from the distortion is overcome by a so-called *quantile formulation* which changes the decision variable from the random variable X to its quantile function $G(\cdot)$. This quantile formulation has been used by several authors, such as Schied (2004, 2005), Dana (2005), Carlier and Dana (2005), in ad hoc ways to deal with problems with convex/concave probability distortions. It has recently been further developed by He and Zhou (2010) into a general paradigm of solving non-expected utility maximisation models. The loss part problem, on the other hand, is more subtle and difficult to handle even with the quantile formulation, because it is to *minimise* a *concave* functional (thanks, of course, to the original S -shaped utility function). Hence it is essentially a combinatorial optimisation in an infinite dimension. The problem is solved by noting that such a problem must have corner-point solutions, which are step functions in a function space. Once the gain and loss part problems are solved, their solutions are then appropriately pasted by optimising the parameters introduced in the first step.

The rest of this article is organised as follows. Section 2 presents the continuous-time CPT portfolio selection model and the approach to solve the model. Motivated by the gain part problem of the CPT model, Section 3 discusses about the quantile formulation that is a powerful tool in dealing with many non-expected utility models. Section 4 is concerned with the loss part problem and its solution procedure. Finally, Section 5 concludes.

2. The CPT Model

2.1. Model formulation. Consider a CPT agent with an investment planning horizon $[0, T]$ and an initial endowment $x_0 > 0$, both exogenously fixed

throughout this paper, in an arbitrage-free economy². Let $(\Omega, \mathcal{F}, P, \{\mathcal{F}_t\}_{t \geq 0})$ be a standard filtered complete probability space representing the underlying uncertainty, and ρ be the *pricing kernel* (also known as the *stochastic discount factor* in the economics literature), which is an \mathcal{F}_T -measurable random variable, so that any \mathcal{F}_T -measurable and lower bounded contingent claim ξ has a unique price $E[\rho\xi]$ at $t = 0$ (provided that $E[\rho\xi] < +\infty$). The technical requirements on ρ throughout are that $0 < \rho < +\infty$ a.s., $0 < E\rho < +\infty$, and ρ admits no atom, i.e. $P(\rho = x) = 0$ for any $x \in \mathbb{R}^+$.

The key underlying assumption in such an economy is that “the price is linear”. The general existence of a pricing kernel ρ can be derived, say, by Riesz’s representation theorem under the price linearity in an appropriate Hilbert space. Hence, our setting is indeed very general. It certainly covers the continuous-time complete market considered in Jin and Zhou (2008) with general Itô processes for asset prices, in which case ρ is the usual pricing kernel having an explicit form involving the market price of risk. It also applies to a continuous-time *incomplete* market with a deterministic investment opportunity set, where ρ is the minimal pricing kernel; see, e.g., Föllmer and Kramkov (1997).

The agent risk preference is dictated by CPT. Specifically, she has a reference point B at the terminal time T , which is an \mathcal{F}_T -measurable random variable with $E[\rho B] < +\infty$. The reference point B determines whether a given terminal wealth position is a gain (excess over B) or a loss (shortfall from B). It could be interpreted as a liability the agent has to fulfil (e.g. a house downpayment), or an aspiration she strives to achieve (e.g. a target profit aspired by, or imposed on, a fund manager). The agent utility (value) function is S -shaped: $u(x) = u_+(x^+) \mathbf{1}_{x \geq 0}(x) - u_-(x^-) \mathbf{1}_{x < 0}(x)$, where the superscripts \pm denote the positive and negative parts of a real number, u_+, u_- are *concave* functions on \mathbb{R}^+ with $u_{\pm}(0) = 0$, reflecting risk-aversion on gains and risk-seeking on losses. There are also probability distortions on both gains and losses, which are captured by two nonlinear functions w_+, w_- from $[0, 1]$ onto $[0, 1]$, with $w_{\pm}(0) = 0, w_{\pm}(1) = 1$ and $w_{\pm}(p) > p$ (respectively $w_{\pm}(p) < p$) when p is close to 0 (respectively 1).

The agent preference on a terminal wealth X (which is an \mathcal{F}_T -random variable) is measured by the prospective functional

$$V(X - B) := V_+((X - B)^+) - V_-((X - B)^-),$$

where $V_+(Y) := \int_0^{+\infty} w_+(P(u_+(Y) \geq y))dy$, $V_-(Y) := \int_0^{+\infty} w_-(P(u_-(Y) \geq$

²In our model the agent is a “small investor”; so her preference only affects her own utility function – and hence *her* portfolio choice – but not the overall economy. Therefore issues like “the limit of arbitrage” and “market equilibrium” are beyond the scope of this article.

$y))dy$. Thus, the CPT portfolio choice problem is to

$$\begin{aligned} & \underset{X}{\text{Maximise}} && V(X - B) \\ & \text{subject to} && \begin{cases} E[\rho X] = x_0 \\ X \text{ is } \mathcal{F}_T \text{-measurable and lower bounded.} \end{cases} \end{aligned} \quad (4)$$

Here the lower boundedness corresponds to the requirement that the admissible portfolios be “tame”, i.e., each of the admissible portfolios generates a lower bounded wealth process, which is standard in the continuous-time portfolio choice literature (see, e.g., Karatzas and Shreve 1998 for a discussion).

We introduce some notation related to the pricing kernel ρ . Let $F_\rho(\cdot)$ be the cumulative distribution function (CDF) of ρ , and $\bar{\rho}$ and $\underline{\rho}$ be respectively the essential lower and upper bounds of ρ , namely,

$$\begin{aligned} \bar{\rho} &\equiv \text{esssup } \rho := \sup \{a \in \mathbb{R} : P\{\rho > a\} > 0\}, \\ \underline{\rho} &\equiv \text{essinf } \rho := \inf \{a \in \mathbb{R} : P\{\rho < a\} > 0\}. \end{aligned} \quad (5)$$

The following assumption is introduced in Jin and Zhou (2008) in solving (4).

Assumption 1. $u_+(\cdot)$ is strictly increasing, strictly concave and twice differentiable, with the Inada conditions $u'_+(0+) = +\infty$ and $u'_+(+\infty) = 0$, and $u_-(\cdot)$ is strictly increasing, and strictly concave at 0. Both $w_+(\cdot)$ and $w_-(\cdot)$ are non-decreasing and differentiable. Moreover, $F_\rho^{-1}(z)/w'_+(z)$ is non-decreasing in $z \in (0, 1]$, $\liminf_{x \rightarrow +\infty} \left(\frac{-xu''_+(x)}{u'_+(x)} \right) > 0$, and $E \left[u_+ \left((u'_+)^{-1} \left(\frac{\rho}{w'_+(F_\rho(\rho))} \right) \right) w'_+(F_\rho(\rho)) \right] < +\infty$.

By and large, the monotonicity of the function $F_\rho^{-1}(z)/w'_+(z)$ can be interpreted economically as a requirement that the probability distortion $w_+(\cdot)$ on gains should not be too large in relation to the market (or, loosely speaking, the agent should not be over-optimistic about huge gains); see Jin and Zhou (2008), Section 6.2, for a detailed discussion. Other conditions in Assumption 1 are mild and economically motivated.

2.2. Ill-Posedness. In general we say a maximisation problem is *well-posed* if its supremum is finite; otherwise it is *ill-posed*. Well-posedness is more a modelling issue; an ill-posed model sets incentives in such a way that the decision-maker could achieve an infinitely favourable value without having to consider trade-offs.

In classical portfolio selection literature (see, e.g., Karatzas and Shreve 1998) the utility function is typically assumed to be globally concave along with other nice properties; thus the problem is guaranteed to be well-posed in most cases³.

³Even with a global concave utility function the underlying problem could still be ill-posed; see counter-examples and discussions in Jin, Xu and Zhou (2008).

However, for the CPT model (4) the well-posedness becomes a more significant issue, and that probability distortions in gains and losses play prominent, yet somewhat opposite, roles.

Theorem 1. (Jin and Zhou 2008, Theorems 3.1 and 3.2) *Problem (4) is ill-posed under either of the following two conditions:*

- (i) *There exists a nonnegative \mathcal{F}_T -measurable random variable X such that $E[\rho X] < +\infty$ and $V_+(X) = +\infty$.*
- (ii) *$u_+(+\infty) = +\infty$, $\bar{\rho} = +\infty$, and $w_-(x) = x$.*

Theorem 1-(i) says that the model is ill-posed if one can find a nonnegative claim having a finite price yet an infinite prospective value. In this case the agent can purchase such a claim initially (by taking out a loan if necessary) and reach the infinite value at the end. Here we reproduce Example 3.1 in Jin and Zhou (2008) for the existence of such a claim in a simple case with very “nice” parameter specifications. Let ρ be such that $F_\rho(\cdot)$ is continuous and strictly increasing, with $E\rho^3 < +\infty$ (e.g., when ρ is lognormal). Take $w_+(p) := p^{1/4}$ on $p \in [0, 1/2]$ and $u_+(x) := x^{1/2}$. Set $Z := F_\rho(\rho) \sim U(0, 1)$ and define $X := Z^{-1/2} - 1$. Then it is an easy exercise to show that $E[\rho X] < +\infty$ and $V_+(X) = +\infty$. Notice that the culprit of the ill-posedness in this case is the probability distortion $w_+(\cdot)$ which has very large curvatures around 0. In other words, the agent is excessively optimistic in the sense that she over-exaggerates the tiny probability of a huge gain, so much so that her resulting risk-seeking behaviour overrides the risk-averse part of the utility function in the gain domain. This in turn leads to a problem without trade offs (an ill-posed one, that is). So the agent is misled by her own “psychological illusion” (her preference set) to take the most risky exposures.

Theorem 1-(ii) shows that a probability distortion on *losses* is *necessary* for the well-posedness if the market upside potential is unlimited (as implied by $\bar{\rho} = +\infty$). In this case, the agent would borrow an enormous amount of money to purchase a claim with a huge payoff, and then bet the market be “good” leading to the realization of that payoff. If, for the lack of luck, the market turns out to be “bad”, then the agent ends up with a loss; however due to the non-distortion on losses its damage to the prospective value is bounded⁴. In plain words, if the agent has no fear in the sense that she does not exaggerate the small probabilities of huge losses, and the market has an unlimited potential of going up, then she would be lured by her CPT criterion to take the infinite risky exposure (again an ill-posed model).

To exclude the ill-posed case identified by Theorem 1-(i), we introduce the following assumption.

⁴This argument is no longer valid if the wealth is *constrained* to be bounded from below. This is why in Berkelaar *et al.* (2004) the model is well-posed even though no probability distortion is considered, as the wealth process there is constrained to be non-negative.

Assumption 2. $V_+(X) < +\infty$ for any nonnegative, \mathcal{F}_T -measurable random variable X satisfying $E[\rho X] < +\infty$.

2.3. Solutions. The original problem (4) is solved in two steps involving three sub-problems, which are described in what follows.

Step 1. In this step we consider two problems respectively:

- *The Gain Part Problem:* A problem with parameters (A, x_+) :

$$\begin{aligned} & \underset{X}{\text{Maximise}} && V_+(X) = \int_0^{+\infty} w_+(P(u_+(X) > y))dy \\ & \text{subject to} && E[\rho X] = x_+, \quad X \geq 0 \text{ a.s.}, \quad X = 0 \text{ a.s. on } A^C, \end{aligned} \tag{6}$$

where $x_+ \geq (x_0 - E[\rho B])^+ (\geq 0)$ and $A \in \mathcal{F}_T$ are given. Thanks to Assumption 2, $V_+(X)$ is a finite number for any feasible X . We define the optimal value of Problem (6), denoted $v_+(A, x_+)$, in the following way. If $P(A) > 0$, in which case the feasible region of (6) is non-empty [$X = (x_+ \mathbf{1}_A) / (\rho P(A))$ is a feasible solution], then $v_+(A, x_+)$ is defined to be the supremum of (6). If $P(A) = 0$ and $x_+ = 0$, then (6) has only one feasible solution $X = 0$ a.s. and $v_+(A, x_+) := 0$. If $P(A) = 0$ and $x_+ > 0$, then (6) has no feasible solution, where we define $v_+(A, x) := -\infty$.

- *The Loss Part Problem:* A problem with parameters (A, x_+) :

$$\begin{aligned} & \underset{X}{\text{Minimise}} && V_-(X) = \int_0^{+\infty} w_-(P(u_-(X) > y))dy \\ & \text{subject to} && \begin{cases} E[\rho X] = x_+ - x_0 + E[\rho B], \quad X \geq 0 \text{ a.s.}, \quad X = 0 \text{ a.s. on } A, \\ X \text{ is upper bounded a.s.}, \end{cases} \end{aligned} \tag{7}$$

where $x_+ \geq (x_0 - E[\rho B])^+$ and $A \in \mathcal{F}_T$ are given. Similarly to the gain part problem we define the optimal value $v_-(A, x_+)$ of Problem (7) as follows. When $P(A) < 1$ in which case the feasible region of (7) is non-empty, $v_-(A, x_+)$ is the infimum of (7). If $P(A) = 1$ and $x_+ = x_0 - E[\rho B]$ where the only feasible solution is $X = 0$ a.s., then $v_-(A, x_+) := 0$. If $P(A) = 1$ and $x_+ \neq x_0 - E[\rho B]$, then there is no feasible solution, in which case we define $v_-(A, x_+) := +\infty$.

Step 2. In this step we solve

$$\begin{aligned} & \underset{(A, x_+)}{\text{Maximise}} && v_+(A, x_+) - v_-(A, x_+) \\ & \text{subject to} && \begin{cases} A \in \mathcal{F}_T, \quad x_+ \geq (x_0 - E[\rho B])^+, \\ x_+ = 0 \text{ when } P(A) = 0, \quad x_+ = x_0 - E[\rho B] \text{ when } P(A) = 1. \end{cases} \end{aligned} \tag{8}$$

The interpretations of the gain and loss part problems, as well as the parameters (A, x_+) , are intuitive. If X is any feasible solution to (4), then its deviation from the reference point B can be decomposed by $X - B = (X - B)^+ - (X - B)^-$.

Let $A := \{X \geq B\}$, the event of ending up with gains, and $x_+ := E[\rho(X - B)^+]$, the price of the gains, then $(X - B)^+$ and $(X - B)^-$ are respectively feasible solutions to (6) and (7) with the parameters (A, x_+) .

If Step 1 above is to “divide” – to decompose the original problem into two sub-problems, then Step 2 is to “conquer” – to combine the solutions of the sub-problems in the best way so as to solve the original one. Problem (8) is to find the “best” split between good states and bad states of the world, as well as the corresponding price of the gains. Mathematically, it is an optimisation problem with the decision variables being a real number, x_+ , and a random event, A , the latter being very hard to handle. However, Jin and Zhou (2008), Theorem 5.1, shows that one needs only to consider the type of events $A = \{\rho \leq c\}$, where c is a real number in certain range, when optimising (8). This important result in turn suggests that the event of having gains is completely characterised by the pricing kernel and a critical threshold.

With all these preliminaries at hand, we can now state the solution to (4) in terms of the following two-dimensional mathematical programme with the decision variables (c, x_+) , which is intimately related to (but not the same as) Problem (8):

$$\begin{aligned} \text{Maximise}_{(c, x_+)} \quad & v(c, x_+) = E \left[u_+ \left((u'_+)^{-1} \left(\frac{\lambda(c, x_+) \rho}{w'_+(F_\rho(\rho))} \right) \right) w'_+(F_\rho(\rho)) \mathbf{1}_{\rho \leq c} \right] \\ & - u_- \left(\frac{x_+ - (x_0 - E[\rho B])}{E[\rho \mathbf{1}_{\rho > c}]} \right) w_-(1 - F_\rho(c)) \end{aligned} \tag{9}$$

$$\text{subject to} \quad \begin{cases} \rho \leq c \leq \bar{\rho}, & x_+ \geq (x_0 - E[\rho B])^+, \\ x_+ = 0 & \text{when } c = \underline{\rho}, \quad x_+ = x_0 - E[\rho B] \text{ when } c = \bar{\rho}, \end{cases}$$

where $\lambda(c, x_+)$ satisfies $E[(u'_+)^{-1}(\frac{\lambda(c, x_+) \rho}{w'_+(F_\rho(\rho))}) \rho \mathbf{1}_{\rho \leq c}] = x_+$, and we use the following convention:

$$u_- \left(\frac{x_+ - (x_0 - E[\rho B])}{E[\rho \mathbf{1}_{\rho > c}]} \right) w_-(1 - F_\rho(c)) := 0 \quad \text{when } c = \bar{\rho} \text{ and } x_+ = x_0 - E[\rho B]. \tag{10}$$

Theorem 2. (Jin and Zhou 2008, Theorem 4.1) *We have the following conclusions:*

- (i) *If X^* is optimal for Problem (4), then $c^* := F_\rho^{-1}(P\{X^* \geq B\})$, $x_+^* := E[\rho(X^* - B)^+]$, are optimal for Problem (9).*
- (ii) *If (c^*, x_+^*) is optimal for Problem (9), then $\{X^* \geq B\}$ and $\{\rho \leq c^*\}$ are identical up to a zero probability event. In this case*

$$X^* = \left[(u'_+)^{-1} \left(\frac{\lambda \rho}{w'_+(F_\rho(\rho))} \right) + B \right] \mathbf{1}_{\rho \leq c^*} - \left[\frac{x_+^* - (x_0 - E[\rho B])}{E[\rho \mathbf{1}_{\rho > c^*}]} - B \right] \mathbf{1}_{\rho > c^*}$$

is optimal for Problem (4).

The explicit form of the optimal terminal wealth profile, X^* , is sufficiently informative to reveal the key qualitative and quantitative features of the corresponding optimal portfolio⁵. The following summarise the economical interpretations and implications of Theorem 2, including those of c^* and x_+^* :

- The future world at $t = T$ is divided by two classes of states: “good” ones (having gains) or “bad” ones (having losses). Whether the agent ends up with a good state is *completely* determined by $\rho \leq c^*$, which in statistical terms is a simple hypothesis test involving a constant c^* , à la Neyman–Pearson’s lemma (see, e.g., Lehmann 1986).
- Optimal strategy is a *gambling* policy, betting on the good states while accepting a loss on the bad. Specifically, at $t = 0$ the agent needs to sell the “loss” lottery, $\left[\frac{x_+^* - (x_0 - E[\rho B])}{E[\rho \mathbf{1}_{\rho > c^*}]} - B\right] \mathbf{1}_{\rho > c^*}$, in order to raise fund to purchase the “gain” lottery, $\left[(u'_+)^{-1}\left(\frac{\lambda \rho}{w'_+(F_\rho(\rho))}\right) + B\right] \mathbf{1}_{\rho \leq c^*}$.
- The probability of finally reaching a good state is $P(\rho \leq c^*) \equiv F_\rho(c^*)$, which in general depends on the reference point B , since c^* depends on B via (9). Equivalently, c^* is the quantile of the pricing kernel evaluated at the probability of good states.
- x_+^* is the price of the terminal gains.
- The magnitude of potential losses in the case of a bad state is a *constant* $\frac{x_+^* - (x_0 - E[\rho B])}{E[\rho \mathbf{1}_{\rho > c^*}]} \geq 0$, which is endogenously *dependent* of B .
- $x_+^* + E[\rho B \mathbf{1}_{\rho \leq c^*}] \equiv E[\rho X^* \mathbf{1}_{\rho > c^*}]$ is the $t = 0$ price of the gain lottery. Hence, if B is set too high such that $x_0 < x_+^* + E[\rho B \mathbf{1}_{\rho \leq c^*}]$, i.e., the initial wealth is not sufficient to purchase the gain lottery⁶, then the optimal strategy *must* involve a leverage.
- If $x_0 < E[\rho B]$, then the optimal $c^* < \bar{\rho}$ (otherwise by the constraints of (9) it must hold that $x_+^* = x_0 - E[\rho B] < 0$ contradicting the non-negativeness of x_+^*); hence $P(\rho > c^*) > 0$. This shows that if the reference point is set too high compared with the initial endowment, then the odds are not zero that the agent ends up with a bad state.

⁵The specific optimal trading *strategy* depends on the underlying economy, in particular the form of the asset prices. For instance, for a complete continuous-time market, the optimal strategy is the one that replicates X^* in a Black–Schole way. If the market is incomplete but with a deterministic investment opportunity set, then ρ involved is the minimal pricing kernel, and X^* in Theorem 2-(ii) is automatically a monotone functional of ρ and hence replicable. However, we do not actually need the form of the optimal strategy in our subsequent discussions.

⁶It is shown in Jin and Zhou (2009), Theorems 4 and 7, that $P(\rho \leq c^*)$ converges to a constant when B goes to infinity, in the case when the utility function is two-piece CRRA and the pricing kernel is lognormal. So $x_+^* + E[\rho B \mathbf{1}_{\rho \leq c^*}]$ will be sufficiently large when B is sufficiently large.

2.4. An example: Two-piece CRRA utility. We now illustrate the general results of Theorem 2 by a benchmark case where ρ is lognormal, i.e., $\log \rho \sim N(\mu, \sigma^2)$ with $\sigma > 0$, and the utility function is two-piece CRRA, i.e.,

$$u_+(x) = x^\alpha, \quad u_-(x) = kx^\alpha, \quad x \geq 0$$

where $k > 0$ (the loss aversion coefficient) and $0 < \alpha < 1$ are constants. In this case $\bar{\rho} = +\infty$ and $\underline{\rho} = 0$. This setting is general enough to cover, for example, a continuous-time economy with Itô processes for multiple asset prices (Karatzas and Shreve 1998, Jin and Zhou 2008) and Kahneman–Tversky’s utility functions (Tversky and Kahneman 1992) with $\alpha = 0.88$.

In this case, the crucial mathematical programme (9) has the following more specific form (see Jin and Zhou 2008, eq. (9.3)):

$$\begin{aligned} & \underset{(c, x_+)}{\text{Maximise}} && v(c, x_+) = \varphi(c)^{1-\alpha} x_+^\alpha - \frac{k w_-(1 - F_\rho(c))}{(E[\rho \mathbf{1}_{\rho > c}])^\beta} (x_+ - \tilde{x}_0)^\alpha, \\ & \text{subject to} && \begin{cases} 0 \leq c \leq +\infty, & x_+ \geq \tilde{x}_0^+, \\ x_+ = 0 \text{ when } c = 0, & x_+ = \tilde{x}_0 \text{ when } c = +\infty, \end{cases} \end{aligned} \tag{11}$$

where $\tilde{x}_0 := x_0 - E[\rho B]$ and

$$\varphi(c) := E \left[\left(\frac{w'_+(F_\rho(\rho))}{\rho} \right)^{1/(1-\alpha)} \rho \mathbf{1}_{\rho \leq c} \right] \mathbf{1}_{c > 0}, \quad 0 \leq c \leq +\infty.$$

It turns out that (11) can be solved explicitly.

Introduce the following function:

$$k(c) := \frac{k w_-(1 - F_\rho(c))}{\varphi(c)^{1-\alpha} (E[\rho \mathbf{1}_{\rho > c}])^\alpha} > 0, \quad c > 0.$$

We state the results for two different cases: one when the agent is initially in the gain domain and the other in the loss domain.

Theorem 3. (Jin and Zhou 2008, Theorem 9.1) Assume that $x_0 \geq E[\rho B]$.

(i) If $\inf_{c > 0} k(c) \geq 1$, then the optimal solution to (4) is

$$X^* = \frac{x_0 - E[\rho B]}{\varphi(+\infty)} \left(\frac{w'_+(F_\rho(\rho))}{\rho} \right)^{1/(1-\alpha)} + B.$$

(ii) If $\inf_{c > 0} k(c) < 1$, then (4) is ill-posed.

Theorem 4. (Jin and Zhou 2008, Theorem 9.2) Assume that $x_0 < E[\rho B]$.

(i) If $\inf_{c > 0} k(c) > 1$, then (4) is well-posed. Moreover, (4) admits an optimal solution if and only if the following optimisation problem attains an optimal solution

$$\text{Min}_{0 \leq c < +\infty} \left[\left(\frac{k w_-(1 - F_\rho(c))}{(E[\rho \mathbf{1}_{\rho > c}])^\alpha} \right)^{1/(1-\alpha)} - \varphi(c) \right]. \tag{12}$$

Furthermore, if an optimal solution c^* of (12) satisfies $c^* > 0$, then the optimal solution to (4) is

$$X^* = \frac{x_+^*}{\varphi(c^*)} \left(\frac{w'_+(F_\rho(\rho))}{\rho} \right)^{1/(1-\alpha)} \mathbf{1}_{\rho \leq c^*} - \frac{x_+^* - (x_0 - E[\rho B])}{E[\rho \mathbf{1}_{\rho > c^*}]} \mathbf{1}_{\rho > c^*} + B, \quad (13)$$

where $x_+^* := \frac{-(x_0 - E[\rho B])}{k(c^*)^{1/(1-\alpha)} - 1}$. If $c^* = 0$ is the only minimiser in (12), then the unique optimal solution to (4) is $X^* = \frac{x_0 - E[\rho B]}{E\rho} + B$.

- (ii) If $\inf_{c>0} k(c) = 1$, then the supremum value of (4) is 0, which is however not achievable.
- (iii) If $\inf_{c>0} k(c) < 1$, then (4) is ill-posed.

As seen from the preceding theorems the characterising condition for well-posedness in both cases is $\inf_{c>0} k(c) \geq 1$, which is equivalent to

$$k \geq \left(\inf_{c>0} \frac{w_-(1 - F_\rho(c))}{\varphi(c)^{1-\alpha} (E[\rho \mathbf{1}_{\rho > c}])^\alpha} \right)^{-1} := k_0.$$

Recall that k is the loss aversion level of the agent ($k = 2.25$ in Tversky and Kahneman 1992). Thus the agent must be *sufficiently* loss averse in order to have a well-posed portfolio choice model.

Another interesting observation is that the optimal portfolios behave fundamentally different depending on whether the agent starts with a gain or loss situation (determined by the initial wealth in relation to the discounted reference point). If she starts in a gain territory, then the optimal strategy is simply to spend $x_0 - E[\rho B]$ buying a contingent claim that delivers a payoff in excess of X , reminiscent of a classical utility maximizing agent (although the allocation to stocks is “distorted” due to the probability distortion). If the initial situation is a loss, then the agent needs to get “out of the hole” soonest possible. As a result, the optimal strategy is a gambling policy which involves raising additional capital to purchase a claim that delivers a higher payoff in the case of a good state of the market and incurs a fixed loss in the case of a bad one. Finally, if $x_0 = E[\rho B]$, then the agent simply buy the claim B at price x_0 . If in particular B is the risk-free payoff, then the optimal portfolio is *not* to invest in risky asset at all. Notice that this case underlines a natural psychological reference point – the risk-free return – for many people. This, nonetheless, does explain why most households do not invest in equities at all⁷.

As described by Theorem 4-(i), the solution of (4) relies on some attainability condition of a minimisation problem (12), which is rather technical (or, shall we say, mathematical) without clear economical interpretation. The following

⁷A similar result is derived in Gomes (2005) for his portfolio selection model with loss averse investors, albeit in the single-period setting without probability distortions.

Theorem 5, however, gives a sufficient condition in terms of the probability distortion on losses.

Theorem 5. *(Jin and Zhou 2009, Theorem 3) Assume that $x_0 < E[\rho B]$, and $\inf_{c>0} k(c) > 1$. If there exists $\gamma < \alpha$ such that $\liminf_{p\downarrow 0} \frac{w_-(p)}{p^\gamma} > 0$, or equivalently (by l'Hôpital's rule), $\liminf_{p\downarrow 0} \frac{w'_-(p)}{p^{\gamma-1}} > 0$, then (12) must admit an optimal solution $c^* > 0$ and hence (13) solves (4).*

The conditions of Theorem 5 stipulate that the curvatures of the probability distortion on losses around 0 must be sufficiently significant in relation to her risk-seeking level (characterised by α). In other words, the agent must have a strong fear on the event of huge losses, in that she exaggerates its (usually) small probability, to the extent that it overrides her risk-seeking behavior in the loss domain.

If, on the other hand, the agent is not sufficiently fearful of big losses, then the risk-seeking part dominates and the problem is ill-posed, as stipulated in the following result.

Theorem 6. *(Jin and Zhou 2009, Proposition 1) Assume that $x_0 < E[\rho B]$. If there exists $\gamma \geq \alpha$ such that $\limsup_{p\downarrow 0} \frac{w_-(p)}{p^\gamma} < +\infty$, then $\inf_{c\geq 0} k(c) = 0 < 1$, and hence Problem (4) is ill-posed.*

3. Choquet Maximisation and Beyond: Quantile Formulation

3.1. The gain part problem. To solve the gain part problem (6), we may consider a more general maximisation problem involving the Choquet integral:

$$\begin{aligned} & \underset{X}{\text{Maximise}} && C(X) := \int_0^{+\infty} w(P(u(X) > y)) dy \\ & \text{subject to} && E[\rho X] = a, \quad X \geq 0, \end{aligned} \tag{14}$$

where $a \geq 0$, $w(\cdot) : [0, 1] \mapsto [0, 1]$ is a non-decreasing, differentiable function with $w(0) = 0$, $w(1) = 1$, and $u(\cdot)$ is a strictly concave, strictly increasing, twice differentiable function with $u(0) = 0$, $u'(0) = +\infty$, $u'(+\infty) = 0$.

Although $u(\cdot)$ in this case is concave (instead of S-shaped), the preference functional $C(X)$ is still non-concave/non-convex in X , due to the probability distortion. The technique to overcome this difficulty is what we call the “quantile formulation”, namely to change decision variable of Problem (14) from the random variable X to its quantile function $G(\cdot)$ (which is an appropriate inverse function of the CDF of X). This transformation will recover the concavity (in terms of $G(\cdot)$) for (14), as will be shown shortly.

The key properties of Problem (14) that make the quantile formulation work are the *law-invariance* of the preference functional $C(X)$ (namely $C(X) = C(Y)$

if $X \sim Y$) and the *monotonicity* of its supremum value with respect to the initial wealth a (as both $w(\cdot)$ and $u(\cdot)$ are increasing functions). The general logic of the quantile formulation goes like this: since $X \sim G_X(Z)$ for *any* $Z \sim U(0, 1)$, where G_X is the quantile of X and $U(0, 1)$ is the uniform distribution on $(0, 1)$, we can replace X by $G_X(Z)$ without altering the value of $C(X)$. Now, since the value of Problem (14) is increasing in the initial price a , the optimal $G_X(Z)$ is necessarily the one that has the *cheapest* price, namely, one that makes $E[\rho G_X(Z)]$ the smallest. There is a beautiful result which states that $E[\rho G_X(Z)]$ achieves its minimum (over all possible $Z \sim U(0, 1)$) at $Z_\rho := 1 - F_\rho(\rho)$. The precise statement of the result is as follows.

Lemma 1. $E[\rho G_X(Z_\rho)] \leq E[\rho X]$ for any lower bounded random variable X whose quantile is $G_X(\cdot)$. Furthermore, if $E[\rho G_X(Z_\rho)] < \infty$, then the inequality becomes equality if and only if $X = G_X(Z_\rho)$, a.s..

This lemma was originally due to Dybvig (1988) where a detailed proof for a finite discrete probability space was provided. The exact form of the lemma for general probability spaces needed for the present article was proved, with a different proof than Dybvig (1988), in Jin and Zhou (2008). The proof is based upon a lemma (Jin and Zhou 2008, Lemma B.1), which is closely related to the so-called Hardy–Littlewood’s inequality (Hardy, Littlewood and Pòlya 1952, p. 278) in an integral form.

It follows from $Z_\rho := 1 - F_\rho(\rho)$ that $\rho = F_\rho^{-1}(1 - Z_\rho)$. Substituting this to (14) we can therefore consider the following problem

$$\begin{aligned} & \underset{G_X(\cdot)}{\text{Maximise}} && C(G_X(Z_\rho)) \\ & \text{subject to} && E[F_\rho^{-1}(1 - Z_\rho)G_X(Z_\rho)] = a, \quad G(\cdot) \in \mathbb{G}, \quad G(0+) \geq 0, \end{aligned} \tag{15}$$

where \mathbb{G} is the set of quantile functions of lower bounded random variables.

It may appear as if (15) were more complicated than (14), but it is actually not. Recall

$$\begin{aligned} C(X) &= \int_0^{+\infty} u(x)d[-w(1 - F_X(x))] \\ &= \int_0^{+\infty} u(x)w'(1 - F_X(x))dF_X(x) \\ &= \int_0^1 u(G_X(z))w'(1 - z)dz \\ &= E[u(G_X(Z_\rho))w'(1 - Z_\rho)], \end{aligned} \tag{16}$$

indicating that $C(X)$, while not concave in X , is indeed *concave* in $G_X(\cdot)$ and the presence of the distortion $w(\cdot)$ now becomes harmless.

We can then rewrite Problem (15) as follows

$$\begin{aligned} &\text{Maximise}_{G(\cdot)} \quad \tilde{C}(G(\cdot)) = \int_0^1 u(G(z))w'(1-z)dz \\ &\text{subject to} \quad \int_0^1 F_\rho^{-1}(1-z)G(z)dz = a, \quad G(\cdot) \in \mathbb{G}, \quad G(0+) \geq 0. \end{aligned} \tag{17}$$

The above problem can be solved rather thoroughly via the Lagrange approach (see the next subsection). Finally, if $G^*(\cdot)$ solves (17), then we can recover the optimal terminal wealth X^* by the following formula

$$X^* = G^*(1 - F_\rho(\rho)). \tag{18}$$

3.2. General solution scheme for quantile formulation. Indeed, the law-invariance and monotonicity are inherent and common in many different continuous-time portfolio choice models, including expected utility maximisation, mean-variance, goal reaching, Yaari’s dual model, Lopes’ SP/A model, as well as those explicitly involving VaR and CVaR in preferences and/or constraints. Thus, like the gain part problem (6), these models all have quantile formulation and can be solved in a similar manner (although there may be technical subtleties with some of them); see He and Zhou (2010).

Let us consider the following general quantile formulation

$$\begin{aligned} &\text{Maximise}_{G(\cdot)} \quad U(G(\cdot)) = \int_0^1 u(G(z))\psi(z)dz \\ &\text{subject to} \quad \int_0^1 F_\rho^{-1}(1-z)G(z)dz = x_0, \quad G(\cdot) \in \mathbb{G} \cap \mathbb{M}, \end{aligned} \tag{19}$$

where $\psi(z) \geq 0$ satisfies $\int_0^1 \psi(z)dz = 1$ and \mathbb{M} specifies some other constraints on quantiles.

The solution scheme starts with removing the budget constraint in (19) via a Lagrange multiplier $\lambda \in \mathbb{R}$ and considering the following problem

$$\begin{aligned} &\text{Maximise}_{G(\cdot)} \quad U_\lambda(G(\cdot)) := \int_0^1 u(G(z))\psi(z)dz - \lambda \left(\int_0^1 F_\rho^{-1}(1-z)G(z)dz - x_0 \right) \\ &\text{subject to} \quad G(\cdot) \in \mathbb{G} \cap \mathbb{M}. \end{aligned} \tag{20}$$

In solving the above problem one usually ignores the constraint, $G(\cdot) \in \mathbb{G} \cap \mathbb{M}$, in the first instance, since in many cases the optimal solution of the resulting unconstrained problem could be modified to satisfy this constraint under some reasonable assumptions. For some cases such a modification could be technically challenging; see for example the SP/A model tackled in He and Zhou (2008). In other cases the constraint may need to be dealt with separately, via techniques specific to each problem.

Once (20) is solved with an optimal solution $G_\lambda^*(\cdot)$, one then finds $\lambda^* \in \mathbb{R}$ that binds the original budget constraint, namely,

$$\int_0^1 F_\rho^{-1}(1-z)G_{\lambda^*}^*(z)dz = x_0.$$

The existence of such λ^* can usually be obtained by examining the monotonicity and continuity of $f(\lambda) := \int_0^1 F_\rho^{-1}(1-z)G_\lambda^*(z)dz$ in λ . Moreover, if the strict monotonicity can be established, then λ^* is unique.

Finally, $G^*(\cdot) := G_{\lambda^*}^*(\cdot)$ can be proved to be the optimal solution to (19). This is shown in the following way. Let $v(x_0)$ and $v_\lambda(x_0)$ be respectively the optimal value of (19) and (20). By their very definitions we have the following *weak duality*

$$v(x_0) \leq \inf_{\lambda \in \mathbb{R}} v_\lambda(x_0) \quad \forall x_0 \in \mathbb{R}.$$

However,

$$v(x_0) \leq \inf_{\lambda \in \mathbb{R}} v_\lambda(x_0) \leq v_{\lambda^*}(x_0) = U_{\lambda^*}(G^*(\cdot)) = U(G^*(\cdot)) \leq v(x_0).$$

This implies that $G^*(\cdot)$ is optimal to (19) (and, therefore, the *strong duality* $v(x_0) = \inf_{\lambda \in \mathbb{R}} v_\lambda(x_0)$ holds).

The uniqueness of the optimal solution can also be derived from that of (20). Indeed, suppose we have established the uniqueness of optimal solution to (20) for $\lambda = \lambda^*$, and λ^* is such that $G_{\lambda^*}^*(\cdot)$ binds the budget constraint. Then $G_{\lambda^*}^*(\cdot)$ is the unique optimal solution to (19). To see this, assume there exists another optimal solution $\tilde{G}^*(\cdot)$ to (19). Then

$$U_{\lambda^*}(\tilde{G}^*(\cdot)) \leq U_{\lambda^*}(G_{\lambda^*}^*(\cdot)) = v(x_0) = U(\tilde{G}^*(\cdot)) = U_{\lambda^*}(\tilde{G}^*(\cdot)).$$

Hence, by the uniqueness of optimal solution to (20), we conclude $\tilde{G}^*(\cdot) = G_{\lambda^*}^*(\cdot)$.

Finally, once (19) has been solved with the optimal solution $G^*(\cdot)$, the corresponding optimal terminal cash flow can be recovered by

$$X^* = G^*(Z_\rho) \equiv G^*(1 - F_\rho(\rho)). \quad (21)$$

The general expression (21) shows that the optimal terminal wealth is *anti-comonotonic* with respect to the pricing kernel. One of its implications is that the mutual fund theorem holds in any market (complete or incomplete, with possible conic constraints on portfolios) having a deterministic opportunity set so long as all the agents follow the general model (19); see He and Zhou (2010), Theorem 5. Note that such a model covers a very diversified risk–return preferences including those of the classical utility maximisation, mean-variance and various behavioural models. Hence, the mutual fund theorem is somewhat inherent in financial portfolio selection, at least in markets with deterministic opportunity sets. As a consequence, the same risky portfolio is being held across neoclassical (rational) and behavioural (irrational) agents in the market. This, in turn, will shed light on the market equilibrium and capital asset pricing in markets where rational and irrational agents co-exist.

3.3. An example: Goal-reaching model. Let us demonstrate the preceding solution scheme by solving the following goal-reaching model:

$$\begin{aligned} & \underset{X}{\text{Maximise}} && P(X \geq b) \\ & \text{subject to} && E[\rho X] = x_0, X \geq 0, X \text{ is } \mathcal{F}_T\text{-measurable,} \end{aligned} \tag{22}$$

where $b > 0$ is the *goal* (level of wealth) intended to be reached by time T . This is called the *goal-reaching* problem, which was proposed by Kulldorff (1993), Heath (1993), and studied extensively (including various extensions) by Browne (1999, 2000).

First, if $x_0 \geq bE[\rho]$, then a trivial optimal solution is $X^* = b$ and the optimal value is 1. Therefore we confine us to the only interesting case $0 < x_0 < bE[\rho]$. Notice

$$P(X \geq b) = \int_0^{+\infty} \mathbf{1}_{\{x \geq b\}} dF_X(x) = \int_0^1 \mathbf{1}_{\{G(z) \geq b\}} dz,$$

and $X \geq 0$ is equivalent to $G(0+) \geq 0$. Hence problem (22) can be formulated in the following quantile version:

$$\begin{aligned} & \underset{G(\cdot)}{\text{Maximise}} && U(G(\cdot)) = \int_0^1 \mathbf{1}_{\{G(z) \geq b\}} dz \\ & \text{subject to} && \int_0^1 F_\rho^{-1}(1-z)G(z) dz = x_0, \\ & && G(\cdot) \in \mathbb{G}, G(0+) \geq 0. \end{aligned} \tag{23}$$

This, certainly, specialises the general model (19) with a non-convex/concave “utility” function $u(x) = \mathbf{1}_{\{x \geq b\}}$ and $\psi(z) \equiv 1$.

Introducing the Lagrange multiplier $\lambda > 0$ (as will be evident from below in this case we need only to consider positive multipliers), we have the following family of problems

$$\begin{aligned} & \underset{G(\cdot)}{\text{Maximise}} && U_\lambda(G(\cdot)) := \int_0^1 [\mathbf{1}_{\{G(z) \geq b\}} - \lambda F_\rho^{-1}(1-z)G(z)] dz + \lambda x_0 \\ & \text{Subject to} && G(\cdot) \in \mathbb{G}, G(0+) \geq 0. \end{aligned} \tag{24}$$

Ignore the constraints for now, and consider the pointwise maximisation of the integrand above in the argument $x = G(z)$: $\max_{x \geq 0} [\mathbf{1}_{\{x \geq b\}} - \lambda F_\rho^{-1}(1-z)x]$. Its optimal value is $\max\{1 - \lambda F_\rho^{-1}(1-z)b, 0\}$ attained at $x^* = b \mathbf{1}_{\{1 - \lambda F_\rho^{-1}(1-z)b \geq 0\}}$. Moreover, such an optimal solution is unique whenever $1 - \lambda F_\rho^{-1}(1-z)b > 0$. Thus, we define

$$G_\lambda^*(z) := b \mathbf{1}_{\{1 - \lambda F_\rho^{-1}(1-z)b \geq 0\}}, \quad 0 < z < 1,$$

which is nondecreasing in z . Taking the left-continuous modification of $G_\lambda^*(\cdot)$ to be the optimal solution of (24), and the optimal solution is unique up to a null Lebesgue measure.

Now we are to find $\lambda^* > 0$ binding the budget constraint so as to conclude that $G_{\lambda^*}^*(\cdot)$ is the optimal solution to (23). To this end, let

$$\begin{aligned} f(\lambda) &:= \int_0^1 F_\rho^{-1}(1-z)G_\lambda^*(z)dz \\ &= b \int_0^1 F_\rho^{-1}(1-z)\mathbf{1}_{\{F_\rho^{-1}(1-z) \leq 1/(\lambda b)\}}dz \\ &= b \int_0^{+\infty} x\mathbf{1}_{\{x \leq 1/(\lambda b)\}}dF_\rho(x) \\ &= bE[\rho\mathbf{1}_{\{\rho \leq 1/(\lambda b)\}}], \quad \lambda > 0. \end{aligned}$$

It is easy to see that $f(\cdot)$ is nonincreasing, continuous on $(0, +\infty)$, with $\lim_{\lambda \downarrow 0} f(\lambda) = bE[\rho]$ and $\lim_{\lambda \uparrow +\infty} f(\lambda) = 0$. Therefore, for any $0 < x_0 < bE[\rho]$, there exists $\lambda^* > 0$ such that $f(\lambda^*) = x_0$ or the budget constraint holds. As per discussed in the general solution scheme the corresponding $G_{\lambda^*}^*(\cdot)$ solves (23) and the terminal payment $X^* = G_{\lambda^*}^*(1 - F_\rho(\rho)) = b\mathbf{1}_{\{\rho \leq c^*\}}$, where $c^* \equiv (\lambda^*b)^{-1}$ is such that the initial budget constraint binds, solves the original problem (22). Finally, the optimal solution is unique and the optimal value is $P(X^* \geq b) = P(\rho \leq c^*) = F_\rho(c^*)$.

To summarise, we have

Theorem 7. (He and Zhou 2010, Theorem 1) Assume that $0 < x_0 < bE[\rho]$. Then the unique solution to the goal-reaching problem (22) is $X^* = b\mathbf{1}_{\{\rho \leq c^*\}}$ where $c^* > 0$ is the one such that $E[\rho X^*] = x_0$. The optimal value is $F_\rho(c^*)$.

The solution above certainly reduces to that of Browne (1999) when the investment opportunity set is deterministic. However, the approach in Browne (1999) is rather ad hoc, in that a value function of the problem is *conjectured* and then verified to be the solution of the HJB equation. In contrast, the quantile approach *derives* the solution (without having to know its form a priori). Thus it can be easily adapted to more general settings. Indeed, the HJB equation fails to work with a stochastic investment opportunity set, which however can be treated by the quantile formulation at ease.

The quantile-based optimisation is proposed by Schied (2004, 2005) to solve a class of convex, robust portfolio selection problems, and employed by Dana (2005) and Carlier and Dana (2006) to study calculus of variations problems with law-invariant concave criteria. The results presented here are mainly taken from He and Zhou (2010) where the quantile approach is systematically developed into a general paradigm in solving non-expected, non-convex/concave utility maximization models, including both neoclassical and behavioural ones. The technique has been further applied to solve a continuous-time version of the SP/A model (He and Zhou 2008), a general risk-return model where the risk is quantified by a coherent risk measure (He, Jin and Zhou 2009) and an optimal stopping problem involving probability distortions (Xu and Zhou 2009).

4. Choquet Minimization: Combinatorial Optimisation in Function Spaces

The loss part problem (7) is a special case of the following general Choquet minimisation problem:

$$\begin{aligned} &\text{Minimise}_X \quad C(X) := \int_0^{+\infty} w(P(u(X) > y))dy \\ &\text{subject to} \quad E[\rho X] = a, \quad X \geq 0, \end{aligned} \tag{25}$$

where $a \geq 0$, $w(\cdot) : [0, 1] \mapsto [0, 1]$ is a non-decreasing, differentiable function with $w(0) = 0$, $w(1) = 1$, and $u(\cdot)$ is strictly increasing, concave, strictly concave at 0, with $u(0) = 0$.

A quantile formulation transforms (25) into

$$\begin{aligned} &\text{Minimise}_{G(\cdot)} \quad \tilde{C}(G(\cdot)) = \int_0^1 u(G(z))w'(1 - z)dz \\ &\text{subject to} \quad \int_0^1 F_\rho^{-1}(z)G(z)dz = a, \quad G(\cdot) \in \mathbb{G}, \quad G(0+) \geq 0. \end{aligned} \tag{26}$$

Compared with (17), a critically different feature of (26) is that a *concave* functional is to be *minimised*. This, of course, originates from the *S-shaped* utility function in the CPT portfolio selection problem. The solution of (26) must have a very different structure compared with that of (17), which in turn requires a completely different technique (different from the Lagrange approach) to obtain. Specifically, the solution should be a ‘‘corner point solution’’; in other words, the problem is essentially a combinatorial optimisation in an infinite dimensional space, which is a generally very challenging problem even in a finite dimension.

The question now is how to characterise a corner point solution in the present setting. A bit of reflection reveals that such a solution must be a step function, which is made precise in the following result.

Proposition 1. (*Jim and Zhou 2008, Propositions D.1 and D.2*) *The optimal solution to (26), if it exists, must be in the form $G^*(z) = q(b)\mathbf{1}_{(b,1)}(z)$, $z \in [0, 1)$, with some $b \in [0, 1)$ and $q(b) := \frac{a}{E[\rho\mathbf{1}_{\{F_\rho(\rho) > b\}}]}$. Moreover, in this case, the optimal solution to (25) is $X^* = G^*(F_\rho(\rho))$.*

Since $G(\cdot)$ in Proposition 1 is uniformly bounded in $z \in [0, 1)$, it follows that any optimal solution X^* to (25) must be uniformly bounded from above.

Proposition 1 suggests that we only need to find an optimal *number* $b \in [0, 1)$ so as to solve Problem (26), which motivates the introduction of the following problem

$$\begin{aligned} &\text{Minimise}_b \quad f(b) := \int_0^1 u(G(z))w'(1 - z)dz \\ &\text{subject to} \quad G(\cdot) = \frac{a}{E[\rho\mathbf{1}_{\{F_\rho(\rho) > b\}}]}\mathbf{1}_{(b,1]}(\cdot), \quad 0 \leq b < 1. \end{aligned} \tag{27}$$

Problem (25) is then solved completely via the following result.

Theorem 8. (Jin and Zhou 2008, Theorem D.1) Problem (25) admits an optimal solution if and only if the following problem

$$\min_{0 \leq c < \bar{\rho}} u \left(\frac{a}{E[\rho \mathbf{1}_{\{\rho > c\}}]} \right) w(P(\rho > c))$$

admits an optimal solution c^* , in which case the optimal solution to (25) is $X^* = \frac{a}{E[\rho \mathbf{1}_{\{\rho > c^*\}}]} \mathbf{1}_{\rho > c^*}$.

5. Concluding Remarks

A referee who reviewed one of our mathematical behavioural finance papers questioned, ‘There is a fundamental inconsistency underlying the problem being considered in this paper. The CPT is a theory that explains how investors are “irrational” - by over emphasising losses over gains, and by under emphasising very high and very low probabilities. In this paper the authors propose that the investor *rationally* account for their irrationalities (implicit in the CPT value function). How is this justified?’

A very good question indeed. Here is our response to the question.

‘Although *irrationality* is the central theme in behavioural finance, irrational behaviours are by no means random or arbitrary. As pointed out by Dan Ariely, a behavioural economist, in his best-seller *Predictably Irrational* (Ariely 2008), “misguided behaviors ... are systematic and predictable – making us predictably irrational”. People working in behavioural finance have come up with various particular CPT values functions and probability weighting functions to examine and investigate the consistency, predictability, and rationality behind what appear as inconsistent, unpredictable and irrational human behaviours. These functions are dramatically different from those in a neoclassical model so as to systematically capture certain aspects of irrationalities such as risk-seeking, and hope and fear (reflected by the probability distortions). Tversky and Kahneman (1992) themselves state “a parametric specification for CPT is needed to provide a ‘parsimonious’ description of the the data”. As in many other behavioural finance papers, here we use CPT and specific value functions as the carrier for exploring the “predictable irrationalities”.’

To explore the consistent inconsistencies and the predictable unpredictabilities – it is the principal reason why one needs to research on “mathematising behavioural finance”. The research is still in its infancy, but the potential is unlimited – or so we believe.

References

- [1] M. Allais. Le comportement de l’homme rationnel devant le risque, critique des postulats et axiomes de l’école americaine, *Econometrica*, 21:503–546, 1953.

-
- [2] D. Ariely. *Predictably Irrational*, HarperCollins, New York, 2008.
- [3] G.W. Bassett, Jr., R. Koenker and G. Kordas. Pessimistic portfolio allocation and Choquet expected utility, *Journal of Financial Econometrics*, 2:477–492, 2004.
- [4] S. Benartzi and R. H. Thaler. Myopic loss aversion and the equity premium puzzle, *The Quarterly Journal of Economics*, 110(1):73–92, 1995.
- [5] A. Berkelaar, R. Kouwenberg, and T. Post. Optimal portfolio choice under loss aversion, *The Review of Economics and Statistics*, 86(4):973–987, 2004.
- [6] C. Bernard and M. Ghossoub. Static portfolio choice under cumulative prospect theory, Working paper, Available at <http://ssrn.com/abstract=1396826>, 2009.
- [7] F. Black and M. Scholes. The pricing of options and corporate liability, *Journal of Political Economy*, 81:637–659, 1973.
- [8] S. Browne. Reaching goals by a deadline: Digital options and continuous-time active portfolio management, *Advances in Applied Probability*, 31(2):551–577, 1999.
- [9] S. Browne. Risk-constrained dynamic active portfolio management, *Management Science*, 46(9):1188–1199, 2000.
- [10] G. Carlier and R.A. Dana. Rearrangement Inequalities in Non-Convex Insurance Models, *Journal of Mathematical Economics* 41(4–5): 485–503, 2005.
- [11] J. Cvitanić and I. Karatzas. Convex duality in constrained portfolio optimization, *Annals of Applied Probability*, 2(4):767–818, 1992.
- [12] R.A. Dana. A representation result for concave Schur functions, *Mathematical Finance*, 15(4):613–634, 2005.
- [13] E. De Giorgi and T. Post. Second order stochastic dominance, reward-risk portfolio selection and CAPM, *Journal of Financial and Quantitative Analysis*, 43:525–546, 2008.
- [14] D. Denneberg. *Non-Additive Measure and Integral*, Kluwer, Dordrecht, 1994.
- [15] P. H. Dybvig. Distributional analysis of portfolio choice, *Journal of Business*, 61(3):369–398, 1988.
- [16] D. Ellsberg. Risk, ambiguity and the Savage axioms, *Quarterly Journal of Economics*, 75:643–669, 1961.
- [17] H. Föllmer and D. Kramkov. Optional decompositions under constraints, *Probability Theory and Related Fields*, 109(1):1–25, 1997.
- [18] M. Friedman and L.J. Savage, The utility analysis of choices involving risk, *Journal of Political Economy*, 56:279–304, 1948.
- [19] F. J. Gomes. Portfolio choice and trading volume with loss-averse investors, *Journal of Business*, 78(2):675–706, 2005.
- [20] G.H. Hardy, J. E. Littlewood and G. Pòlya. *Inequalities*, Cambridge University Press, Cambridge, 1952.
- [21] J. M. Harrison and D. M. Kreps. Martingales and arbitrage in multiperiod security markets. *Journal of Economic Theory*, 20(3):381–408, June 1979.

- [22] J. M. Harrison and S. R. Pliska. Martingales and stochastic integrals in the theory of continuous trading, *Stochastic Processes and their Applications*, 11(3):215–260, 1981.
- [23] X. D. He, H. Jin and X. Y. Zhou. Portfolio selection under a coherent risk measure, Working Paper, University of Oxford, 2009.
- [24] X. D. He and X. Y. Zhou. Hope, fear, and aspiration, Working Paper, University of Oxford, 2008.
- [25] X. D. He and X. Y. Zhou. Behavioral portfolio choice: An analytical treatment, Working paper, Available at <http://ssrn.com/abstract=1479580>, 2009.
- [26] X. D. He and X. Y. Zhou. Portfolio choice via quantiles, To appear in *Mathematical Finance*, 2010.
- [27] D. Heath. A continuous time version of Kulldorff’s result, Unpublished manuscript, 1993.
- [28] H. Jin, Z. Xu and X.Y. Zhou. A convex stochastic optimization problem arising from portfolio selection, *Mathematical Finance*, 81:171–183, 2008.
- [29] H. Jin and X. Y. Zhou. Behavioral portfolio selection in continuous time, *Mathematical Finance*, 18:385–426, 2008. *Erratum*, To appear in *Mathematical Finance*, 2010.
- [30] H. Jin and X. Y. Zhou. Greed, leverage, and potential losses: A prospect theory perspective, Working paper, Available at <http://ssrn.com/abstract=1510167>, 2009.
- [31] D. Kahneman and A. Tversky. Prospect theory: An analysis of decision under risk, *Econometrica*, 47:263–291, 1979.
- [32] I. Karatzas and S. E. Shreve. *Methods of Mathematical Finance*, Springer, New York, 1998.
- [33] M. Kulldorff. Optimal control of favourable games with a time limit, *SIAM. Journal of Control and Optimization*, 31(1):52–69, 1993.
- [34] E. Lehmann. *Testing Statistical Hypotheses* (2nd edition), Wiley, 1986.
- [35] L. L. Lopes. Between hope and fear: The psychology of risk, *Advances in Experimental Social Psychology*, 20:255–295, 1987.
- [36] L. L. Lopes and G. C. Oden. The role of aspiration level in risk choice: A comparison of cumulative prospect theory and sp/a theory, *Journal of Mathematical Psychology*, 43(2):286–313, 1999.
- [37] H. Markowitz. Portfolio selection, *Journal of Finance*, 7(1):77–91, 1952.
- [38] R. Mehra and E.C. Prescott. The equity premium: A puzzle, *Journal of Monetary Economics*, 15:145–161, 1985.
- [39] R. C. Merton. Lifetime portfolio selection under uncertainty: the continuous-time case, *Review of Economics and Statistics*, 51(3):247–257, 1969.
- [40] R. C. Merton. Optimum consumption and portfolio rules in a continuous-time model, *Journal of Economic Theory*, 3:373–413, 1971.
- [41] R. C. Merton. Theory of rational option pricing, *Bell Journal of Economics and Management Sciences*, 4:141–183, 1973.

-
- [42] J. von Neumann and O. Morgenstern. *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, 1944.
 - [43] A. Schied. On the Neyman-Pearson problem for law-invariant risk measures and robust utility functionals, *Annals of Applied Probability*, 14:1398–1423, 2004.
 - [44] A. Schied. Optimal investments for robust utility functionals in complete market models, *Mathematics of Operations Research*, 30:750–764, 2005.
 - [45] H. Shefrin and M. Statman. Behavioral portfolio theory, *Journal of Financial and Quantitative Analysis* 35(2):127–151, 2000.
 - [46] A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty, *Journal of Risk and Uncertainty*, 5:297–323, 1992.
 - [47] Z. Xu and X. Y. Zhou. Optimal stopping with distorted probabilities, Working Paper, University of Oxford, 2009.
 - [48] M. E. Yaari. The dual theory of choice under risk, *Econometrica*, 55(1):95–115, 1987.

This page is intentionally left blank

Section 19

**Mathematics Education
and Popularization of
Mathematics**

This page is intentionally left blank

Professional Knowledge Matters in Mathematics Teaching[†]

Jill Adler*

Abstract

In this paper, I argue that mathematics teachers' professional knowledge matters, and so requires specific attention in mathematics teacher education. Two examples from studies of mathematics classrooms in South Africa are described, and used to illustrate what mathematics teachers use, or need to use, and how they use mathematics in their practice: in other words, the substance of their mathematical work. Similarities and differences across these examples, in turn, illuminate mathematics teachers' professional knowledge, enabling a return to, and critical reflection on, mathematics teacher education.

Mathematics Subject Classification (2010). 97C60, 97C70 and 97D99

Keywords. Mathematics for teaching; Mathematics teacher education; mathematical reasoning; mathematical objects and processes.

1. Introduction

This paper explores mathematics teachers' professional knowledge, and illuminates how and why this matters in teaching and so for teacher education. This exploration builds on the seminal work of Lee Shulman. In the mid-1980s Shulman argued cogently for a shift in understanding, in research in particular, of the professional knowledge base of teaching. He highlighted the importance of content knowledge in and for teaching, criticising research that examined teaching activity without any concern for the content of that teaching. He described the various components of the knowledge base for teaching, arguing that content knowledge for teaching (what I refer to in this paper as

[†]This paper builds on the paper presented at the Australian Association of Mathematics Teachers Conference in Perth, July, 2009, entitled Mathematics for Teaching Matters.

*This material is based upon work supported by the National Research Foundation under Grant number FA2006031800003.

School of Education, University of the Witwatersrand, Private Bag 3, WITS 2050, South Africa. E-mail: jill.adler@wits.ac.za.

professional knowledge) included subject matter knowledge (SMK), pedagogical content knowledge (PCK) – that particular blend of mathematics with concerns of learners and learning, and curriculum knowledge (cf. [14], [15]). Shulman's work, and particularly his notion of PCK, set off a research agenda, with a great deal focused on mathematics. This paper draws from and builds on the mathematical elaboration of Shulman's work.

The profound insight of Shulman's work was that being able to reason mathematically, for example, was necessary but not sufficient for being able to teach others to reason mathematically. Being able to teach mathematical reasoning involves recognising mathematical reasoning in others' discourse, and at various curriculum levels, being able to design and adapt tasks towards purposes that support mathematical reasoning, and critically working with or mediating the development of such in others. We could say the same for being able to solve algebraic or numeric problems. Most mathematics teachers and mathematics teacher educators would agree with this assertion. Yet, in the particular case of mathematical reasoning, its actuality in curricular texts, classroom practices and learner performances remains a challenge in many, if not most, classrooms [16]. We could say the same for learner performance in many areas of mathematics, including algebra. Despite the longevity and consistency of elementary algebra in school mathematics curricula worldwide, large numbers of learners experience difficulty with this powerful symbolic system [8].

In this paper I argue that strengthening our understanding of mathematics teachers' professional knowledge is a critical dimension of enhancing its teaching and learning. Mathematics teachers' professional knowledge matters, as do its implications for mathematics teacher education. I will develop this argument through examples from school mathematics classrooms that focus on generalizing as a key aspect of mathematical reasoning, together with comment on developments in mathematics teacher education in South Africa. Ultimately, the argument in this paper poses considerable challenges for mathematics teacher education.

2. Teaching and Learning Mathematics in South Africa

Post-apartheid South Africa has witnessed rapid and intense policy and curriculum change. New mathematics curricula are being implemented in schools across Grades 1-12, where there is greater emphasis than before on sense-making, problem-solving and mathematical processes, including mathematical reasoning, as well as on new topics related to data handling and financial mathematics. New education policy and curricula have strong equity goals, a function of the deep and racialised inequality produced under apartheid that affected teachers and learners alike. New policies and qualifications have been introduced into teacher education, with goals for improving the quality of teachers

and teaching. In the case of mathematics, there is also a quantitative goal - of need to address enduring critical shortages of qualified secondary mathematics teachers. Tertiary institutions have responded, offering new degree and diploma programs for upgrading teachers in service, retraining into teaching, and preparing new teachers.

It is in moments of change that taken-for-granted practices are unsettled, in both inspiring and disconcerting ways. Moments of change thus provide education researchers and practitioners with challenging opportunities for learning and reflection. Of pertinence to this paper is that the challenge of new curricula in schools and thus new demands for learning and teaching, on top of redress, bring issues like the selection of knowledges for teacher education development and support to the fore. Mathematics teacher educators in all tertiary institutions have had the opportunity and challenge to make decisions on what knowledge(s) to include and exclude in their programs, and how these are to be taught/learned. This has meant deliberate attention to what mathematics, mathematics education and teaching knowledge¹ teachers need to know and be able to use to teach well. This is no simple task: in South Africa, teaching well encompasses the dual goals of equity and excellence. At the same time as strengthening the pool of mathematics school leavers entering the mathematical sciences and related professions, high quality teaching also entails catering for diverse learner populations, and inspiring school learners in a subject that all too often has been alienating.

Hence the question: what selections from mathematics, mathematics education and teaching are needed to provide the greatest benefit to prospective and in-service teachers?

Shulman's categories provide a starting point to answering this question. Others, particularly Ball and her colleagues working on mathematical knowledge for teaching in Michigan USA, have argued that these categories need elaboration; and that elaboration requires a deeper understanding of mathematics *teaching*, and hence, of teachers' mathematical work. In [5], Ball, Thames and Phelps have elaborated Shulman's categories, distinguishing within subject matter knowledge, between *Common* and *Specialised Content Knowledge* where the latter is what teachers in particular need to know and be able to use. Within Pedagogical Content Knowledge, they distinguish *knowledge of mathematics and students*, and *knowledge of mathematics and teaching*. PCK is embedded in (and so integrated with) tasks of teaching, that is, a set of practices teachers routinely engage in or need to engage in. In their more recent work where they examine case studies of teaching [7], Hill, Blunk, Charalambos, Lewis, Phelps, Sleep and Ball note that while their elaboration is robust, compelling and helpful, they underestimated the significance of what Shulman identified as Curriculum Knowledge. In case studies of ten teachers, they were confronted

¹Mathematics education here refers to texts related to research on the teaching and learning of mathematics; teaching refers to professional practice.

with contradictory evidence of the ways different teachers used particular curriculum materials, and how this related to their mathematical knowledge. What this reflects is that all teaching always occurs in a context and set of practices, of which curricular discourses are critical elements.

Ball et al.'s elaboration of Shulman's categories is useful, particularly as it has been derived from studies of mathematics classroom practice. They provide a framework with which to think about and make selections for teacher education. At immediate face value, they suggest that mathematical content in teacher education and for teaching requires considerable extension beyond knowing mathematics for oneself.

I go further to say we need to understand what and how such selections take shape in mathematics teacher education practice. As in school, teacher education occurs in a context and set of practices, and is shaped by these. In addition, as intimated above, in mathematics teacher education, mathematics as an "object" or "focus" of learning and becoming, is integrated with learning to teach. The research we have been doing in the QUANTUM² project in South Africa (that now has a small arm in the UK) has done most of its work in teacher education as an empirical site, complemented by studies of school mathematics classroom practice. The goal is to understand the substance of opportunities to learn mathematics in teacher education, and how this relates to mathematics teachers' professional work in their school classrooms.

In this paper, I select two examples from studies of mathematics classrooms in South Africa. I use these to illustrate what mathematics teachers use, or need to use, and how they use it in their practice: in other words, the mathematical substance of their professional work. Similarities and differences across these examples, in turn, illuminate professional knowledge for mathematics teaching, enabling a return to, and critical reflection on, mathematics teacher education.

3. Productive Mathematics Tasks

Example 1: Angle properties of a triangle. The episode discussed below is described in detail in [3]³, and takes place in a Grade 8 classroom. This teacher was particularly motivated by a participatory pedagogy, and developing her learners' broad mathematical proficiency [9]. She paid deliberate attention to supporting her learners' participation in mathematical discourse [13], which in practice involved having them learn to reason mathematically, and verbalise this. It is interesting to note that the empirical data here date back to the early

²For details on QUANTUM, a project focused on Qualifications for Teachers Underqualified in Mathematics, see [4], [6]; [1]

³The focus of the study reported in [3] was on teaching and learning mathematics in multilingual classrooms. There I discuss in detail the learners' languages, and how and why talking to learn worked in this class. I have since revisited this data, reflecting on the teachers' mathematical work (see [2]).

1990s and long before curriculum reform as it appears today in South Africa was underway.

As part of a sequence of tasks related to properties of triangles, the teacher gave the activity in Figure 1 to her Grade 8 class. The questions I will address in relation to this task are: What mathematical work is entailed in designing this kind of task, and then mediating it in a class of diverse learners?

- If any of these is impossible, explain why; otherwise, draw it.
- ◆ Draw a triangle with 3 acute angles.
 - ◆ Draw a triangle with 1 obtuse angle.
 - ◆ Draw a triangle with 2 obtuse angles.
 - ◆ Draw a triangle with 1 reflex angle.
 - ◆ Draw a triangle with 1 right angle.

Figure 1. A triangle task

The task itself evidences different elements of important mathematical work entailed in teaching learners to reason mathematically. Firstly, this is not a “typical” task on the properties of triangles. A more usual task to be found in text books, particularly at the time of the research, would be to have learners recognise (identify, categorise, name) different types of triangles, defined by various sized angles in the triangle. What the teacher has done here is recast a “recognition” task based on angle properties of triangles into a “reasoning” task (generalising about properties and so relationships). She has constructed the task so that learners are required to reason in order to proceed. In so doing, she sets up conditions for producing and supporting mathematical reasoning in the lesson and related proficiencies in her learners. In particular, whereas a recognition task will refer only to examples (of triangles), the task above incorporates non-examples. Secondly, in constructing the task so that learners need to respond whether or not particular angle combinations are “impossible” in forming a triangle, the task demands proof-like justification—an argument or explanation that, for impossibility, will hold in all cases. In this task, content (properties of triangles) and processes (reasoning, justification, proof) are integrated. The question, of course, is what and how learners attend to these components of the task, and how the teacher then mediates their thinking.

In preparation for this lesson and task, the teacher would have had to think about the mathematical resources available to this classroom community with which they could construct a general answer (one that holds in all cases). Any single task necessarily falls within some sequence or ordering of learning, and thus the need for curriculum considerations by the teacher. For example, if as was the case, learners had worked with angle sum in a triangle, what else might come into play as learners go about this task? What is it about the triangle as a mathematical object that the teacher needs to have considered and that she needs to be alert to as her learners engage in reasoning about its properties?

Before engaging further with the details of the teachers' mathematical work, let us move to the actual classroom, where students worked on their responses in pairs. The teacher moved between groups, probing with questions like: "Explain to me what you have drawn/written here?", "Are you sure?", "Will this always be the case?" She thus pushed learners to verbalise their thinking, as well as justify their solutions or proofs. I foreground here learners' responses to the second item: Draw a triangle with two obtuse angles. Interestingly, three different responses were evident.

- Some said, "It is impossible to draw a triangle with two obtuse angles, because you will get a quadrilateral." They drew the shape shown in Figure 2.



Figure 2. Student drawing of a triangle with two obtuse angles

- Others reasoned as follows: "An obtuse angle is more than 90 degrees and so two obtuse angles give you more than 180 degrees, and so you won't have a triangle because the angles must add up to 180 degrees."
- One learner (Joe) and his partner reasoned in this way: "If you start with an angle say of 89 degrees, and you stretch it [to make it larger than 90 degrees], the other angles will shrink and so you won't be able to get another obtuse angle." Their drawing is shown in Figure 3.

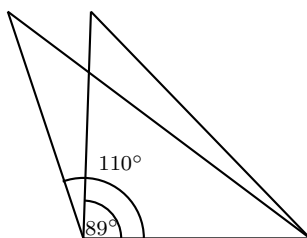


Figure 3. Joe and his partners response.

The range of learner responses to this task is indicative of a further task-based teaching skill. The task is designed with sufficient openness, and so diverse learner responses are possible and indeed elicited. In addition, the third, unexpected, response produced much interest in the class, for the teacher, and for myself as researcher. The first two responses were common across learners and more easily predicted by the teacher.

Having elicited these responses, it is the teacher's task to mediate within and across these responses, and enable her learners to reason whether each of these responses is a general one, one that holds in all cases⁴. In the many contexts where I have presented the study and this particular episode, much discussion is generated both in relation to the mathematical status of the responses, and their levels of generality, as well as simultaneous arguments as to what can be expected of learners at a Grade 8 level (13 - 14 years). What constitutes a generalised answer at this level? Are all three responses equally general? Is Joe's response a generalised one? How does the teacher value these three different responses, supporting and encouraging learners in their thinking, and at the same time judging/evaluating their mathematical worth?

These are mathematical questions, and the kind of work this teacher did on the spot as she worked to *evaluate* and *value* what the learners produced was also mathematical work. The point here is that this kind of mathematical work i.e. working to provoke, recognise and then mediate notions of proof and different kinds of justification, is critical to effective teaching of "big ideas" (like proof) in mathematics. In Ball et al.'s terms, this work entails *specialised content knowledge* (judging the mathematical generality of the responses), knowledge of *mathematics and teaching* (designing productive tasks) and *mathematics and students* (and mediating between these and learners' mathematics).

We need to ask further questions about subject matter knowledge, or content in this example, and specifically questions about the angle properties of triangles. The insertion of a triangle with a reflex angle brought this to the fore in very interesting ways. Some learners drew the following, as justification for why a triangle with a reflex angle was possible; and so provoked a discussion of concavity, and interior and exterior angles.

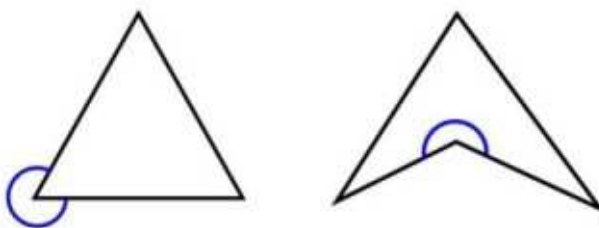


Figure 4. Learner drawings to justify triangles with reflex angles

The tasks of teaching illuminated in this example are: task design where content (angle properties of triangles) and process (reasoning, justifying) are integrated; mediation of both mathematical content and processes; and valuing and evaluating diverse learner productions. The mathematical entailments of

⁴The interesting interactions that followed in the class are described and problematised in [3] and will not be focused on here.

this work are extensive, and are illustrative of both subject matter knowledge and pedagogical content knowledge. The teacher here, in requiring learners to consider when different angle configurations are always or never possible, and mediating their partial reasoning, reflects an appreciation of mathematical proof, and an awareness of defining and generalizing as mathematical processes to be developed in relation to a specific mathematical object and its properties. To effectively mediate Joe's response and the two above, she would also need to ask suitable questions or suggest productive ways forward for these learners, so that their notions of proof and of the mathematical triangle are strengthened and progressed. Indeed, as learners in the class engaged with the second triangle drawn above, their focus was that the answer was incorrect because there were three reflex angles not one, and the teacher had a difficult time shifting them from this focus and onto the interior angles.

In [3], I show that as the teacher mediated the three different responses to the triangle with two obtuse angles, she worked explicitly to value each contribution and probe learner thinking, thus creating opportunities for all in the class to engage in valued mathematical practices. What was further interesting was that her judgment of the relative mathematical worth of each of the contributions was left implicit. She accepted the first two responses above, but probed Joe's, with questions to Joe that implied she was not convinced of the generality of his argument. At the public level however, all three responses were left as if they were all equally valid responses. A question then arises as to whether all learners in the class could read in her interactions, the implicit question of the validity of Joe's argument. In Cobb's terms, the socio-mathematical norms (whereby teachers' practices make mathematical criteria explicit in their interactions with learners) might not be sufficiently well established for all learners in this class [18].

The example here is compelling in a number of ways, and provokes the question: Where, when and how does a mathematics teacher learn to do this kind of work, and in ways that are of benefit to all learners? Before attempting to answer this and so shift back into teacher education, we need to look at additional and different examples of the mathematical work of teaching.

Example 2: Polygons and diagonals - or a version of the "mystic rose". The second example is taken from a Grade 10 class (see [11]), where the teacher posed the following task for learners to work on in groups: *How many diagonals are there in a 700-sided polygon?*

Here too, the teacher has designed or adapted a task and presented learners with an extended problem. They have to find the number of diagonals in a 700-sided polygon, a sufficiently large number to require generalising activity, and so mathematical reasoning. I pose the same questions here as for Example 1: What mathematical work is entailed in designing this kind of task, and mediating it in a class of diverse learners?

Many teachers will recognise the “mystic rose” investigation in this problem. The mathematical object here is a polygon and its properties related to diagonals. Yet the problem has been adapted from a well known (perhaps not to the teacher) mathematical investigation of points on a circle and connecting lines - a different, though related object. Here learners are not asked to investigate the relationship between the number of points on a circle and connecting lines, but instead to find an actual numerical solution to a particular polygon, albeit with a large number of sides and so approaching a circle. I have discussed this case in detail in [1], where I point out that unlike triangles and their properties, the general polygon and its properties is not an explicit element of the secondary school curriculum. However, the processes and mathematical reasoning required for learners to solve the problem are desired mathematical processes in the new curriculum.

My concern in this paper is not with the merits of the problem and its adaptation in an isolated way. Rather, I wish to reflect on the mathematical work of the teacher in presenting the problem, mediating learner progress, valuing and evaluating their responses, and managing the integration of mathematical content and mathematical processes as foci in the lesson. I present selections from the transcript of the dialogue in the classroom to illuminate these four components of the teachers’ mathematical work.

The teacher (Tr), standing in the front of the class, explained what the class had to do.

Tr: I want you to take out a single page quickly. Single page and for the next five minutes no discussion. I want you to think about how would you possibly solve this problem? (pointing to the projected problem: *How many diagonals are there in a 700-sided polygon?*)

After seven minutes, the Teacher calls the class’ attention. (Learners are referred to as Lr A, B, etc.)

Tr: Ok! Guys, time’s up. Five minutes is over. Who of you thinks they solved the problem? One, two, three, four, five, six.

Lr A: I just divided 700 by 2.

Tr: You just divided 700 by 2. (Coughs).

Lr A: Sir, one of the sides have, like a corner. Yes... [inaudible], because of the diagonals. Therefore two of the sides makes like a corner. So I just divided by two... [Inaudible].

Tr: So you just divide the 700 by 2. And what do you base that on? ...

[]

Tr: Lets hear somebody elses opinion.

Lr B: Sir what Ive done sir is . . . First 700 is too many sides to draw. So if there is four sides how will I do that sir? Then I figure that the four sides must be divided by two. Four divided by two equals two diagonals. So take 700, divide by two will give you the answer. So thats the answer I got.

Tr: *So you say that, theres too many sides to draw. If I can just hear you clearly; that 700 sides are too many sides, too big a polygon to draw. Let me get it clear. So you took a smaller polygon of four sides and drew the diagonals in there. So how many diagonals you get?*

Lr B: In a four-sided shape sir, I got two.

Tr: *Two. So you deduced from that one example that you should divide the 700 by two as well? So you only went as far as a 4 sided shape? You didnt test anything else.*

Lr B: Yes, I dont want to confuse myself.

Tr: *So you dont want to confuse yourself. So youre happy with that solution, having tested only one polygon?*

Lr B: [Inaudible response.]

Tr: *Ok! You say that you have another solution. [Points to learner D] Lets hear.*

[]

Lr A: I just think its right It makes sense.

Tr: *What about you [Lr D]? You said you agree.*

Lr D: *He makes sense. . . He proved it. . . He used a square.*

Tr: *He used a square? Are you convinced by using a square that he is right?*

Lr E: But sir, here on my page I also did the same thing. I made a 6-sided shape and saw the same thing. Because a six thing has six corners and has three diagonals.

Lr A: So what about a 5-sided shape, then sir?

Tr: *What about a 5-sided shape? You think it would have 5 corners? How many diagonals?*

I have underlined the key mathematical contributions by learners, and italicised the teachers' mediating comments and questions. These highlight the learners' reasoning and the teacher's probing for further mathematical justification.

At this point in the lesson, the teacher realises that some of the learners are confusing terms related to polygons, as well as some of the properties of a general polygon and so deflects from the problem for a while to examine with learners, various definitions (of a polygon, pentagon, a diagonal, etc.). In other words, at this point, the mathematical object in which the problem is embedded comes into focus. It is interesting to note here that at no point was there reflection on the polygons in use in developing responses to the problem. All were regular and convex. A little later in the lesson, another learner offers a third solution strategy. The three different solution representations are summarised in Figure 5, illustrating the varying orientations students adopted as they attempted to work towards the solution for a 700-sided polygon.



<u>Learner A</u>	<u>Learner B</u>	<u>Learner C</u>
700-sided polygon $700 / 2 = 350$ diagonals	4-sided polygon $4 / 2 = 2$ diagonals	7-sided polygon 14 diagonals $14 \times 100 = 1400$ diagonals
Representation: Verbal description	Representation: 	Representation: 
Reasoning: Because of sides - corners. $700/2 = 350$ corners and 175 diagonals	Reasoning: Too big a number therefore use a quadrilateral. $4/2 = 2$ diagonals therefore $700/2$	Reasoning: 7-sided polygon has 14 diagonals therefore multiply by 100 which equals 1400.

Figure 5. Three different representations and reasoning

As with Example 1, we see four tasks of teaching demanded of the teacher: task design or adaptation; mediation of learners' productions; valuing and evaluating their different responses; and managing mathematics content and processes opened up by the task.

The representations offered by learners give rise to interesting and challenging mathematical work for the teacher. All responses are mathematically flawed, though the approaches of Learners B and C show attempts at specialising and then generalising [10]. While this is an appropriate mathematical practice, the move from the special case to the general case in both responses is problematic, though in different ways. Does the teacher move into discussion about specialising and generalising in mathematics (and if so, how)? Open-ended investigations and problem-solving as described above open up possibilities for this kind of mathematical work in class. Here, the mediation by the teacher (which included

challenging questions of other learners, as well as his own questions) focused on counter examples (what about six sides? Or five sides), and did not move on to more general elaboration. Could and should the teacher have taken up such elaborations, and if so, how?

4. The Mathematics Involved in Particular Teaching Tasks

In selecting and presenting two different examples from different secondary school classrooms in South Africa, I have highlighted four inter-related tasks of teaching, each of which entails considerable mathematical skill and understanding over and above (or underpinning) the teaching moves that will ensue. The four tasks (two of which are discussed in each of the bulleted sections below) further illustrate categories of professional knowledge developed by Shulman and elaborated by Ball et al. in mathematics.

Managing processes and objects in task design and adaptation. In the first example, the process of mathematical reasoning was in focus, as was the triangle and its angle properties. I will call this an *object-and-process-focused* task. Angle properties of triangles are the focus of reasoning activity. Learners engage with and consolidate knowledge of these properties through reasoning activity, and vice versa. Here the integration of learning content and process appears to keep them both in focus, and thus provides opportunities for learning both. Example 2 is also focused on mathematical reasoning. It is a *process-focused* task, having been adapted (what I would refer to as recontextualised) from an investigation and re-framed as a problem with a solution. The mathematical object of the activity, the polygon and its diagonals, are backgrounded. At a few points in the lessons, these come into focus, when understanding polygons, their properties, and the definition of a diagonal are required for learners to make progress with the problem. Some learners, for example, assume a polygon to be regular, with an even number of sides, and a diagonal as passing through the centre; some generalise from one specific case (a four-sided figure); while others over-generalise multiplicative processes from number to polygons (if I need 700 sides, then I can start with 7 sides, draw in all diagonals, and then multiply by 100 to get all the diagonals). Picking up on the pertinence of examples and non-examples in working with mathematical objects, the contributions of learners here remain in the realm of particular examples, with most learners considering only regular convex polygons.

The intricate relationship between mathematical objects and processes has been an area of extensive empirical research in the field of mathematics education. It appears from studying two examples of teaching that selecting, adapting or designing tasks to optimise teaching and learning entails an understanding

of mathematical objects and processes and how these interact within different kinds of tasks. The teaching of mathematical content and mathematical processes is very much in focus today. Reform curricula in many countries promote the appreciation of various mathematical objects, their properties and structure, conventions (how these are used and operated on in mathematical practice), as well what counts as a mathematical argument, and the mathematical processes that support such. In Example 1, we see opportunity for developing reasoning skills, and understanding of proof at the same time as consolidating knowledge about triangles. In Example 2, it is not apparent whether and how either proof or reasoning will flourish through this example and its mediation. The relevance of the mathematical object in use is unclear. Thus the question: *Do we need a mathematics teaching curriculum that includes task interpretation, analysis and design with specific attention to intended mathematical objects and processes and their interaction?*

In other words, should a curriculum for professional knowledge for mathematics teaching include attention to the mediation of mathematical content and processes as these unfold in and through engagement with varying tasks? If so, is this to be part of the mathematics curriculum, or part of the teaching curriculum? And hidden in this last question is a question of who teaches these components of the curriculum in teacher education, because they are not traditional components of mathematics curricula, nor do they typically enter courses on teaching methodology? What competences and expertise would best support this teaching?

Valuing and evaluating diverse learner productions. Diverse learner productions are particularly evident in Examples 1 and 2, given their more open or extended nature. Thus, in each example, the teacher dealt with responses from learners that they predicted, and then those that were unexpected. In Example 1, the teacher needed to consider the mathematical validity of Joe's argument for the impossibility of a triangle with two obtuse angles, and then how to encourage him to think about this himself, and convince others in the class. Similarly, we can ask in Example 2: what might be the most productive question to ask Learner C and so challenge the proportional-type reasoning used here (that, since 700 can be factored into 7×100 , finding the diagonals in a 7-sided figure is the route to the solution to a 700-sided figure)? Such questioning in teaching needs to be mathematically informed.

Together these examples illuminate how teachers need to exercise mathematical judgment as they engage with what learners do (or do not do). This is particularly so if teachers are building a pedagogical approach and classroom environment that encourages mathematical practices where error, and partial meanings are understood as fundamental to learning mathematics. In earlier work I referred to this as a teaching dilemma, where managing both the valuing of learner participation and evaluation of the mathematical worth of their responses was important [3]; and illuminated the equity concerns if and when

evaluation of diverse responses – i.e., judgments as to which are mathematically more robust or worthwhile – are left implicit.

So, a further question needs to be asked of the curriculum in mathematics teacher education. Learner errors and misconceptions in mathematics are probably the most developed research areas in mathematics education. We know a great deal about persistent errors and misconceptions that are apparent in learners' mathematical productions across contexts. These provide crucial insight into the diverse responses that can be anticipated from learners. Yet, as Stacey [17] argues, the development of this research into contents for teacher education has been slow. We have shown elsewhere that the importance of learner mathematical thinking in mathematics teacher education is evident in varying programs in South Africa (see [6]; [12]). Yet there are significant differences in the ways this is included in such programs, and so with potential effects on who is offered what in their teacher education. *How should curriculum for mathematics teaching then include such content?*

5. Professional Knowledge Matters in Mathematics Teaching

I have argued that curriculum for mathematics teaching matters. I have suggested that what matters are task design and mediation, as well as attention to mathematical content, objects and processes within these. I have also suggested that there are equity issues at stake. I now return to the context of teacher education in South Africa where various innovative teacher education programs are grappling with a curriculum for mathematics teachers that appreciates the complexity of professional knowledge for teaching and its critical content or subject basis. I will focus here on what we have observed as objects of attention (and so meanings) shift from classrooms to teacher education and back again, observations that support the argument in this paper, that we need to embrace a deeper understanding of the complexities of teaching and so our task in teacher education.

In more activity-based, participative or discursively rich classroom mathematics practice, there is increased attention to mathematical processes as critical to developing mathematical proficiency and inducting learners into a breadth of mathematical practices. The examples in this paper illustrate how mathematical processes are always related to or based on some mathematical object. If the latter is not well understood, in the first instance by the teacher, in ways that enable her to notice when it goes out of focus or is completely missed by students, then their reasoning is likely to be flawed or mathematically empty. This phenomenon is apparent in classrooms in South Africa, and more so in historically disadvantaged settings, thus perpetuating rather than attacking inequality. Mathematical objects and processes and their interaction are the central “matter” in curricular for mathematics teaching. The shift in

new curricula to mathematical processes creates conditions for diminished attention to mathematical objects. Attention to *objects and processes* need to be embraced *in the context of teaching* if access to mathematics is to be possible for all learners.

Herein lies considerable challenge. In each of the two examples in this paper, a mathematical object was embedded in a task that worked varyingly to support mathematical reasoning processes. What the teacher in each case faced was different learner productions as responses to the task. These become the focus of the teachers' work, requiring integrated and professional based knowledge of mathematics, teaching tasks and learner thinking. So what then, is or comes into focus in teacher education, and not only into teacher education, but into school curricula? What we have observed (and I have seen elements of this in elementary mathematics teacher education in the UK), is that learner thinking and the diversity of their responses become the focus, with the mathematical objects and tasks that give rise to these, out of focus. What one might see in the case of the triangle properties is a task that requires learners to produce three different arguments for why a triangle cannot have two obtuse angles. And there is a subtle shift of attention: from how to mediate diverse responses, to multiple answers or solutions being the required competence in learners. Simply, there are curricular texts that now require learners to produce multiple solutions to a problem. I leave this somewhat provocative assertion for discussion and further debate.

In conclusion, there is an assumption at work throughout this paper that teacher education is crucial to quality teaching. In South Africa, all pre-service and formal in-service teacher education has become the responsibility of universities. Tensions between theory and practice abound. I hope in this paper to have provided examples that illuminate the mathematical work of teaching, and through these opened up challenges for mathematics teacher education. Professional knowledge for mathematics teaching, and its place in mathematics teacher education, particularly in less resourced contexts, matters profoundly.

Acknowledgements

This paper forms part of the QUANTUM research project on Mathematics for Teaching, directed by Jill Adler, at the University of the Witwatersrand. Dr Zain Davis from the University of Cape Town is a co-investigator and central to the theoretical and methodological work in QUANTUM. The elaboration into classroom teaching was enabled by the work of Masters students at the University of the Witwatersrand. This material is based upon work supported by the National Research Foundation under Grant number FA2006031800003. Any opinion, findings and conclusions or recommendations expressed in this material are those of the author and do not necessarily reflect the views of the National Research Foundation.

References

- [1] J. Adler, A methodology for studying mathematics for teaching. *Recherches en Didactique des Mathématiques*, 2009.
- [2] J. Adler, Mathematics teacher education in post-apartheid South Africa: A focus on the mathematical work of teaching across contexts. In M. Borba (Ed.) *Trends in mathematics education*, Brazil (in Portuguese, pp. 45-64). São Paulo: Autênt, 2006.
- [3] J. Adler, *Teaching mathematics in multilingual classrooms* Dordrecht: Kluwer, 2001.
- [4] J. Adler & Z. Davis, Opening another black box: Researching mathematics for teaching in mathematics teacher education. *Journal for Research in Mathematics Education* **37** (4), 270-296, 2006.
- [5] D.L. Ball, M. H. Thames & G. Phelps, Content knowledge for teaching: What makes it special? *Journal of Teacher Education*, **59**, 389-407, 2008.
- [6] Z. Davis, J. Adler & D. Parker, Identification with images of the teacher and teaching in formalized in-service mathematics teacher education and the constitution of mathematics for teaching. *Journal of Education* **42** 33-60, 2007.
- [7] H. Hill, M. Blunk, Y. Charalambos, J. Lewis, G. Phelps, L. Sleep, & D. Ball, Mathematical knowledge for teaching and the mathematical quality of instruction: An exploratory study. *Cognition and Instruction* **26**, 430-511, 2008.
- [8] J. Hodgen, D. Kuchemar, M. Brown & R. Coe, Childrens understandings of algebra 30 years on. In Joubert, M. (Ed.) *Proceedings of the British Society for Research into Learning Mathematics* **28** (3) November 2008. Available at www.bsrlm.org.uk.
- [9] J. Kilpatrick, J. Swafford & B. Findell (Eds) *Adding it up: Helping children learn mathematics* Washington: National Academy Press, 2001.
- [10] J. Mason, Generalisation and algebra: Exploiting childrens powers. In L. Haggerty (Ed.), *Aspects of teaching secondary mathematics: Perspectives on practice* (pp. 105-120) London: Routledge Falmer, 2002.
- [11] S. Naidoo, *Mathematical knowledge for teaching geometry to Grade 10 learners* Johannesburg, South Africa: School of Education, The University of the Witwatersrand, 2008.
- [12] Parker, *The specialisation of pedagogic identities in initial mathematics teacher education in post-apartheid South Africa* Unpublished PhD Thesis, University of the Witwatersrand, Johannesburg, 2009.
- [13] A. Sfard, *Thinking as communicating: Human development, the growth of discourses, and mathematizing*. Cambridge, UK: Cambridge University Press, 2008.
- [14] L. Shulman, Those who understand: knowledge growth in teaching, *Educational Researcher*, **15** (2), 4-14, 1986.
- [15] L. Shulman, Knowledge and teaching: Foundation of the new reform, *Harvard Educational Review*, **57** (1), 1-22, 1987.
- [16] K. Stacey & J. Vincent, Modes of reasoning in explanations in Australian eighth-grade mathematics textbooks, *Educational Studies in Mathematics*, published online 25 March 2009, Springer.

-
- [17] K. Stacey, International perspectives on the nature of mathematical knowledge for secondary teaching: Progress and dilemmas. In M. J. Hoines & A. B. Fuglestad (Eds) *Proceedings of the 28th Conference of the International Group for the Psychology of Mathematics Education PME 28*, Bergen, Norway, 167-196, 2004.
- [18] E. Yackel & C. Cobb, Sociomathematical norms, argumentation, and autonomy in mathematics. *Journal for Research in Mathematics Education* **27** (4), 458-477, 1996.

This page is intentionally left blank

Section 20

History of Mathematics

This page is intentionally left blank

History of Convexity and Mathematical Programming: Connections and Relationships in Two Episodes of Research in Pure and Applied Mathematics of the 20th Century

Tinne Hoff Kjeldsen*

Abstract

In this paper, the gradual introduction of the concept of a general convex body in Minkowski's work and the development of mathematical programming, are presented. Both episodes are exemplary for mathematics of the 20th century, in the sense that the former represents a trend towards a growing abstraction and autonomy in pure mathematics, whereas the latter is an example of the many new disciplines in applied mathematics that emerged as a consequence of efforts to develop mathematics into a useful tool in a wider range of subjects than previously. It will be discussed, how and why these two new areas emerged and developed through different kinds of connections and relations; and how they at some point became connected, and fed and inspired one another. The examples suggest that pure and applied mathematics are more intertwined than the division in 'pure' and 'applied' signals.

Mathematics Subject Classification (2010). Primary 01A60; Secondary 52-03, 90-03.

Keywords. History of 20th century mathematics, the theory of convexity, positive definite quadratic forms, convex sets, the lattice point theorem, mathematical programming, linear programming, nonlinear programming, the Kuhn-Tucker theorem, Minkowski, Fenchel, Tucker, Kuhn, the military-university complex, the Second World War.

*IMFUFA, NSM, Roskilde University, P.O. Box 260, 4000 Roskilde, Denmark.
E-mail: thk@ruc.dk.

1. Introduction

Mathematical ideas that have led to new developments often seem to have emerged in one piece of mathematics, from where they have become connected to other parts of mathematics. Through these connections they have initiated developments into and opened whole new areas of mathematical research - areas that then tend to become completely disconnected from the field in which the original ideas came from. The history of mathematics is filled with episodes that illustrate this pattern, and mathematical monographs are full of footnotes, stating that the idea of such and such originated in this or that work on something, that seems to belong to an entirely different area of mathematics. In monographs on the moment problem e.g., one can often read that this problem originated towards the end of the 19th century from Stieltjes' work on continued fractions - a piece of information, that will leave many readers puzzled, wondering how and why the moment problem originated in connection with continued fractions, how and why it became connected with completely different areas of mathematics, and how and why it developed into its own mathematical theory.¹ To dig out such connections and relationships, and to understand how they govern developments of and changes within mathematics, is what makes history of mathematics so interesting.

Looking at 20th century mathematics there are at least two things that leap to the eye: a trend towards abstraction and autonomy, and a migration of mathematics into a much wider range of science and social life than hitherto. The first has been analysed and discussed in terms of a 'modernism' development in mathematics in the early 20th century [16], and the second has given rise to developments of many new disciplines in applied mathematics. In this talk I will follow the development of two new areas of research in mathematics that emerged in the 20th century, and which represent these two trends, namely the theory of convexity and the emergence of mathematical programming.² I will follow their trajectories, and discuss how and why they through different kinds of connections and relations emerged, developed, became connected, and fed and inspired one another; suggesting that pure and applied mathematics are more intertwined than the division in 'pure' and 'applied' signals. At the outset, the two examples seem to have developed due to very different circumstances and through very different kinds of connections. On a closer look, however, they also share some particular traits that will be brought out, discussed and compared.

¹For answers to these questions see [19].

²For more elaborated and detailed analyses from various kind of perspectives of the historical developments of these two episodes see [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31].

2. From Number Theory to Convexity - an Episode of Connections in Pure Mathematics

The modern notion of a convex set was coined by Hermann Minkowski at the turn of the 19th century. Eighty years later the Danish mathematician Werner Fenchel wrote in his paper “Convexity through the ages” that:

Minkowski’s interest in convexity originated, *strange to say*, from the Theory of Numbers. ([10], p. 126. My emphasis)

signalling that by then, convexity was totally disconnected from number theory.

In the following, Minkowski’s work will be analysed with respect to how the concept of a general convex body gradually emerged and took form in his mathematical practice, in order to understand how he was led to define the property of convexity, and what this “strange” connection between number theory and convexity was all about. The analysis will show that it is possible to indentify three phases in the parts of Minkowski’s work, where he introduced the concept of a convex body and turned convex bodies into research objects in their own right. These phases will be explained below, and it will become clear that the dynamics of knowledge production in this particular research episode of Minkowski’s mathematical career, can be characterised as an interplay between on the one hand, posing and answering new questions, and on the other hand, treating and investigating known problems in new ways by using new or different methods, techniques or tools.

Phase 1 - the minimum problem.

In the first phase, Minkowski worked on a well-known and important problem in the reduction theory of positive, definite quadratic forms in n -variables, namely the so-called minimum problem of finding the minimum value of such a form for integer values of the variables not all zero. A quadratic form f in n variables

$$f(x_1, \dots, x_n) = \sum_{h,k=1}^n a_{h,k} x_h x_k, \quad a_{h,k} \in \mathfrak{R}, \quad a_{h,k} = a_{k,h}$$

is said to be positive definite if

$$f(x_1, \dots, x_n) > 0 \quad \text{for all } (x_1, \dots, x_n) \neq (0, \dots, 0).$$

In a letter to Jacobi, published in Crelle’s Journal in 1850, Hermite had found that there exists an integer point (x_1, \dots, x_n) such that

$$f(x) \leq \left(\frac{4}{3}\right)^{\frac{1}{2}(n-1)} \sqrt[n]{D}$$

where D denotes the determinant of the form [17]. In the same volume of Crelle’s journal there is a paper by Dirichlet, in which he presented a geometrical foundation for the theory of positive definite quadratic forms in three variables [7].

Inspired by the papers of Hermite and Dirichlet, Minkowski approached the minimum problem geometrically, by associating a positive definite quadratic form in n -variables with a n -dimensional lattice, build up of congruent (standard) parallelotopes.

The idea of interpreting a positive definite quadratic form geometrically goes back to Gauss, who in 1831 had given an outline of such an interpretation ([12], p. 188-196).³ For forms in two variables:

$$f(x, y) = ax^2 + 2bxy + cy^2$$

a system of coordinates, in which the angle ϕ between the axes is determined by the equation

$$\cos \phi = \frac{b}{\sqrt{ac}}$$

and where \sqrt{a} and \sqrt{c} measure the units of the axes, can be associated with f . For integer values of x and y the points $(x\sqrt{a}, y\sqrt{c})$ will form a pattern - a lattice - build up of equal (standard) parallelograms. The lattice points are then the points for integer values of the variables, and they form the vertices of the parallelograms. The square of the area of such a parallelogram is equal to the determinant of the form. Since the square of the distance from a lattice point $(x_0\sqrt{a}, y_0\sqrt{c})$ to the origin is equal to $f(x_0, y_0)$, the minimum problem is to determine the distance between the origin and the lattice point which is closest to the origin - in other words - to determine the smallest distance between points in the lattice, see Figure 1.

The significance of this geometrical interpretation for determining bounds for the minimum is laid out in Minkowski's probationary lecture for his Habilitation in Bonn in March, 1887.⁴ Here he explained the interpretation of such a form in three variables as a lattice of parallelotopes. He then reached an upper bound for the minimum of a positive definite quadratic form in three variables by a very elegant and intuitive geometrical argument. He let \sqrt{M} denote the smallest distance in the lattice, and imagined spheres with \sqrt{M} as diameter placed around each lattice point. He then argued that since these spheres will not overlap, and their volume is smaller than the volume of a standard parallelotope, it is possible to deduce an upper bound for the smallest distance in the lattice simply by comparing the volume of these two geometrical objects:

$$V_{\text{sphere}} < V_{\text{parallelotope}}$$

$$\frac{4}{3}\pi \left(\frac{\sqrt{M}}{2}\right)^3 < V_{\text{parallelotope}}$$

³For historical accounts of number theory, see e.g. [54], [14], [51], [52], [15]

⁴A copy of the manuscript can be found among Minkowski's papers in the Niels Bohr Library, USA. It was published in 1991 by Joachim Schwermer as part of his paper "Räumlich Anschauungen und Minima positive quadratischer Formen", see [54].

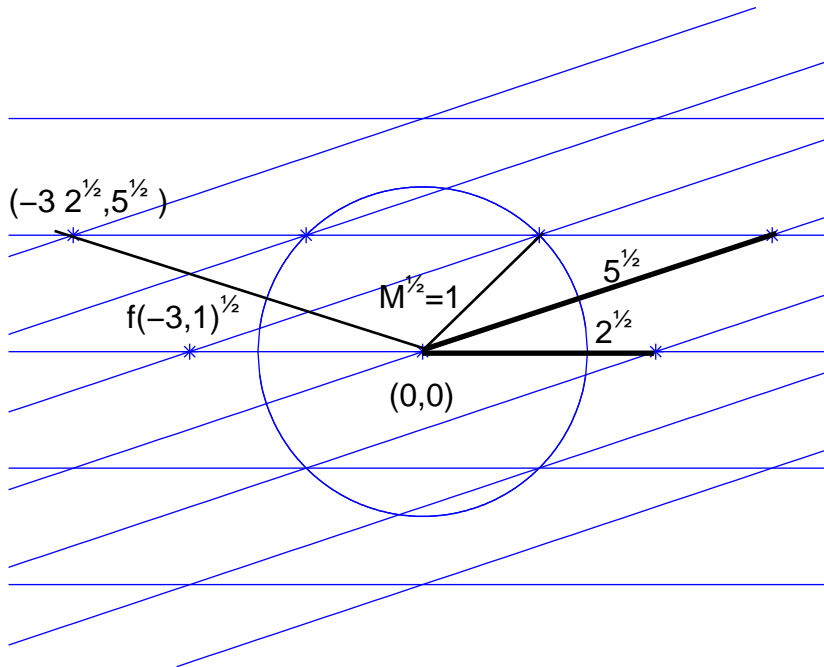


Figure 1. Lattice representing the quadratic form $f(x, y) = 2x^2 + 6xy + 5y^2$. The lattice points are the points where the two sets of parallel lines intersect. The smallest distance in the lattice is \sqrt{M} ([28], p. 65).

Since \sqrt{M} is the smallest distance in the lattice, M is the minimum of the quadratic form, and since the square of the volume of a standard parallelotope is equal to the determinant, D , of the form, the above inequality expresses that

$$M < kD^{\frac{1}{3}}.$$

Minkowski mentioned in the manuscript for his probationary lecture that his result could be generalised to forms in n -variables, but he did not publish a proof until 1891. In 1889, though, he referred to a proof for n -dimensions in a letter to Hilbert:

Now I have come much further in the theory of positive quadratic forms, for larger numbers of variables it becomes much different. Perhaps the following theorem (which I can prove on half a page) will interest you or Hurwitz: In a positive quadratic form with n (≥ 2) variables and determinant D one can always assign integer values to the variables such that the form becomes $< nD^{\frac{1}{n}}$. For the coefficient n Hermite had only $(\frac{4}{3})^{\frac{1}{2}(n-1)}$, which obviously, in general, is a much larger limit. ([44], p. 38).

In this phase, which can be identified as the first of three phases in the work that led Minkowski to introduce the concept of a convex body, Minkowski investigated positive definite quadratic forms with the purpose of solving the well known minimum problem - and he did so in a new way, by using geometrical interpretations and methods.

Phase 2 - investigations of the lattice and associated bodies.

The first shift, the shift into what can be identified as the second phase, can be seen from a summary of a talk Minkowski gave in Halle in 1891. Here Minkowski began to investigate his method, i.e. the lattice and the bodies circumscribing the lattice points, by which he deduced his upper bound for the minimum [36]. The talk was titled “Über Geometrie der Zahlen”, and he explained that he hereby meant geometrical investigations of the lattice and associated bodies as well as extensions into arbitrary dimensions. He introduced the lattice, not as a representation of a positive definite quadratic form, but as points with integer coordinates in a three-dimensional, orthogonal coordinate system. This signals a beginning detachment from positive definite quadratic forms. He made it quite clear, though, that he was still driven by his work in number theory. He considered the lattice and the associated bodies as a method that was useful in number theory, and he investigated them as such, since, as he explicitly pointed out:

Every statement about the grid [lattice] has of course a purely arithmetic core. ([36], p. 264)

Minkowski explained that he considered a very general category of bodies consisting of, as he phrased it himself:

all the bodies that have the origin as a middle point and whose boundary is nowhere concave.([36], p. 264)

He explained his investigations of the lattice in more details in the manuscript “Über eigenschaften von ganzen Zahlen, die durch räumliche Anschauungen erschlossen sind” which he wrote for the mathematical congress in Chicago in 1893. Here he pointed out that:

The deeper properties of the lattice are connected with a generalization of the concept of the length of a straight line by which only the theorem, that the sum of two of the sides in a triangle is never less than the third side, is maintained. [37]

In this phase, Minkowski explored the lattice and the associated bodies. They functioned as the objects about which new questions could be asked. His research into these objects led him to generalize, as he said in the quote above, the concept of the length of a straight line. He did so by introducing what he called the radial distance $S(ab)$ between two points a and b , and the corresponding

“Eichkörper” or gauge body, which consists of all points u for which $S(ou) \leq 1$. He explained that:

If moreover $S(ac) \leq S(ab) + S(bc)$ for arbitrary points a , b , and c the radial distance is called *einhellig*. Its “Eichkörper” then has the property that whenever two points u and v belong to the “Eichkörper” then the whole line segment uv will also belong to the “Eichkörper”. On the other hand every *nowhere concave body*, which has the origin as an inner point, is the “Eichkörper” of a certain “einhellig” radial distance function. ([37], p. 272-273)

Today we would call an “einhellig” distance function a metric that induces a norm if the radial distance function is also reciprocal, i.e. if $S(ab) = S(ba)$. In the above quote we recognize the “Eichkörper” as what we today would call the unit ball around the origin, and its property as that of convexity.

Minkowski then gave a proof for a theorem, which he had stated without proof in his talk from Halle in 1891, and which became known as Minkowski's lattice point theorem, namely that if the “Eichkörper” for an “einhellig” and reciprocal radial distance function has volume $\geq 2^3$, then the “Eichkörper” contains a lattice point in addition to the origin. Minkowski wrote about the theorem that:

The hereby gained theorem about nowhere concave bodies with middle point seems to belong to the most fruitful in the whole of number theory. ([37], p. 274)

His proof follows the line of arguments that he gave for the minimum problem. So, probably around 1891, Minkowski had realized that the essential property for the argument to work is that the bodies he used to circumscribe around the lattice points are - as he named them at that time - nowhere concave bodies with middle point.

In this phase Minkowski posed and answered new questions. He generalized the lattice, the associated bodies, and the lattice point theorem to n -dimensions, thereby creating a powerful tool, he could use to prove theorems in number theory. He introduced the general notions of nowhere concave bodies with - and without - middle point, radial distances, and “Eichkörper”. The relation between these concepts comes from the fact, as spelled out by Minkowski in the quote above, that every nowhere concave body with the origin as an inner point is the “Eichkörper” of some “einhellig” radial distance, and vice versa the “Eichkörper” corresponding to an “einhellig” radial distance form a nowhere concave body with the origin as an inner point. The usefulness of the property of convexity - or nowhere concavity - as Minkowski called it at that time, came out of his investigations of the lattice and the associated bodies.

This idea of using methods from geometry to solve problems in number theory initiated an extremely innovative period in Minkowski's career, where he

introduced the concept of a general convex set, and founded the new mathematical discipline that he named “*Geometrie der Zahlen*” (Geometry of Numbers), because, as he wrote in the advertisement for his 1896 book on the subject:

I have chosen the title *Geometry of Numbers* for this work because I reached the methods that gives the arithmetical theorems, by spatial intuition. Yet the presentation is throughout analytic which was necessary for the reason that I consider manifolds of arbitrary order right from the beginning ([42], p. v.)

Minkowski gave a precise mathematical definition of what he understood by a general nowhere concave body towards the end of his book *Geometrie der Zahlen*, and even though he still stressed the importance of nowhere concave bodies, or convex bodies as he soon began to call them, in number theory, he began to think of them as independent mathematical concepts in their own right, which brings us to phase 3.

Phase 3 - investigations of convex bodies for their own sake.

The shift of Minkowski’s focus into phase 3 is seen very clearly from four papers he published in the period 1897-1903, along with an unfinished paper that was published posthumously in his collected works [38], [39], [40], [41], [43]. In these papers Minkowski worked on convex bodies completely detached from quadratic forms and number theory. He treated different aspects of convex sets, and began the work that developed into the modern theory of convexity.

In his first paper solely devoted to the study of convex sets for their own sake, Minkowski gave the following definition:

A *convex body* is completely characterized by the properties that it is a closed set of points, has inner points, and that every straight line that takes up some of its inner points always has two points in common with its boundary. ([38], p. 103)

He then explained that investigating such bodies was mathematically interesting because of their applicability in areas like number theory, but also because, as he phrased it:

The theorems about convex bodies have a special appeal because they as a rule are valid for the whole category of objects without any exceptions. ([38], p. 103)

He also gave a more particular motivation, namely that the material he presented in the paper had came out of his attempts to prove a theorem that he had expected for a long time, namely that

A *convex body*, that is build up by a finite number of sheer *bodies with middle point* that only touch each other at the boundaries, has a middle point as well. ([38], p. 103)

The resemblance to his proof technique for the minimum problem suggests that Minkowski's move into the study of convex bodies for their own sake was a continuation of his previous work in phase 1 and phase 2 on number theory and geometry of numbers.

In his following papers on convex bodies, Minkowski gave a systematic treatment of such bodies in three dimensions. He developed the notions of the length of curves and the area of curved surfaces from the concept of volume, by which he was able:

To give a new and more rigorous proof of the theorem that among all convex bodies with equal volume the sphere has the smallest surface. [39]

He also introduced many of the now standard notions of distance function for convex bodies with the origin as an inner point, supporting hyper planes, separating hyper planes, mixed volumes etc.

Readers, who are familiar with the theory of convexity and its history might wonder, why Hermann Brunn has not been mentioned so far in this narrative. Historical sketches in introductory chapters of textbooks on convexity, intentionally or by chance, convey the impression that convexity was developed in a direct line from Karl Hermann Brunn (1862-1939) to Minkowski and so on. Apparently, Brunn was the first to engage in systematic studies of sets only characterised by the property of convexity. He did so in his inaugural thesis *Ueber Ovale und Eiflächen* from 1887, where he studied geometrical objects that he named ovals and egg-forms [2]. By an oval he understood a closed plane curve that has two and only two points in common with every intersecting straight line in the plane, and by a "volles [full] Oval" he understood an oval together with the inner points. The corresponding spatial objects, which he called egg-surfaces and egg-bodies, were introduced similarly. There are many differences between Brunn's and Minkowski's approach and motivation to the study of convexity - and I will not go into all of them here, interested readers are referred to [30]. Here it will only be mentioned that Brunn and Minkowski worked independently of each other. Brunn became aware of Minkowski's work probably around 1893 where the first announcement of Minkowski's book appeared. A year later he published a revision of parts of his thesis, apparently after Minkowski had read his thesis and pointed out some flaws. According to Brunn:

The occasion to return to this subject [the inaugural thesis] the revision of which for a long time has appeared ungrateful for the author is the knowledge of similar work by Mr. Minkowski in Bonn (soon in Könningsberg). By Teubner Minkowski has published a preannouncement of a book in print entitled *Geometrie der Zahlen* in which an unexpected and fruitful connection between number theory and the geometry of bodies whose boundaries are nowhere

concave is established and thereby also in analytical terms treats the theory of the latter. Thus also from other sides than a geometrical point of view a certain importance is attached to egg-forms and this has encouraged the author to supplement his doctoral thesis in the manner indicated above. ([3], p. 94)

3. From Logistic Problem Solving in the US. Air Force to Mathematical Programming - an Episode of Connections in Applied Mathematics

Mathematical programming is another mathematical discipline that emerged, took form and developed in the 20th century. The history of its emergence and development is probably as far from that of the theory of convexity as it can be. Mathematical programming came from applied mathematics, not from the ‘queen’ of pure mathematics as did Minkowski’s general convex bodies. It was not developed by a mathematician, who was at the centre of the mathematical universe, recognized as one of the most brilliant mathematicians of his time, as was the case with the theory of convexity. It started in the military context of the Second World War with a concrete logistic problem, and a group of people, including a young mathematician who still had to finish his Ph.D., who worked at the US. Air Forces. Its development was driven by a completely different set of issues. It was initiated and highly influenced by the historical circumstances of its time, by the needs of society. This story will show that also external driving forces, influences and connections are present in the development of mathematics.

Mathematical programming started in the military as a model called ‘programming in a linear structure’. Through the post war military-science-complex it got connected to game theory and moved into academic research institutions, Princeton University being one of them, where it became exposed to fundamental research. It turned into the theory of linear programming, sparked new interests and developments in the theories of linear inequalities and convexity, and was extended to non-linear programming. It became included in the toolbox of operations research, which was itself a new subject that rapidly got established as a scientific enterprise in the USA in the first decade after the war. A whole variety of programming problems were introduced, and through all these connections the mathematical theory of programming problems expanded and finally, with the founding in 1972 of the Mathematical Programming Society, became established as the mathematical discipline called mathematical programming.

This short outline prompts several interesting questions: How could an Air Force logistic problem turn into a mathematical research discipline? How did

the Air Force problem cross the boundary to mathematics proper? Why, and by whom, was it deemed important enough that mathematicians found it worth while exploring? What role did the military play, and what influence did it have for the emergence of mathematical programming as a mathematical research area?

The story that will be told on the following pages has been guided by these questions as well as by the different kinds of connections and networks that appeared crucial along the path to answers.

The Air Force Programming Problem.

During the Second World War the US mobilized scientists in great numbers. The mathematician George B. Dantzig was employed by the Air Forces to work on what they called “programming planning methods”, and to teach Air Force personnel how to calculate these programmes. An Air Force programme was a tool for handling huge logistic planning. Dantzig and his co-worker Marshall K. Wood, who was an expert on military programming procedures, characterised the programming problem in a paper presented in 1948 as follows:

Programming, or program planning, may be defined as the construction of a schedule of actions by means of which an economy, organisation, or other complex of activities may move from one defined state to another, or from a defined state toward some specifically defined objective. ([58], p.15)

This was a post war, or rather a post computer, definition. Initially, the possibility of searching for a programme fulfilling some kind of defined objective was not part of the problem. The assignment during the war was to construct a consistent programme:

The levels of various activities such as training, maintenance, supply, and combat had to be adjusted in such a manner as not to exceed the availability of various equipment items. Indeed, activities should be so carefully phased that the necessary amounts of these various equipment items were available when they were supposed to be available, so that the activity could take place. ([58], p.18)

One of the major difficulties was that it took too long to calculate a programme: “Even with the most careful scheduling, it took about seven months to complete the process.” ([13], p. 191)

After the war Dantzig went back to Berkeley University to complete his Ph.D., but already in 1946 he was back in the armed forces working until 1952 as mathematical advisor for the U.S.A.F. Headquarters. The original job assignment was to “develop some kind of analogue device which would accept, as input, equations of all types, basic data, and ground rules, and use these to generate as output a consistent Air Force plan.” ([6], p. 12).

However, it soon became clear that the computer would be realized, and that it probably could be used for the calculation of the Air Force programmes, so the work changed character. Project SCOOP (Scientific Computation of Optimum Programmes) was initiated and a project group was formed. The focus was on the construction and development of computers and the building of a mathematical model for the programming problem. The group reformulated the model in an axiomatic mathematical language, and mathematized the air force problem as follows:

the minimization of a linear form subject to linear equations and inequalities. ([5], p. 44)

which is identical with the formulation of a linear programming problem given in most textbooks today.

In December 1948 Dantzig and Wood pointed towards the problem of how to solve this model:

we seek to determine that program which will, in some sense, most nearly accomplish objectives without exceeding stated resource limitations. So far as is known, there is so far no satisfactory procedure for solution of the type of problem. ([4], p. 195)

That the task wasn't easy can be seen from the following description given by two of the protagonists Wood and Geisler:

These complexities [of the Air Force programming problem] have been spelled out to indicate a whole range of planning problems which, because of the present difficulties of computing alternative programs, receive little or no consideration. *So much time and effort is now devoted to working out the operational program that no attention can be given to the question whether there may not be some better program that is equally compatible with the given conditions.* It is perhaps too much to suppose that this difference between programs is as much as the difference between victory and defeat, but it is certainly a significant difference with respect to the tax dollars and the division of the total national product between military and civilian uses.

Consideration of the practical advantages to be gained by comparative programming, and particularly by the selection of best programs, leads to a requirement for a technique for handling all program elements simultaneously and for introducing the maximization process directly into the computation of programs. Such a technique is now in prospect. ([13], p. 194 (Italic in the origin))

Dantzig was advised to consult with John von Neumann at the Institute for Advanced Study at Princeton to discuss the problem of solving the model.

The establishment of such a meeting between von Neumann and Dantzig was a consequence of the organisation and mobilization of scientists in the USA during the war. The mobilization of science did not only happen through the involvement of scientists who, like Dantzig, were hired directly by the armed forces. That was only one part of the scientific mobilization; another part was the mobilization of civilian scientists, which was an extraordinary feature of the scientific mobilization. These scientists were not under military command. They remained civilians, and they stayed at their universities and worked on problems requested by the armed forces through contracts with the OSRD (Office of Scientific Research and Development) [59], [48].

The American Mathematical Society and Mathematical Association of America took initiatives right from the beginning to get mathematicians involved in the war effort [47]. In 1940 they “appointed a committee known as the War Preparedness Committee, to prepare the two societies to be useful to our nation in time of war” ([46], p. 293-294). But the leaders of OSRD were slow in bringing the mathematicians on stage. There was no joint coordination of their services until late 1942, when the Applied Mathematics Panel (AMP) was established with Warren Weaver as leader. The reason for this late involvement seems to be rooted in the dichotomy between pure and applied mathematics. The leaders of OSRD had problems seeing how practical problems could be treated in the realm of abstract mathematics. They regarded the Committee as being too pure. But even though Weaver had problems with some of the more egocentric mathematicians, the overall evaluation of the mathematicians’ contribution to the war effort was positive. Especially the Panel’s education of mathematicians to operations research (OR) personnel is often highlighted as a success.

The scientific mobilisation opened new communication channels. Mathematicians held consulting jobs for the military and they got involved with practical problem solving. John von Neumann was one of these mathematicians. He was in many committees and held a lot of consulting jobs both during and after the war.⁵ The scientific successes during the war created an atmosphere of optimism and belief in science that extended well into the post war period. As a result of the organisation of civilian scientists during the war and in the post war period, mathematicians working in the military became connected with mathematicians in academia.

Dantzig’s contact with von Neumann to be advised on how to solve the linear programming problem was such a connection. The first meeting between the two took place in October 1947 [5], [6]. According to Dantzig, von Neumann immediately realized that the Air Force programming problem was connected to game theory and thereby also to the theory of linear inequalities and convexity. Hence, von Neumann provided the air force programming problem with a

⁵For an incomplete list of von Neumann’s involvement with the military during and after the war see ([55], p. 42).

mathematical foundation, and it became connected with mathematical research [21],[22].

This connection between game theory and the Air Force problem was not at all obvious, and von Neumann was probably the only one, who could have made this connection at that time. He had just completed the book *Game Theory and Economic Behavior* co-authored with the Austrian economist Oskar Morgenstern, where they presented the theory of two-person zero-sum games and its main result, the minimax theorem about existence of optimal solutions, in what they themselves characterised as the “mathematico-geometrical theory of linearity and convexity” ([57], p. 128).

From problem to theory - the significance of military funding. The Air Force problem moved into academic research in mathematics through a research project that was set up at the mathematics department at Princeton University in the summer of 1948. The aims of the project were to study the connections between the linear programming problem, as the mathematical model of the Air Force programming problem was now called, and game theory, as well as the underlying mathematical theories of linear inequalities and convexity. The project was financed by ONR (Office of Naval Research).

Albert W. Tucker, who was a mathematics professor at Princeton, was appointed principal investigator of the project, and he hired two graduate students, David Gale and Harold W. Kuhn, to work with him. Dantzig had met Tucker at one of his meetings with von Neumann. Tucker showed an interest in the problem, and that was probably why he was asked to undertake the project. While it was kind of a coincidence that Tucker got involved with the project, the project itself did not come into existence by a coincidence. It was the result of a well thought plan for how the military, through ONR, could effectively promote and to some extent control directions for research at the universities.

Mina Rees, who had been the technical aid to Warren Weaver at the Applied Mathematics Panel during the war, became the leader of ONR’s mathematics programme. In 1977 she wrote the following about her memories of this summer project, of how it came in to being on the request of the ONR, and how it initiated the establishment of a separate Logistics Branch in ONR with its own research programme:

[...] when, in the late 1940’s the staff of our office became aware that some mathematical results obtained by George Dantzig, [...] could be used by the Navy to reduce the burdensome costs of their logistics operations, the possibilities were pointed out to the Deputy Chief of Naval Operations for Logistics. His enthusiasm for the possibilities presented by these results was so great that he called together all those senior officers who had anything to do with logistics, as well as their civilian counterparts, to hear what we always referred to as a presentation. The outcome of this meeting was the establishment

in the Office of Naval Research of a separate Logistics Branch with a separate research program.

This has proved to be a most successful activity of the Mathematics Division of ONR, both in its usefulness to the Navy, and in its impact on industry and the universities. [49]

After Dantzig's first consultation with von Neumann the latter wrote a note "Discussion of a Maximum Problem", in which he rewrote a linear programming problem of maximizing a linear form subject to linear inequality constraints into a problem of solving a system of linear inequalities [56]. This note was circulated privately, and together with von Neumann's and Morgenstern's book on game theory it furnished the point of departure for the work done by Tucker and his co-workers at Princeton that summer.

The first results of their work are reported in the paper "Linear Programming and the Theory of Games" which they presented the following summer at the first conference on linear programming [11]. That this new area of research was a true child of the military-university cooperation is reflected in the list of participants and their sponsors. The conference itself and the research done by a majority of the participants were supported by the military. From the organisation of the proceedings and Koopmans introduction to it, it follows that linear programming was mainly perceived as an economic theory at the time, and the theory of convexity as a tool "relatively new to economics" ([32], p. 10).

Tucker, Gale, and Kuhn proved the duality theorem⁶ and the existence theorems - not for the 'basic' linear programming problem, but for a generalized 'matrix' problem that have the ordinary 'basic' problem as a special case. They also gave a new proof of von Neumann's minimax theorem in game theory, and showed that the optimal strategies for a two-person zero-sum game constitute a solution to the corresponding 'basic' linear programming problem and its dual. Here we see the effects of the military-university complex on research in mathematics. Kuhn, Gale, and Tucker treated the linear programming problem as a mathematical research field. Instead of just working on the 'basic' linear programming problem for the purpose of practical problem solving, they immediately generalized it without any consideration of the applicability of the generalized 'matrix' version of the problem. This approach is typical for basic research in mathematics as it is conducted in academia.

The duality result is an interesting mathematical result, and it caught the further interest of Tucker, who took it to the next level [21], [26]. He asked Gale

⁶To a linear programming problem another linear programming problem can be formulated using the same data such that if the original (called the primal) problem is a minimum problem then the other (called the dual) problem will be a maximum problem. The duality theorem states that the primal problem has a finite optimal solution if and only if the dual problem has a finite optimal solution, in which case the minimum of the primal problem is equal to the maximum of the dual problem.

and Kuhn whether they wanted to continue working on the project, to see if they could extend the duality result from the linear to the quadratic case [33]. Gale declined, but Kuhn went along with Tucker, and in 1950 they presented their work at the Second Berkeley Symposium on Mathematical Statistics and Probability. It was published in the proceedings from the meeting with the title “Nonlinear Programming” [34]. As the title indicates, they changed the focus from the quadratic to the general nonlinear case. They did not succeed in proving a duality result for non-linear programming, but they did prove what immediately became known as the Kuhn-Tucker conditions for the existence of a solution to a non-linear programming problem. Contrary to the emergence of linear programming, there was no direct external applicational motive for Kuhn’s and Tucker’s initial inquiry into the theory of nonlinear programming.

Kuhn’s and Tucker’s paper initiated the new research field of nonlinear programming. Their theorem is considered to be an important result, and a lot of significance was (and still is) attached to it. Later it turned out that two other mathematicians, William Karush and Fritz John, had proven the “same” theorem as Kuhn and Tucker in 1939 and 1948, respectively, and both times it went almost unnoticed in the mathematical community. Karush proved his result in his master’s thesis from the mathematics department at Chicago University. His thesis was a finite dimensional investigation of questions related to the calculus of variation, and it can be seen as a continuation of a work done by Gilbert Ames Bliss (1876-1951) on a similar finite dimensional problem, where the constraints were given as equalities instead of inequalities [1]. Within the context of the “Chicago School” of the calculus of variation, Karush’s theorem of existence of solutions to finite-dimensional versions of the interesting problems in the calculus of variation was simply not considered to be important.⁷

Fritz John’s result was published in 1948 in a paper with the title “Extremum Problems with Inequalities as Subsidiary Conditions” in a collection of papers put together to honour Courant at his 60’s birthday [18]. John had earlier tried to publish the paper in *Duke Mathematical Journal*, but they had turned it down. John’s paper is structured in two parts: a theoretical one in which he proved a result similar to the one by Kuhn and Tucker (though without the constraint qualification), and a second part with two geometrical applications to convex sets. Even though Fritz John’s structure of the paper suggests that the theoretical part is the important one, there are several circumstances that indicate that the geometrical applications were the main focus of attention.⁸ John himself has expressed that his inspiration and motivation behind the work should be found in the applications, both of which belong to the theory of convexity. Again, the theorem in itself was not considered important.

Nevertheless, just two years after Fritz John’s paper appeared in Courant’s birthday publication, Kuhn and Tucker wrote their names into the “Hall of

⁷For further details, see [20].

⁸For further details, see [20].

Fame” of operations research with their Kuhn-Tucker theorem, and as we have seen, the explanation for that cannot be found in the result alone, considered in isolation. The explanation lies in the local context of time and place - in the military-university complex of science support in the USA following the Second World War.

Linear and non-linear programming was not only supported by the logistic branch at ONR, the new field of mathematical programming also benefited from the connection to game theory. Eventhough von Neumann proved the main theorem, the minimax theorem, for two-person zero-sum games in 1928,⁹ the really significant development took place during the war, where he and Morgenstern wrote their huge book *Theory of Games and Economic Behavior* [57]. The book was written with the economists in mind [35], not the war, but given the time and place, and the concept of optimal strategies for winning a game, which fitted perfectly with the war context, and given von Neumann’s multiple connections, reputation, and influence within the military-science complex, game theory became embedded in the military context [45]. The RAND Corporation became the most important one. In the first decade after the war, RAND was the centre for mathematical research in game theory. The group at RAND held lengthy summer sessions in game theory and collaborated with Tucker’s logistic project in Princeton which, besides Tucker and Kuhn, also included people like John Nash and Martin Shubik. Both places - the ONR project and RAND - were staffed with mathematicians, who were brought up in the culture of pure mathematical research in academia. They took that training with them, and what they did at RAND and in Tucker’s ONR project, they themselves described as they were doing research in pure mathematics sponsored by the military. Game theory and linear programming with its extensions into non-linear and convex programming were on their research agenda.

We can now give answers to the questions of how the Air Force logistic problem crossed the boundary to academic research in mathematics, and what role the military played in the emergence of mathematical programming. The answers to these two questions are connected, because this move into university mathematical research can only be understood within the war and the post war context of military supported science in a university culture of academic research. The military establishment could see the usefulness of linear programming and be convinced of a future use of game theory. Given the success of the collaboration with scientists during the war there was a belief that fundamental science was the optimal foundation for war fare. In order to be strong militarily the country needed to be strong scientifically ([53], p. 11). By setting up contract-based projects like Tucker’s where the scientists were not constrained by practical problem solving, but were free to explore what ever they found interesting as long as it was just remotely connected to the topics of linear programming and game theory, the soil was prepared for fundamental

⁹For further details on von Neumann’s conception of the minimx theorem, see [22].

mathematical research. More specifically, the duality result for linear programming and its connection to the minimax theorem for two-person zero-sum games was an important source of inspiration for Tucker. It was interesting from a pure mathematical point of view. It opened the model of the Air Force problem and made it an interesting research area in itself. Tucker and his group took the results to the next level and launched the theory of non-linear programming.

The inquires into the foundations of these new mathematical research areas created a framework that embedded game theory and the various programming problems in a mixture of pure and applied mathematics. Another important factor of a more sociological nature, that probably should also be taken into account, is the excitement and promising career opportunities that lies in exploring new areas of research where results come fast, and where one might have the possibility of being counted as one the founding “fathers” or “mothers”.

4. Convexity Meets Mathematical Programming at Princeton - a Mutual Beneficial Relationship

Kuhn and Tucker did not succeed in deriving a duality result for nonlinear programming, but in their work with the Kuhn-Tucker-theorem, they realised that full equivalence between solutions to nonlinear programming problems and to the corresponding saddle value problems for the associated Lagrangian function could be obtained under certain convexity assumptions.¹⁰

It must therefore have been very exciting for Kuhn and Tucker when they became aware that Werner Fenchel from Copenhagen University, who was the leading expert on convexity at the time, was visiting the Institute for Advanced Study in Princeton as part of a sabbatical year in the USA in 1950/51. Tucker invited Fenchel to give a series of lectures on the theory of convexity at the mathematics department at Princeton University within his ONR project.

Through this connection with Tucker and his group at Princeton, Fenchel became aware of the mathematical problems of nonlinear programming. These problems inspired him to further developments in the theory of convexity, and he succeeded in deriving the first duality theorem for nonlinear programming, the so-called Fenchel-duality. Fenchel’s lectures were published by Tucker’s ONR project, and in the acknowledgement Fenchel wrote:

The author [Fenchel] wishes to express his gratitude to Professor A. W. Tucker for giving him this opportunity to write this report and for calling his attention to the problems dealt with in the final sections (pp 105-137). ([9], Acknowledgement)

¹⁰For further details, see [26].

In the final sections Fenchel treated what he called a generalized programming problem, and it is also here we find Fenchel's duality result for nonlinear programming - a duality result that he based on the notion of conjugate convex functions. In 1949, that is a couple of years earlier, Fenchel had published a small paper on convex functions, where he had introduced the notion of conjugate convex functions as a consequence of some investigations of the underlying mathematical structure beneath inequalities in analysis [8]. Fenchel had noted that these inequalities often can be interpreted as consequences of the convexity of some functions. He proved that to every convex function, defined on a certain convex set, and fulfilling some continuity conditions, there corresponds a convex set, and a convex function defined on it, with the same properties as the original convex function, such that a certain inequality is fulfilled. Fenchel called the two convex functions for each others conjugate [26], [25].

In the Princeton notes Fenchel argued that, similarly to what he had shown in his 1949 paper, it is possible to find the conjugate to a concave function. By considering a closed convex function defined on a convex set, and a closed concave function defined on another convex set, he formulated a maximum problem of the difference between the concave and the convex functions as well as a minimum problem of the difference between the corresponding conjugate functions. Fenchel then proved the first duality result in non-linear programming, namely that under certain conditions, the supremum of the difference between the concave and the convex function is equal to the infimum of the difference between the corresponding conjugate functions.

Fenchel was motivated by problems and connections in pure mathematics, and his lecture notes became highly influential for further developments within the theory of convexity, especially through the work of R. T. Rockafellar [50].

Here we see that not only did the applied field of mathematical programming develop due to already existing theories of pure mathematics, also pure mathematics benefitted and developed due to inspiration from mathematical programming - suggesting that the distinction between 'pure' and 'applied' mathematics is much more complex and blurred than these two notions indicate.

5. Discussion and Conclusion

We have traced the emergence and development of two mathematical theories: one that originated at the turn of the 20th century, and one that materialized in the middle of the century. They are in a certain sense representative for 20th century mathematics which, on the one hand, is characterized by a move into an autonomous enterprise - a 'modernism' movement in early twentieth century mathematics; and on the other hand, is characterized by a migration of mathematics into areas such as the life sciences, medicine, and the social sciences to name a few, with the purpose of developing mathematics into a tool in those areas inspired by the effectiveness of mathematics in physics.

It is possible to distinguish modernistic features in Minkowski's development of the concept of a convex body and the beginning of a general theory of convexity. He generalized the ideas, he conceived through geometrical intuition, to n -dimensional space for them to be useful in the study of positive definite quadratic forms in n -variables. He also axiomatized the notion of the "length of a straight line". But, if we also look at further developments of the theory of convexity, analysed with respect to connections and relationships that governed these developments, a much messier picture becomes visible, where 'pure' and 'applied' mathematics interact and develop in mutual beneficial relationships. This becomes visible when we take the development of mathematical programming into account and look at how, why, and through which connections, this theory took form and became an established discipline of 'applied' mathematics. As we have seen, this did not happen only by using already existent mathematics and applying it to "programming problems", it also happened by developing and proving new mathematical theorems, and by providing existing mathematical theories in 'pure' mathematics with new mathematical problems and challenges.

Regarding Minkowski's introduction of the concept of a general convex body into mathematics, we have seen that this concept gradually crystallized into a stable mathematical object through connections between different mathematical disciplines - number theory and geometry - a connection that 'forced' Minkowski to work in and to develop his geometrical method for n -dimensional space. As a result of this 'mixing' or crossing of disciplinary boundaries, Minkowski developed a new tool that, on the one hand, was very effective for number theoretical investigations and, on the other hand, gave rise to new questions and problems, thereby creating no less than two new areas for mathematical research: geometry of numbers and the theory of convexity.

When we switch to mathematical programming, other kinds of connections and relationships emerge that can be seen to have initiated and guided developments of mathematics. First there are, again, the mathematical connections. The connection between the programming problem and game theory provided the programming problem with a mathematical foundation in the theories of linear inequalities and convexity. These connections made it possible to "look" for a duality theorem in linear programming. The question of a dual programming problem could be asked, because solutions to two-person zero-sum games come in pairs - one optimal strategy for each player. Realizing the connections between linear programming and two-person zero-sum games initiated inquiries into how this 'duality' could be understood in the context of linear programming, thereby opening the Air Force problem for mathematical research. As a consequence, the Air Force programming problem, which originated directly from an urgent need in a time of war to solve a concrete, practical logistic problem in the Air Force, was followed by a development into nonlinear programming which then eventually was divided into many kinds of programming problems, the theories of which became Mathematical Programming. The generalization

into nonlinear programming did not, as we have seen, originate in practical problem solving. This development was not motivated by an urgent need to solve an existing problem here and now, rather it followed the lines of basic research in pure mathematics. It was driven by a longing for understanding and generalization within the realm of abstract mathematics.

The possibility for this development to take place can be found in another kind of connection, a sociological relationship between the US military and scientists. The military-university complex that had been developed in the USA as part of the war effort created new channels of communication and new funding possibilities for research in academia. The connection between the Air Force programming problem and game theory was discovered through a personal connection between Dantzig and von Neumann, and the further developments were made possible through military funding of research in academia. In both situations, the military-university complex served as the mediating link.

Finally, the organisation and sociology of academic science created the opportunity for establishing the relationship between Werner Fenchel and Tucker's group at Princeton University. The personal connection between Fenchel and Tucker appeared because of the social structure of sabbatical travels to visit other/foreign research institutions and scientists. The interaction between Fenchel's research in the theory of convexity, and mathematical programming as it was developing in Tucker's group at Princeton, became possible because the programming problem - again due to the military-science complex - crossed the boundary to academic research in university institutions.

The two episodes in the history of mathematics seem to have developed due to very different circumstances and through very different kind of connections. On a closer look, however, they also share some very particular traits that tell us something about the significance of the context for the importance attached to mathematical results, for their 'fruitfulness' or capabilities to initiate and drive new developments in mathematics.

In the first example, a theory of convexity in itself, as it appeared in Brunn's work, did not have the potential to success, but in the right context it did. Minkowski's approach in what has been described as the two first phases highlighted the usefulness of convex bodies in interaction with other mathematical theories - in connection with number theory, but just as importantly, with the beginning general functional analysis with its definition of metric and normed spaces. Thereby, Minkowski's convex bodies became centrally placed in the mathematical universe.

In the second example we saw no less than two instances, illustrating that whether a mathematical result becomes acknowledged as important or not depend on the mathematical context in which it is derived. What doesn't appear to be an interesting and important theorem in one mathematical setting at a particular place at a particular time, might be evaluated quite differently in another mathematical setting at another place in another time. The case with the Kuhn-Tucker theorem is a clear example of this. Within the mathematical

context of the calculus of variation at the Chicago School in the 1930s, Karush' result neither solved an important open problem nor opened for new interesting paths and research questions - and likewise with the result proved by Fritz John. None of these two appearances of what later became known as the Kuhn-Tucker theorem caused any new big developments in mathematics, but when Kuhn and Tucker proved their theorem, the situation was completely different. Their result answered a fundamental question in the new research field of mathematical programming and spurred further interest in this area that, due to the circumstances in the society at the time, was under rapid development.

References

- [1] A.G. Bliss, Normality and Abnormality in the Calculus of Variations, *Transactions of the American Mathematical Society*, 43, 1938, 365–376.
- [2] H. Brunn, *Ueber Ovale und Eiflächen*, Inaugural-Dissertation, Munich, Akademische Buchdruckerei von F. Straub, 1887.
- [3] H. Brunn, Referat über eine Arbeit: Exacte Grundlagen für eine Theorie der Ovale, *Sitzungsberichte der Mathematik und Physik*, XXIV, 1894, 93–111.
- [4] G.B. Dantzig, and M.K. Wood, Programming of Interdependent Activities, I General Discussion, *Econometrica*, 17, 1949, 193–199.
- [5] G.B. Dantzig, Reminiscences about the Origins of Linear Programming, *Operations Research Letters*, 1, 1982, 43–48.
- [6] G.B. Dantzig, Impact of Linear Programming on Computer Development, *OR/MS Today*, 1988, 12–17.
- [7] G.L. Dirichlet, Über die Reduction der positiven quadratischen Formen mit drei unbestimmten ganzen Zahlen, *Journal für die reine und angewandte Mathematik*, 40, 1850, 209–227.
- [8] W. Fenchel, On Conjugate Convex Functions, *Canadian Journal of Mathematics*, 1, 1949, 73–77.
- [9] W. Fenchel, *Convex Cones, Sets, and Functions*, Lecture Notes, Department of Mathematics, Princeton University, 1953.
- [10] W. Fenchel, Convexity Through the Ages, in P.M. Gruber and J.M. Wills (ed.) *Convexity and Its Applications*, Birkhäuser VerlagBasel, Bosten, Stuttgart, 1983, 121–130.
- [11] D. Gale, H.W. Kuhn, and W. Tucker, Linear Programming and the Theory of Games, in T.C. Koopmans (ed.) *Activity Analysis of Production and Allocation*, Cowles Commission Monograph, 13, New York, Wiley, 1951, 317–329.
- [12] C.F. Gauss, *Collected Works II*, Göttingen, 1863.
- [13] M.A. Geisler, and M.K. Wood, Development of dynamic models for program planning, in T.C. Koopmans (ed.) *Activity Analysis of Production and Allocation*, Cowles Commission Monograph, 13, New York, Wiley, 1951, 189–215.
- [14] J.R. Goldman, *The Queen of Mathematics*, Wellesley, Massachusetts, A K Peters, Ltd., 1998.

-
- [15] C. Goldstein, N. Schappacher, and J. Schwermer, (eds.) *The Shaping of Arithmetic after C.F. Gauss's Disquisitiones Arithmeticae*, Berlin-Heidelberg-New York, Springer, 2007.
- [16] J.J. Gray, *Plato's ghost: the modernist transformation of mathematics*, Princeton Univeristy Press, 2008.
- [17] C. Hermite, Extraits de letters d. M. Ch. Hermite a M. Jacobi sur diffrents objects de la thorie des nombres, *Journal für die reine und angewandte Mathematik*, 40, 1850, 261–315.
- [18] F. John, Extremum Problems with Inequalities as Subsidiary Conditions, in *Studies and Essays, Presented to R. Courant on his 60th Birthday January 8, 1948*. New York, Interscience, 1948, 187–204.
- [19] T.H. Kjeldsen, The Early History of the Moment Problem, *Historia Mathematica*, 20, 1993, 19–44.
- [20] T.H. Kjeldsen, A Contextualized Historical Analysis of the Kuhn-Tucker Theorem in Nonlinear Programming: The Impact of World War II, *Historia Mathematica*, 27, 2000, 331–361.
- [21] T.H. Kjeldsen, The Emergence of Nonlinear Programming: Interactions between Practical Mathematics and Mathematics Proper, *The Mathematical Intelligencer*, 22, 2000, 50–54.
- [22] T.H. Kjeldsen, John von Neumann's Conception of the Minimax Theorem: A Journey Through Different Mathematical Contexts, *Archive for History of Exact Sciences*, 56, 2001, 39–68.
- [23] T.H. Kjeldsen, Different Motivations and Goals in the Historical Development of the Theory of Systems of Linear Inequalities, *Archive for History of Exact Sciences*, 56, 2002, 469–538.
- [24] T.H. Kjeldsen, New Mathematical Disciplines and Research in the Wake of World War II, in B. Booss-Bavnbek, and J. Hyrup (eds.) *Mathematics and War*, Birkhäuser Verlag, Basel-Boston-Berlin, 2003, 126–152.
- [25] T.H. Kjeldsen, Fenchel's Duality Theorem, (in Danish), *Matilde - Newsletter of the Danish Mathematical Society*, 15, 2003, 14–17.
- [26] T.H. Kjeldsen, The Development of Nonlinear Programming in Post War USA: Origin, Motivation, and Expansion, in H. B. Andersen, F. V. Christiansen, K. F. Jrgensen, and V. Hendricks (eds) *The Way Through Science and Philosophy: Essays in Honour of Stig Andur Pedersen*, College Publications, London, 2006, 31–50.
- [27] T.H. Kjeldsen, Albert W. Tucker, in N. Koertge (ed.) *The New Dictionary of Scientific Biographies*, 7, Charles Scribner's Sons, Detroit, 2008, 80–82.
- [28] T.H. Kjeldsen, From Measuring Tool to Geometrical Object: Minkowski's Development of the Concept of Convex Bodies, *Archive for History of Exact Sciences*, 62, 2008, 59–89.
- [29] T.H. Kjeldsen, Operations Research and Mathematical Programming: From War to Academia - A Joint Venture, in V. Lundsgaard Hansen, and J. Gray (eds.) *History of Mathematics in Encyclopedia of Life Support Systems (EOLSS)*. Developed under the Auspices of the UNESCO, Eolss Publishers, Oxford, UK, 2008, 20 pp. [<http://www.eolss.net>]

- [30] T.H. Kjeldsen, Egg-forms and Measure Bodies: Different Mathematical Practices in the Early History of the Development of the Modern Theory of Convexity, *Science in Context*, 22, 2009, 85–113.
- [31] T.H. Kjeldsen, Abstraction and application: new contexts, interpretations in twentieth-century mathematics, in E. Robson, and J. Stedall (eds.) *The Oxford Handbook of the History of Mathematics*, New York, Oxford University Press, 2009, 755–778.
- [32] T.C. Koopmans (ed.) *Activity Analysis of Production and Allocation*, Cowles Commission Monograph, 13, New York, Wiley, 1951.
- [33] H.W. Kuhn, Nonlinear Programming: A Historical View, *SIAM-AMS Proceedings*, 9, 1976, 1–26.
- [34] H.W. Kuhn, and A.W. Tucker, Nonlinear Programming, in J. Neyman (ed.) *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, Berkeley, 1950, 481–492.
- [35] R.J. Leonard, Creating a Context for Game Theory, in E. Roy Weintraub (ed.) *Towards a History of Game Theory*, Durham and London, Duke University Press, 1992, 29–76.
- [36] H. Minkowski, Über Geometrie der Zahlen, *Gesammelte Abhandlungen*, I, Leipzig, Berlin, B. G. Teubner, 1911, 264–265.
- [37] H. Minkowski, Über Eigenschaften von ganzen Zahlen, die durch räumliche Anschauung erschlossen sind, *Gesammelte Abhandlungen*, I, Leipzig, Berlin, B. G. Teubner, 1911, 271–277.
- [38] H. Minkowski, Allgemeine Lehrsätze über die konvexen Polyeder, *Gesammelte Abhandlungen*, II, Leipzig, Berlin, B. G. Teubner, 1911, 103–121.
- [39] H. Minkowski, Über die Begriffe Länge, Oberfläche und Volumen, *Gesammelte Abhandlungen*, II, Leipzig, Berlin, B. G. Teubner, 1911, 122–127.
- [40] H. Minkowski, Über die geschlossenen konvexen Flächen, *Gesammelte Abhandlungen*, II, Leipzig, Berlin, B. G. Teubner, 1911, 128–130.
- [41] H. Minkowski, Volumen und Oberfläche, *Gesammelte Abhandlungen*, II, Leipzig, Berlin, B. G. Teubner, 1911, 230–276.
- [42] H. Minkowski, *Geometrie der Zahlen*, Leipzig: B. G. Teubner, 1910.
- [43] H. Minkowski, *Gesammelte Abhandlungen*, Leipzig, Berlin, B. G. Teubner, 1911.
- [44] H. Minkowski, *Briefe an David Hilbert*, Berlin-Heidelberg-New York, L. Rüdenberg, H. Zassenhaus, 1973.
- [45] P. Mirowski, When Games Grow Deadly Serious: The Military Influence on the Evolution of Game Theory, in D. G. Goodwin (ed.) *Economics and National Security*, Annual Supplement to Volume 23, History of Political Economy, Durham and London, Duke University Press, 1991, 227–255.
- [46] M. Morse, Mathematics in the Defense Program, *American Mathematical Monthly*, 48, 1941, 293–302.
- [47] L. Owens, Mathematicians at War: Warren Weaver and the Applied Mathematics Panel, 1942–1945, in D. E. Rowe, and J. McCleary (eds.) *The History of Modern Mathematics, vol. II: Institutions and Applications*, San Diego, Academic Press, Inc., 1989, 287–305.

-
- [48] E. Rau, The Adoption of Operations Research in the United States during World War II, in A. C. Hughes, and T. P. Hughes (eds.) *Systems, Experts, and Computers*, Dibner Series, Cambridge MA, MIT Press, 2000.
- [49] M.S. Rees, Mathematics and the Government: The Post-War Years as Augury of the Future, in D. Tarwater (ed.) *The Bicentennial Tribute to American Mathematics, 1776–1976*, The American Association of America, Buffalo, NY, 1977, 101–116.
- [50] R.T. Rockafellar, *Convex Analysis*, Princeton, New Jersey, Princeton University Press, 1970.
- [51] W. Scharlau, A Historical Introduction to the Theory of Integral Quadratic Forms, in G. Crzech (ed.) *Conference on Quadratic Forms*, Queen's University, Kingston, Ontario, Canada, 1977.
- [52] W. Scharlau, and H. Opolka, *From Fermat to Minkowski*, New York, Springer Verlag, 1985.
- [53] S.S. Schweber, The Mutual Embrace of Science and the Military: ONR and the Growth of Physics in the United States after World War II, in E. Mendelsohn, M. R. Smith, and P. Weingart (eds.) *Science, Technology and the Military*, Dordrecht, The Netherlands, Kluwer Academic Publishers, 1988, 3–45.
- [54] J. Schwermer, Räumliche Anschauung und Minima positive definiten quadratischer Formen, *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 93, 1991, 49–105.
- [55] S. Ulam, John von Neumann, 1903–1957, *Bulletin of the American Mathematical Society*, 64, 1958, 1–49.
- [56] J. von Neumann, Discussion of a Maximum Problem, in A. H. Taub (ed.) *John von Neumann Collected Works*, 6, Oxford, Pergamon Press, 1963, 89–95.
- [57] J. von Neumann, and O. Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, Princeton, 1944.
- [58] M.K. Wood, and G.B. Dantzig, The Programming of Interdependent Activities: General Discussion, in T.C. Koopmans (ed.) *Activity Analysis of Production and Allocation*, Cowles Commission Monograph, 13, New York, Wiley, 1951, 15–18.
- [59] P.G. Zachary, *Endless Frontier: Vannevar Bush, Engineer of the American Century*. New York, The Free Press, 1997.

Rewriting Points

Norbert Schappacher*

Abstract

A few episodes from the history of mathematics of the 19th and 20th century are presented in a loose sequence in order to illustrate problems and approaches of the history of mathematics. Most of the examples discussed have to do with some version of the mathematical notion of point. The Dedekind-Weber theory of points on a Riemann surface is discussed as well as Hermann Weyl's successive constructions of the continuum, and the rewriting of Algebraic Geometry between 1925 and 1950. A recurring theme is the rewriting of traditional mathematics, where 'rewriting' is used in a colloquial, non-terminological sense the meaning of which is illustrated by the examples.

Mathematics Subject Classification (2010). Primary 01A55, 01A60; Secondary 03-03, 11-03, 12-02, 14-03.

Keywords. History of mathematics, abstract Riemann surface, Intuitionism, Foundations of Algebraic Geometry

1. Rewriting History

History is all about change and difference; philosophy rethinks things; and mathematics moves on by constantly reinventing its own history.

In a series of lectures on the classification of algebraic threefolds delivered 15 years ago at an instructional meeting in Ankara, Miles Reid announced a new method he was about to present by saying: "In order to go further, we have to rewrite history." Such "rewritings of history" occur time and again, and on varying scales, whenever mathematicians get down to work. The astronomical treatises of the *Siddhantas* for example do not focus on the chord associated to a central angle in the circle, but work with relations between the half chord and its associated angle. In this way, they introduced and computed the sine and other trigonometric quantities, and also coined the terminology of *jya*, *kojya*, etc. which, via confused rewritings in Arabic, produced our

*Norbert Schappacher, IRMA, 7 rue René Descartes, 67084 Strasbourg cedex, France.
E-mail: schappacher@math.unistra.fr.

pseudo-Latin expressions *sine* and *cosine*. This seemingly small step of renormalization brought about a corresponding rewriting, a reorganization of the traditional tables handed down from the Babylonians and from Ptolemy. Another well-known rewriting in the same domain occurred in 18th century Europe when the complex exponential function was brought in to reorder the plethora of trigonometric formulae.

Once adopted for further work, such rewritings tend to stick. If you are an algebraic geometer used to working with cohomology theories, you will find it difficult to imagine how your predecessors have dealt with things without seeing cup products, vanishing $H^i(X, \dots)$, etc. mapping out the geometry. One might think that this is purely psychological or superficial, like asking silly questions such as: How could former generations survive without cellular phones? After all, at least *in principle*, all information carried by a cohomology group could be spelled out in non-cohomological, geometric terms. But in the development of science in general, and particularly in mathematics, no technological advance ever leaves the world intact. Every rewriting of a mathematical theory recreates both its objects and the ways to handle them.

In the example of cohomological methods in algebraic geometry this is highlighted by the very beginning of W.V.D. Hodge's international career with his 1930 paper [46], where he proved that a nonzero holomorphic n -form—an " n -ple integral of the first kind" in Hodge's terminology—on a complex n -dimensional algebraic variety cannot have all its periods equal to zero, thus answering a question posed by Francesco Severi for algebraic surfaces ($n = 2$). Atiyah in [2], p. 104, tells the story how Solomon Lefschetz, whose very methods Hodge was generalizing in his proof, would not get the point, asked Hodge to withdraw the paper, and took months to be convinced. Lacking independent sources for the details of this affair, which earned Hodge a first invitation to Princeton, let me just point out an interesting twist in Atiyah's account of it: On the one hand, after sketching the proof in modern cohomological terms, he justly points out that back in 1930, "complex manifolds (other than Riemann surfaces) were not conceived of in the modern sense, and the simplicity of the proof indicated above owes much to Hodge's work in later years which made complex manifolds familiar to the present generation of geometers." Yet on the other hand, Atiyah also expresses his surprise that Lefschetz did not grasp Hodge's argument immediately. In my view, that this happened to somebody like Lefschetz provides additional, first hand historical evidence for how different the situation in 1930 actually was from what our rewritten account suggests. Once a line of thought—like the argument developed in [46]—has been recast in a universally practised technique, its originality at the time it was conceived becomes very hard for us to appreciate.

Lovers of Western classical music may have encountered this problem. Being familiar, say, with Robert Schumann's piano works that are part of today's standard repertoire (*Carnaval*, *Davidsbündlertänze*, *Kinderszenen*, *Kreisleriana*, ...), one cannot listen to his more rarely performed little *opus* 1, the Abegg-Variations in F-major, without being *reminded* of what Schumann

actually composed *later*. This then makes it genuinely difficult to understand the helpless critiques which Schumann's *opus* 1 received in 1832, when the best available stylistic comparison was probably with the time-honoured composer Johann Nepomuk Hummel.

It is the job of the historian of mathematics to recognize such *remembrance of things to come* in the reading of old documents which the ongoing rewriting of mathematics offers us time and again, and to set the historical record straight as far as the available documentary evidence permits.

2. Rewriting Historiography

Rewritings on all scales make up the very fabric of mathematical activity through the centuries. But the notion of rewriting—which I will use here in a loose, non-terminological fashion—works best on a relatively local, micro-historical level which makes specific comparisons of an original document with rewritten versions of it possible. The objective is then to describe explicit *transformations of epistemic objects and techniques*, where these latter terms are to be taken in the sense that Moritz Epple ([23], pp. 14–17) has extracted from Hans-Jörg Rheinberger's approach to the history of laboratory science [63]. Since the mathematical research we will be looking at works essentially without physical machinery, its epistemic evolution, even if a lot of it may happen orally, will finally be documented almost exclusively by textual documents, published or unpublished papers, notes, correspondence, etc. This may justify speaking of rewriting.

In [30], Catherine Goldstein has not only conducted such analyses of various rewritings—she speaks of different readings (*lectures*) instead—of one of Fermat's marginal notes—the one that proves in particular that there is no right-angled triangle with rational sides whose area is a rational square—but she has in fact reconstructed the history of this note over 350 years as a collection of such readings. If one likes mathematical metaphors, the structure that Goldstein winds up with may remind one of a complicated covering space that history has superimposed on that marginal note written in the first half of the 17th century; each reading is a reading of Fermat's note and thereby related to it, and certain readings are also related among each other. But for the historian all rewritings are created equal; each lives on its own respective sheet.

When it comes to major, general upheavals in the history of mathematics, the analysis of rewritings can easily run into the famous warning sign that every French child reads and learns: *un train peut en cacher un autre*. For instance, Herbert Mehrtens [55] has tried to describe the overall history of mathematics around the turn from the 19th to the 20th centuries as the finally victorious incursion of modernism into mathematics. Understanding this modernist transformation continues to be a currently active research topic in the history

of mathematics.¹ Already certain strands of this macrohistorical phenomenon, such as “the entrance of non-Euclidean geometries” during the 19th century, may appear so cataclysmic that “exceptionally one uses the term ‘revolution’ even for mathematics” to describe them ([55], p. 44). Yet, a closer look usually discovers subtle webs of intertwined rewritings which in any event deserve to be disentangled.

An example from the history of non-Euclidean geometries is provided by Eugenio Beltrami who would insist on building (the planimetry) of his non-Euclidean pseudosphere on the *substrato reale* of C.F. Gauss’s differential geometry of curves (see [29], pp. 48–51; letter to Hoüel of 2 Jan. 1870, p. 117), thereby integrating it into an established part of mathematical analysis, i.e., the differential geometry available to him. This is still a far cry from David Hilbert’s reworking of the Foundations of Geometry at the end of the century (see [43]) which transformed the axioms of geometry from specific claims about known objects, such as points, into implicit definitions of these very (potential) objects, turned geometry into an autonomous mathematical discipline—*which* geometry fit nature was no longer for the geometer to decide—and would in due course give birth to mathematical logic and model theory as a new, equally autonomous sub-discipline of mathematics. Obviously, qualifications like ‘modern’ or ‘abstract’ apply to Hilbert’s approach much better than to Beltrami’s attitude. Yet, Beltrami was a pivotal author of the non-Euclidean ‘revolution’, rendering Riemann’s vision of intrinsic geometry concrete in a crucial case.

There are thus different stories—or histories—to be told here, and a good history of non-Euclidean geometry will of course mention both of them in turn, plus several others, like the ones alluded to in [39], pp. 690–699. The question remains whether one wants the formalist re-interpretation of axioms as implicit definitions, and the logical inspection of various axiom systems, to be, or not to be, a core feature of the rewriting of geometry after the dawning of non-Euclidean geometries. Rouse Ball, for example, included at the end of the fourth, 1908 edition of his *History of Mathematics*, in the discussion of non-Euclidean geometries ([3], pp. 485–489), references to recent works published as late as 1903; but he would not mention Hilbert’s 1899 book on the Foundations of Geometry ([43], chap. 5) at all. Did he maybe think that the problem about the nature of space raised by the new geometries was such a burning issue for our natural philosophy that Hilbert’s purely “logical analysis of our intuition of space” ([43], p. 436) was less relevant in comparison? After all, “there are indeed reasons . . . for suggesting that to see the search for non-Euclidean geometry in this axiom-based way is an artifact of mathematical modernism that distorts the historical record.” ([33], p. 51)

The scattered episodes I am about to present are loosely held together by the mathematical concept of a point and they deal with rewritings. They date

¹See for instance [33], [81], and <http://web.uni-frankfurt.de/fb08/HS/wg/gif.html>.

from before, during, and after World War I. I present these episodes in order to illustrate various historical regards.

3. Arithmetic Points

Just as his teachers Peter G. Lejeune-Dirichlet and Bernhard Riemann, whose conceptual approach to mathematics he consciously emulated and carried further, Richard Dedekind (1831–1916) has been seen by many as a pioneer of modern mathematics, particularly by Emmy Noether who participated in the edition of his collected papers. Her forward-looking comments on many of them, and her own further development of the theory of modules and ideals, make it plausible that she sometimes felt as if “all she had done was to develop Dedekind’s ideas” [1]. But Emmy Noether’s close reading and rewriting of Dedekind’s is not our subject here. I will look at Dedekind’s two famous contributions to the question of what a point is: in his little 1872 brochure [17], he proposed what we all know as Dedekind cuts to define real numbers arithmetically; and in his seminal joint paper with Heinrich Weber ten years later [20], he defined a purely arithmetic avatar of a point on an algebraic Riemann surface.

Both definitions are remarkably similar; both try to conceptualize the intuition of what a concrete point does for you. On the real line, fixing a point can tell you, can it not, where to cut the line in two, and in Dedekind’s analysis [17], the idea of continuity is precisely that every cut in the line is also afforded by a point. So, if you ban intuition but still want to define a point, a real number, in the linearly ordered continuum, with only rational numbers at your disposal, you just define the point by the cut, i.e., as being a partition of the rationals in two subsets, one of which has all its elements smaller than all the elements of the other. Likewise, on an algebraic Riemann surface, a point will be something where you can evaluate (sometimes getting the value $\infty \dots$) rational functions living on the Riemann surface; and you know or postulate that rational functions ought to be sufficiently plentiful to separate points. So, if you can’t see the Riemann surface, but still have its field of rational functions, define a point arithmetically as an evaluation homomorphism (including possible values ∞) on rational functions which leaves constant functions invariant. That is what Dedekind and Weber did in the second part of [20], showing subsequently that such an evaluation mapping defines a prime ideal in the coordinate ring, hence a maximal ideal (we are in dimension one), etc. They carried the theory as far as the Riemann-Roch Theorem, formulated and proved in this purely arithmetical setting. (A point in the sense of Dedekind and Weber was the first special case of what was later called a ‘place’ of a field; see for instance [97], p. 3.)

Both of these conceptual creations of points by Dedekind were *arithmetizations*. This label ‘arithmetization’ has been used in many different ways at the end of the 19th century; we have explored in [60] the panoply of arithmetizing

treatments of the continuum and how they were received and rewritten by various mathematicians and philosophers at the time. Suffice it to emphasize here that Dedekind's arithmetization is not an axiomatization. Dedekind cuts were invented to *define* real numbers in terms of infinite sets of rational numbers—never mind the debates about impredicative definitions or set theoretical paradoxes this approach would encounter later on. Also the definition of a point on an algebraic Riemann surface given in [20] plainly relies on an *arithmetization* of the theory of algebraic Riemann surfaces because Dedekind and Weber prepare this definition by a good fifty pages “of a purely formal nature” ([20], §14) which carry over to rings of (entire) algebraic functions most of the apparatus that Dedekind had developed earlier for the theory of algebraic integers in number fields. It is this theory which then allowed them, among many other things, to quickly deduce that every point—in the sense of evaluation mapping—corresponds to a prime ideal of the coordinate ring.

The paper by Dedekind and Weber has attracted considerable attention, and indeed praise, in the historical and philosophical literature. The philosopher David Corfield, for example, insists that the paper of Dedekind and Weber constitutes “a watershed in the use of analogy” and quotes Jean Dieudonné to the effect that “this article by Dedekind and Weber drew attention for the first time to a striking relationship between two mathematical domains up until then considered very remote from each other . . .” ([15], p. 96) Corfield then further embeds his argument into a “historical claim” about a broad mathematical road towards a “structural outlook.” ([15], p. 98) However, his whole discussion of [20] starts not in the 19th century but as a remembrance of things to come, with a discussion of two well-known pieces by André Weil: the letter to his sister Simone [86], 1940a, from the beginning of World War II, on the analogies that were guiding Weil when he tried to develop the theory of correspondences on curves over finite fields, and Weil's even later variant of it [87], 1960a. Furthermore, the discussion of the article by Dedekind and Weber is mixed with a brief look at Kurt Hensel's different approach to the arithmetic of algebraic function fields.

One may of course sympathize with Dieudonné's statement quoted by Corfield, about the “very remote” mathematical domains that Dedekind and Weber managed to bring together, if one remembers Riemann's geometrical and topological (*analysis situs*) approach to Riemann surfaces, and the unique feature of Riemann's theory of abelian functions which actually obtains the rational functions on the algebraic Riemann surface from a combination of *transcendental* functions with specified local behaviour. But there are also good reasons to rethink—and rewrite—Dieudonné's and Corfield's conclusion. In fact, the tectonics of sub-disciplines of mathematics was very much in flux during the 19th century, and statements about the relative distance of two domains of mathematics at a given point in time raise intricate questions.

To start with the classification adopted by the mathematicians at the time: the *Jahrbuch über die Fortschritte der Mathematik* (vol. 14 for the year 1882,

published in 1885) reviewed the article by Dedekind and Weber neither under the heading *Algebra* (section II) nor under *Number Theory* (section III), nor under any of the chapters on algebraic curves, surfaces, etc. in section IX (*Analytic Geometry*), but under section VII, *Function Theory*, chap. I: *Generalities*. This is reasonable in that the paper explicitly gives a new general treatment of the algebraic functions on Riemann's surfaces, as the reviewer (Otto Toeplitz's father Emil, teacher at a Gymnasium in Breslau) also duly points out. By the way, the subsequent review in this volume of the *Jahrbuch*, in the same section and chapter, is of Felix Klein's very pedagogical exposition of Riemann's theory with appeal to physical intuition. On the other hand, Dedekind published another big paper the same year, on the discriminant of an algebraic number field; it was classified in the *Algebra* section II, chap. 2 on the *Theory of forms*. But the Italian translation of Dirichlet's Lectures on Number Theory edited by Dedekind which also appeared in 1882 did make it into the *Number Theory* section. Clearly the perception of the mathematical sub-disciplines at the time was not ours, looking back from today.

Trying to describe as best we can the 'domain of mathematics' at the time *into which* Dedekind's and Weber's paper [20] integrated part of Riemann's theory of algebraic functions, we have to say that it was the arithmetic (or number theory) in the wake of Kummer's "ideal numbers." This specialty, it turns out, was practised by less than a handful of mathematicians between, say, 1860 and 1880 (see [31], chap. I.2, §3). Among them two researchers clearly stand out: Leopold Kronecker and Richard Dedekind. As for Kronecker, he assures his readers ([50], p. 197) that as early as 1858 he had actually communicated to Riemann the main result of an algebro-arithmetic investigation of his on the discriminant of algebraic functions in one variable, because it provided a better justification than the one Riemann had given for a simplifying assumption concerning the ramification points, which Riemann used throughout his theory of abelian functions. Kronecker also tells us he discussed these ideas with Weierstrass who folded them into his Berlin lectures on abelian functions. At any rate, Kronecker did not publish a paper on this at the time, although he did include his ideas in several lecture courses he gave in Berlin, as I was able to confirm in handwritten notes from Kronecker's lectures which are kept at the Mathematics Library of IRMA, Strasbourg. In 1880 however, when Weber submitted the manuscript of [20] to *Crelle's Journal*, Kronecker apparently decided to only look at it after publishing in the upcoming volume 91 (1881) what he presented as his old write-up from 1862 ([50], pp. 193–236) together with a preface explaining its history. Dedekind's and Weber's paper was thus delayed until the subsequent, 1882 volume of *Crelle's Journal*, only to be again preceded there by yet another paper of Kronecker's closely related to the theory developed by Dedekind and Weber: a reprinting of his momentous 1881 *Grundzüge* [50], pp. 239–387. Now these *Grundzüge*—which the *Jahrbuch* classifies in the section on *Algebra*, chap. 1, *General theory of algebraic equations*—sketch a very complete, unified arithmetic theory of algebraic integers and algebraic functions (of

arbitrary many variables). So Dedekind and Weber—who had come into closer contact when they were both collaborating to prepare Riemann’s works and some of his unpublished papers for the edition of his Collected Papers—were not alone with their explicitly arithmetic approach to algebraic functions; in fact, all the usual suspects, so to say, were publishing along this line in the early 1880s, even if their papers ended up in different drawers of the classifiers.

Furthermore, the arithmetic rewriting of Riemann’s analytic theory was not an unlikely idea at the time because competing digests of Riemann’s theory were around which, even if they were not arithmetic, all started, unlike Riemann, from the explicitly given algebraic functions: see the overview in [11], p. 287. That the two arithmetic attempts were, in spite of the tensions between Dedekind and Kronecker, akin in many ways is also confirmed by Weber’s textbook presentation of the whole paper [20] in [83], pp. 623–707, which follows the original rather faithfully, except that Weber replaced Dedekind’s method of ideals by Kronecker’s forms (just as Hilbert had done in his *Zahlbericht* [44], pp. 63–363, in the very proof of the uniqueness of factorization of ideals into prime ideals in the ring of integers of an algebraic number field, cf. [66]).

Finally, it should not be overlooked—because it emphasizes the algebro-arithmetic nature of the paper—that, in spite of the impressive theorems which they manage to prove so neatly by the arithmetic method, Dedekind’s and Weber’s paper does remain incomplete in that the naked abstract ‘surface’ of which they have defined the points is not endowed with any sort of topological or analytic structure. This is not an anachronistic comment of mine because the last sentences of the introduction show that the authors wanted to come back to this, and it is a pity that neither of them ever did.

For a better understanding of what is at stake here, let us look at a broader time scale. When Catherine Goldstein and I were preparing the first two chapters of [31], we sifted through 19th century papers directly or indirectly taking up Gauss’s *Disquisitiones Arithmeticae*, and discovered for the approximate period 1825–1860 a domain of research connected with Gauss’s work that “knit together reciprocity laws, infinite series with arithmetical interpretations, elliptic functions and algebraic equations.” ([31], p. 52) We called this domain *Arithmetic Algebraic Analysis* and we argued ([31], pp. 24–38, 52–55) that it constituted a (research) *field*, in the sense of Bourdieu: “all the people who are engaged in [this] field have in common a certain number of fundamental interests, viz., in everything that is linked to the very existence of the field,” and one can uncover “the presence in the work of traces of objective relations . . . to other works, past or present, [of the field].” ([7], p. 115) We had indeed found that the main actors of this domain were “linked by a dense communication network, both personal and mathematical. Their published papers would meet with prompt reactions. . . . An interesting characteristic feature was the production of new proofs of the central results.” ([31], p. 52) And we explained ([31], p. 54) that the coming together of very different types of objects, methods and

results does *not* suggest calling the field of Arithmetic Algebraic Analysis a mathematical *discipline*.

Kronecker had participated in that field; Richard Dedekind in his younger years had still seen it in action; and for such a versatile mathematician as Heinrich Weber—who moved from first papers in mathematical physics to being the author of the 3 volume *Lehrbuch der Algebra* where he would rewrite, among other things, all the main results obtained by Arithmetic Algebraic Analysis—the heritage of this field held no secrets. Insofar as Dedekind and Weber managed to provide an alternative proof of Riemann’s theorems by means of an arithmetico-algebraic method, one might therefore be tempted to consider their paper [20] a late contribution to the practice of Arithmetic Algebraic Analysis. But this is not so. Times had changed. Not only had the field of Arithmetic Algebraic Analysis largely died out ([31], chap. I.2); but at least since the 1870s, arguing about the adequacy of the method employed had become an important issue for many mathematicians, esp. in Germany. Dedekind and Kronecker were no exception. Recall how Dedekind argued in [17] that he had uncovered with his cuts the true conceptual essence of the intuitive idea of continuity, or completeness, of the real line. On Kronecker’s side, his bitter controversy of 1874 with Camille Jordan about the proper way to set up the theory of bilinear forms and their normal forms confirms this point perfectly—see Frédéric Brechenmacher’s detailed analysis [10] of this controversy and of the different images of bilinear algebra which it brought to the fore.

In that controversy, Kronecker insisted very much on the *generality* of a theorem of Weierstrass’s which did not have to exclude, or treat separately, the case of characteristic polynomials with multiple roots. He also claimed a superiority of the *arithmetical* point of view, which is more difficult to pinpoint precisely. The same values are also appealed to by Dedekind and Weber when they describe what their treatment avoids: “In previous investigations of the subject, certain restrictive assumptions have usually been imposed on the singularities of the functions studied, and these so-called exceptional cases are either obtained parenthetically as limit cases, or simply excluded. In the same way, certain principles about continuity and existence of [series] expansions are admitted whose evidence relies on various sorts of geometric intuition.” Furthermore, there is the value of *simplicity* which, when it is coupled with the said generality or *completeness*, seems to have captured the highest standards of a scientific treatise (see [35] for a discussion of the scientific values at the time from the point of view of the humanities). Kronecker explicitly refers to the physicist Gustav Kirchhoff for these values ([50], pp. 353–354), and Dedekind & Weber announce their “simple, and at the same time rigorous and completely general point of view” right away in the first sentence of [20]. Finally, Dedekind struck a similar note when he graciously thanked his younger colleague after two years of work on the joint project, quoting from Pascal’s letter to Fermat: “. . . I see that the truth is the same in Toulouse as in Paris”, and then commenting on their collaboration “which after various oscillations increasingly

took on the character of intrinsic necessity.” ([19], p. 488) And sure enough, on the page just quoted from Kronecker, [50], p. 354, he also claims necessity for his method, i.e., the above-mentioned forms as alternative to Dedekind’s ideals.

All this rhetoric leads us away from the inspirational play of fruitful analogies. It points towards establishing arithmetic as a model approach to the theory of algebraic functions. But this model stood not alone; other methods rivalled or complemented it. Apart from the analytic approach and the algebraic geometry of the time, there was in particular the theory of Hensel and Landsberg—their book [42] was dedicated to Dedekind on the occasion of the fiftieth anniversary of his doctorate—who used local series expansions as input for an otherwise purely arithmetic theory, but then went further than Dedekind and Weber, treating abelian integrals. Emmy Noether would call this the “accretion of Weierstrass and Dedekind-Weber” ([57], p. 273), but in view of Hensel’s own mathematical history as a student of Kronecker’s, this is not a final historical assessment.

Even though Emmy Noether would increasingly see herself in the line of thought of Dedekind during the 1920s, in 1919 she published the well-balanced report [57] which supplied the chapters deliberately left out of the report that Brill and her father had compiled a generation earlier [11]. The first part of Emmy Noether’s report sets an impressive example of a dense, virtuoso and impartial comparison of the existing theories, concise and yet explicit down to the different arrangements of proof for corresponding theorems. In the final section 8, her report briefly explores analogies within the theory of algebraic functions, of transcendental problems in the theory of algebraic numbers. These analogies lay outside of the scope of Dedekind and Weber; they had been briefly hinted at in Hilbert’s statement of his twelfth problem in 1900, [45], pp. 312–313. A budding new sub-discipline which started from this sort of analogy, but then converted to basic tenets of the Dedekind-Weber programme was the arithmetic theory of algebraic function fields of one variable over a finite field of constants. It first began after 1900, and then afresh after World War I, from an analogy for fields of algebraic functions of the analytic theory of algebraic number fields, in particular the analytic class number formula. Then, reversing prior practice in this new field, F.K. Schmidt decided in 1926 to work with *all* points afforded by the field, according to the first principles of Dedekind and Weber, and to change the definition of the zeta function of the field accordingly. (See [64], p. 571–572; cf. [31], pp. 174–178)

4. Holistic Points

We have seen that Kronecker and Dedekind attached great importance to certain methodological values. To be sure, all scientists have values they try to live up to in their work and, just like the values shared by Kronecker and Dedekind,

these will usually not be limited to one branch of science and will typically be in tune with the ambient culture. In this section, I take one more look at Dedekind to draw attention to a basic concept of his mathematics which has a holistic ring to it. This will then be the occasion to look at holistic tendencies in mathematics in general, and in particular in the 20th century.

Dedekind's very basic and very successful concept which I am alluding to is *Körper*. Unfortunately, this is nowadays called a *field* in English, which may remind one, if not of Bourdieu, of agriculture or cricket. Earlier, feeble attempts to introduce the word *corpus*, and its plural *corpora*, into English mathematical terminology instead (see for instance [80], [41]) have never caught on. What is totally lost in the translation *field* is Dedekind's motivation for choosing the term *Körper*: "Similarly as in the sciences, in geometry and in the life of human society, this name is to denote also in the present context a system which exhibits a certain completeness, perfection, self-containedness through which it appears as an organic whole, a natural unity." ([19], p. 20)

Dedekind's *Körper* were not our fields, neither mathematically nor philosophically. Mathematically, his *Körper* consisted of complex numbers; and he called "finite field" (*endlicher Körper*) a finite extension of the rational numbers, i.e., an algebraic number field. Never did he take the step—although he knew of course Galois's *imaginaires de la théorie des nombres*—to extend what we view as the general field axioms to infinite and finite sets alike. (The first one who published such a parallel treatment of fields with infinitely many and finitely many elements would be Heinrich Weber [82]. The systematic exploration of the modern axiomatic notion of field is due to Steinitz [79] and was prompted by the advent of a new type of examples: Hensel's *p*-adic fields.) Various possible reasons for Dedekind's failure to really go structural here, are discussed in [38], pp. 108–109. The very intimate connection between his "finite fields" and the theory of algebraic numbers presumably had an important part in it. Dedekind's vision of his *Körper* as a basic object of arithmetic is also reflected in the fact that he described the inclusion of fields as a division. ([19], p. 409; [38], pp. 106–107; [16], Part I, chap. 2.) For Dedekind, the importance of fields was not that they represented a basic algebraic structure, whereas he did appreciate groups for this ([38], pp. 107–108). He treated his *Körper* as the active entities on which algebra and number theory rest. In a letter to Lipschitz from 1976 he alluded to the intrinsic possibility of defining inside an algebraic number field its ring of algebraic integers as the "number-theoretic capabilities of the field" (*zahlentheoretische Fähigkeiten des Körpers*), which are wasted once one fixes a primitive element to deal with a finite extension of the rationals.

For reasons that I find impossible to trace, but which might have to do with the holistic ring of the word, Dedekind's *Körper* had a direct impact on the immediately following generation, unlike Kronecker's terminology from his *Grundzüge*, and unlike other parts of Dedekind's theory. While Hurwitz, Weber and Hilbert would time and again substitute some of Dedekind's arguments in

the theory of ideals by Kroneckerian arguments using the adjunction of indeterminates (a method that Dedekind abhorred as being not intrinsic), Dedekind's "finite fields" were turned into the pivotal notion of Hilbert's *Zahlbericht* ([44], pp. 63–363) whose full title reads "The theory of algebraic number fields." (Addressing their rings of integers directly, as a primary object of study, would become common only as a result of Emmy Noether's works from the 1920s on the axiomatics of commutative rings and their ideals.)

Sticking for simplicity with literature in German, the *Fortschritte* database lists 69 papers between the first one in 1882 (an article by Dedekind) and 1914 which have the word *Körper* (in the algebraic sense) or the word *Zahlkörper* already in the title. The little industry really took off around 1900. The arithmetic model also induced Felix Hausdorff to introduce analogous terminology into set theory "on the basis of a vague analogy from which one should not demand too much": a *Mengenkörper* is a set of sets which, with two member-sets, also contains their union and difference. ([39], p. 115) It was by this bias that the word *Körper* even made it into Kolmogorov's 1933 axiomatization of probability, see [49], p. 2.

And when Deuring, Hasse and his collaborators embarked as of 1936 on translating from the theory of Riemann surfaces into the theory of function fields the notion of a correspondence between two Riemann surfaces, or two algebraic curves, they studied the arithmetic of the field generated over the fixed field of constants K by (algebraically independent) generators of the two given function fields of one variable. They called this field of transcendence degree 2 over K the *Doppelkörper* attached to the situation. André Weil would not tire of deriding the clumsiness of this method—see [86], p. 253, from the text quoted by Corfield; [87], p. 14: "notably unsuccessful paper of Deuring"; [6], p. 104, note 18: "... orthodox successors of Dedekind"; cf. [68]. Weil embarked instead soon after 1940 on a fundamental algebraic rewriting of Algebraic Geometry to which we will return below. At least Deuring himself may not have been such an "orthodox successor of Dedekind" after all; he had apparently proved that the ring of correspondences of an algebraic curve over a finite field had characteristic zero by first lifting the curve to characteristic zero, doing an analytic argument on the associated Riemann surface, and then reducing back to the original finite characteristic. According to [37], p. 347, he deleted this argument—which may appeal to us today but which Hasse qualified as "unfair"—from the galley proofs of the second installment of his paper [21]. (Cf. [68])

Let us stop this field trip here. What would Dedekind say, if he knew that today his organic fields were, each single one of them, just a point, in Grothendieckian Algebraic Geometry, and not even a thick one?

Holism in the sciences has been studied especially in the history of the life sciences. A good example of such a study is Anne Harrington's *Reenchanted Science* [36]. Starting from Kant's Third Critique (*Kritik der Urteilskraft*) and Goethe's *Farbenlehre*, continued in the early 19th century by the idealist

philosophers' quest for completed systems in the face of political fragmentation of Germany, holistic ideas within the sciences in Germany increasingly turned into a revolt against the machine image of life towards the end of the 19th century. Nor was this only a German phenomenon: the French philosopher Henri Bergson clearly owed much of his popularity to the timeliness of his message, when he pointed for example to the incompatibility between our inner sense of duration (*durée*) and movement on the one hand, and our daily life surrounded by time pieces and nature as described in mechanical treatises on the other—see for instance his *Essai sur les données immédiates de la conscience* [4], pp. 58–80. But an important point of Harrington's analysis is that holistic currents in the sciences responded (by way of metaphors) to political agenda in Germany. Harrington's case studies of holistic thinkers—whose careers often evolved outside the scientific mainstream, but who nonetheless marked the history of their disciplines—include the biologists Jakob v. Uexküll with his key notion of *Umwelt* and Hans Driesch; the neurologist Constantin von Monakow; the writer on Wagner and the Aryan race Houston Stuart Chamberlain; a list of *Gestalt*-psychologists from Christian von Ehrenfels to Max Wertheimer, Wolfgang Köhler and others; and the expert of brain research and therapy Kurt Goldstein.

Holistic influences in mathematics have not received comparable attention. And it may not be obvious at first, what holism would mean in mathematics and what it could do for the history of mathematics. Gerolamo Cardano for instance was clearly a holistic scientist; metals were for him inhabited by the soul of the world. But the bearing of this on the history of his publication of Tartaglia's formulae for solving a cubic equation by radicals seems negligible. All through the nineteenth century, it seems very problematic to attribute holistic tendencies to mathematicians, or to try and use this category to better understand major debates. The holistic element we stressed in Dedekind's choice of the word *Körper* is undoubtedly there, as shown by his own comment when he introduces it; but which role this particular emphasis given to the term played for the practice of this notion I find impossible to trace, and so the observation does not seem to add anything exploitable to the analysis of Dedekind's guiding values, and of his conceptual approach which he traced to his teachers Dirichlet and Riemann.

However, I contend that the attribute 'holistic' is well-suited and useful to explain certain mathematicians' attitudes in the 20th century. Since the phenomenon clearly touched different branches of science at that time, something may then be gained by looking across disciplinary boundaries, and the history of philosophy during the period ought to be simply integrated into the history of sciences for the purpose. Two examples of mathematicians immediately suggest themselves: Luitzen Egbertus Jan Brouwer and Hermann Weyl. One might think of adding other mathematicians, Andrei Kolmogorov for instance (in view of the constructivist side of his *œuvre*) or Erich Kähler (considering

his *magnum opus* [48]). But at least for this talk, I will concentrate just on Hermann Weyl.

It is well-known that Weyl went through different periods in his thinking about the foundations of mathematics. (Cf. [69]) In spite of the fact that as early as 29 July 1910 he would write to Piet Mulder in Holland (I owe this quote to Dirk van Dalen): “I have recently thought about the foundations of set theory and were led to views which diverge rather strongly from Zermelo’s, coming close in a certain sense to the point of view of Borel and Poincaré which is generously derided around here” (i.e., in Göttingen), his *Habilitation* lecture that same year, “On the definition of the fundamental notions of mathematics”, did not come to a very skeptical conclusion about the possibility of founding all of mathematics on set theory, but rather ended on a note which for me is typical of the first period of Weyl’s thinking: “May we say—as is suggested by what we have developed—that mathematics is the science of ϵ [i.e., the element relation in set theory] and of those relations which can be defined from this notion via the principles discussed? Maybe such an explanation does actually determine mathematics correctly *as for its logical substance*. However, I see the proper value and the meaning proper of the system of notions of logi-cised mathematics thus constructed in that its notions may also be interpreted *intuitionwise* without affecting the truth of the statements about them. And I believe that the human spirit has no other way to ascend to mathematical notions but by digesting the given reality.” ([91], p. 304) And Weyl would play on the same theme in the 1913 preface to his book *Die Idee der Riemannschen Fläche* [88]. In this book, Weyl famously rewrote Riemann’s ideas on the basis of an abstract notion of two-dimensional manifold ([88], §4), and used very recent analytic results to secure the existence of functions via Dirichlet’s principle. (A slightly different axiomatic description of a topological manifold, also in terms of neighbourhoods, was conceived independently at about the same time by Hausdorff—see the comments in [39], pp. 712–718.) One key message from Weyl’s preface is that, for reasons of rigour, there is no alternative to building up very technical, abstract theories, even though one has to be aware of the fact that this necessity has “also brought about unhealthy phenomena. Part of the mathematical production has lost . . . the connection with the living stream of science.” Therefore, “. . . to grasp what accounts for the life, the true substance, the inner value of the theory: for this a book (and even a teacher) can only provide scanty indications; here everyone has to wrestle himself afresh to gain understanding.”

In other words, Weyl suffered from the apparent incompatibility between the human, intuitive, ideal core of mathematics, and the artificial scaffolding we have to erect in order to obtain a sound scientific theory. But he is not prepared yet to move the holistic reaction against this difficult, potentially inhuman state of affairs into the formal mathematical work. The holistic conception remains an individual task to be mastered beyond and in spite of the modern, distorting

presentation of the theory. The dilemma is relegated to prefaces or concluding exhortations.

World War I would change this, but at first in the direction of an even bigger divide between rigorous mathematics and the human mathematical activity. This is characteristic of the second phase in Weyl's thinking about the foundations. In 1913, Hermann Weyl married Helene Joseph in Göttingen and was appointed professor at ETH Zürich. However, in May 1915 he was drafted into German military service. It did not involve actual fighting though, just a stay at a garrison near Saarbrücken. An article on Riemann surfaces that would appear in 1916 ([91], pp. 600–613) was written there without access to mathematical literature. A year later, the Swiss authorities managed to obtain his release from his military duties, the Weyls returned to Zürich, but work did not go on where Weyl had left it. In 1916, Weyl started reading J.G. Fichte and Meister Eckart with a philosopher colleague in Zürich, Fritz Medicus. Meanwhile the courses that he taught indicate his new orientation: In the summer of 1917 he lectured on *Raum, Zeit, Materie*, and in the following winter on the *Logical foundations of mathematics*. ([27]) These courses gave each rise to a book published in 1918: [89] and [90]. And it is the latter which reflects the war experience within the foundations of mathematics.

Mathematically, the *Kontinuum* [90] constructs a viable but deliberately poor version of the continuum which systematically and carefully avoids all impredicative definitions, i.e., all quantification over sets of primitively defined objects. (See Solomon Feferman's analysis in his article "Weyl vindicated" in [25], pp. 249–283, which elaborates on the surprising logical efficiency of Weyl's poor analysis.) Since the notion of upper bound of an infinite set of real numbers, analysed in terms of Dedekind cuts, involves such a higher order quantification, the existence of an upper bound cannot be proved for *every* bounded set in Weyl's poorer continuum, although it can be established for denumerable subsets of real numbers. Two lines of philosophical and rhetorical arguments stand out in Weyl's book: (i) the claim that the uncontrolled, impredicative usage of Dedekind cuts introduces a vicious circle into analysis, and that in order to prevent that theory from falling to pieces, one has no choice but to be content with the poorer continuum presented in this book; (ii) discussions of the problematic applications of the poor continuum to physics, i.e., to an arithmetized model of space-time.

Point (i) was criticized already at the time—see Weyl's reply to a letter from Otto Hölder [92], pp. 43–50—and until very recently; Paolo Mancosu, for instance, expressed the opinion that Weyl's contention "is a far cry from pointing out a vicious circle in the foundations." ([54], p. 75) His Zürich colleague Georg Pólya apparently did not believe in the vicious circle either, as is confirmed by a wager between Weyl and him, which other Zürich colleagues signed as witnesses, dated 9 February 1918 (see [65], p. 15, for the original text). As is to be expected from the book [90], the text of the wager confirms that Weyl at the beginning of 1918 saw no way of proving along traditional lines the existence

of the precise upper bound of a bounded set of real numbers. But the wager went beyond the book in making predictions about the rewriting of the foundations of mathematics to be expected over 20 years, until 1937. Here Weyl seems prepared for coming research which could produce new, precise theories of the continuum where the existence theorem of the upper bound would not hold in general. If, however, that general existence theorem could actually be established by 1937 in a rigorous way—without any *circulus vitiosus*—then this would only be possible because of a truly original rewriting of the foundations in a way impossible to imagine now, i.e., in 1918.

An illuminating perspective on Weyl's ideas at the time can be gleaned from a postcard he wrote to Pólya on 29 December 1919. I found it quoted in Reinhard Siegmund-Schultze's thorough analysis of the correspondence between Pólya and Richard von Mises [75], p. 472. In my reading of this postcard, Weyl compares "two things": his own earlier debate with Pólya which led to the wager, and Pólya's ongoing debate with Richard von Mises about the latter's foundations of probability theory. Weyl writes that these two issues are more closely related than Pólya may have originally thought, but that in his ongoing debate with von Mises, Pólya finds himself on the side which corresponds to Weyl's part in their earlier debate about the foundations of analysis. Now Pólya's foremost criticism of von Mises's axiomatics concerned the irregularity axiom for collectives, [56], p. 57, *Forderung II*; see also [61], p. 184, and [40], pp. 825–833. I therefore think that Weyl is alluding to the analogy that, both in the definition of the upper bound of a set of real numbers (given as Dedekind cuts) and in von Mises's second axiom, a property has to be checked for an infinity of sequences or sets of objects satisfying certain requirements. If Pólya finds this "mathematically not viable" (*mathematisch untauglich*, as he writes to von Mises, [75], p. 501), then this strikes Weyl as very much analogous to his own criticism of a vicious circle on the ground that the notion "property of rational numbers" is not extensionally definite (*umfangsdefinit*, [92], p. 45). A much later rewriting, from the 1960s, of the theory of collectives at the hands of Kolmogorov and in terms of the algorithmic complexity of subsequences would resuscitate the theory of collectives in a new mathematical outfit. ([61], pp. 233–237) As for the rewritings of the foundations of the continuum predicted by Hermann Weyl, we shall encounter one anon.

We have seen that Weyl's claim of a vicious circle met with various criticisms. But I am reading his line of thought (i) in the book on the continuum as a reaction to the abysmal cultural experience of World War I, transposed into the problems about the foundation of mathematics. The way I see it, Weyl was closing the shutters because of the storm outside. He had been reading Fichte's relentless scrutiny of the act of judgment and the potential evidence provided by intellectual intuition which Fichte construed in analogy with proofs by geometric construction ([90], p. 2). Fichte was also the author of the *Reden an die deutsche Nation* which had helped to rally resistance against the French troops in 1807–1808. I conjecture that, in a similar vein, the return to the rock-bottom

of absolute evidence in the face of potentially shaky foundations, and the restriction to tightly controlled methods of object construction was for Weyl a natural rejoinder to a war whose visible effects were increasingly hard to reconcile with the origins of the civilization that had unleashed it. This part of the book was not a move towards a more holistic, humane way of doing mathematics; it was a rescue operation, faced with a world which was threatening to go to pieces.

But the second line of thought (ii) mentioned above went beyond the immediate purpose of saving a minimal secure form of analysis. Here Weyl took stock of just how far mathematics had gone astray as a consequence of its modern development: “If we make precise the notion of set in the way here proposed then the claim that to every point on the line . . . correspond a real number, and vice versa, acquires a profound content. It establishes a peculiar link between what is given in our intuition of space and what is construed in a logical-conceptual manner. But this claim obviously leaves entirely the scope of what intuition teaches us or may teach us about the continuum; it is no longer a morphological description of what intuition offers us . . .” ([90], p. 37) And on p. 68: “It is the great merit of Bergson’s philosophy to have emphasized this profound alienation of the world of mathematical concepts from the immediately experienced continuity of the phenomenon of time.” The security of sound foundations is thus obtained in *Das Kontinuum* at the high price of violating even more our intuition of space and time. The continuum is all but holistically satisfying for the mathematician-physicist Hermann Weyl. Seeing no way out of this dilemma during the war, he resigned himself to sketching the principles of physical applications of the poor continuum. To even start doing this, to determine a point, one has to refer to a coordinate system: “The coordinate system is the inevitable residue of the annihilation of the ego in that geometrical-physical world which reason carves out of what is given under the norm of ‘objectivity’; the last meagre symbol even in this objective sphere for the fact that existence is only given and can only be given as intentional content of the conscious experience of a pure, sense-creating ego.” ([90], p. 72) Never did the reality of space seem further from our mathematical models of it to Hermann Weyl than at the end of World War I. The choosing of a point, i.e., the very beginning of the soulless arithmetization of what once was a lived intuition, is the only act remaining to remind the mathematician-physicist of his creative self.

Meeting Brouwer in the Engadin in the Summer of 1919 liberated Weyl from this dehumanized, atomistic mathematical universe and started the short period, his third, during which he believed in holistic analysis. The articles he wrote to propagate this view are full of political metaphors reflecting the collapse of the German empire as well as the ensuing revolution and inflation; in this way Weyl’s holistic turn is made to reflect, to match the historical moment. These passages are well-known or easy to find. Let me rather quote from the holistic rhetoric here: “Mathematics is, as Brouwer occasionally puts it,

more of an activity than a doctrine. . . . Brouwer's view ties together the highest intuitive clarity with freedom. It must have the effect of a deliverance from a nightmare for whoever has maintained any sense for intuitively given facts in the abstract formalism of mathematics." ([92], p. 157/179) "The ice cover has burst into floes, and now the element of flux was soon altogether master over the solid. L.E.J. Brouwer sketches a rigorous mathematical theory in which . . . the continuum is not conceived as a rigid being but as medium of free becoming. With this we also regain our freedom as concerns number sequences and sets of numbers. We no longer try to gain a yes or no answer . . . by stretching the sequences on the Procrustean bed of construction principles. With Brouwer, mathematics gains the highest intuitive clarity . . ." ([92], p. 528/530)

The technical gadget which Weyl received from Brouwer were the choice sequences, *Wahlfolgen*, which in general are eternally in the making; only finite beginnings of them can be considered given. A point is defined by a choice sequence of natural numbers that encode nested intervals: "The whole admits parts" replaces the principle that a set has elements. ([92], p. 177) "The continuum appears as something which is infinitely in the making inside." ([92], p. 172) A precise point, for instance $x = 0$, does not cut the continuum in two, because whether an arbitrary other point is or is not equal to x may be undecided. The new continuum is uncuttable. When we want to study a function on it, we have to "hover over" the new continuum, because we cannot "sit down" on an arbitrary point of it. ([92], p. 179) If ever there was 're-enchanted mathematics', an enchanted continuum—in analogy with the title of [36]—this is one. Considering how much more readily a biologist or a neurologist can deliver scientific verdict on organic connections or expressions of life, Weyl's "medium of free becoming", as he calls the intuitionist continuum, strikes me as a remarkably coequal holistic notion for a mathematician. (Cf. [36], pp. xxviii–xix)

Weyl's allusions to the postwar situation place his holistic articles in a period of time which was favourable for holistic writers, at least in Germany. The following is an extract from a petition of his students in Zürich dated 6 May 1920 which was written in order to prevent Weyl's leaving Zürich for Göttingen or Berlin: "Our conviction that Herr Prof. Weyl is irreplaceable has its source in the following reasons: We admire in him the ingenious creator of new cultural values which consist in that the exact sciences come into fruitful interaction with life itself. It is this exceedingly fortunate fusion of the man and the scholar in Herr Prof. Weyl which inspires in each one of us a sense of liberation . . . and seems to us to guarantee most surely that whole men will emerge from the eighth section", i.e., the Mathematics Institute. ([27], pp. 43–44)

Possibly the most far-reaching consequence that Hermann Weyl was seriously considering in pursuing his holistic mathematics was the *inherently probabilistic universe*. In a way, Weyl carried the comparison he had made in his postcard to Pólya over to his continuum based on choice sequences, which begin to look like random variables: "the quantitative data in a piece of the (space-

time) world S are known only approximately, with a certain margin, not only because my sensory organs have limited precision, but *they are affected in themselves by such a vagueness*. . . the future will continue working on the present; the past is not terminated. This lifts the rigid pressure of natural causality and opens up—irrespective of the validity of the laws of nature—a *space for autonomous decisions which are causally totally independent from each other* and which according to me take place in the elementary quanta of matter.” ([92], p. 121–122; cf. p. 173; cf. [61], p. 68–70)

This may remind one of Paul Forman’s old and oft debated thesis [26] to the effect that German physicists let go of traditional deterministic principles after World War I in order to accommodate to the Weimar Republic whose cultural climate was hostile to traditional scientific ideas such as determinism and whose societal reorganisation threatened the academic elite. I have not worked on such a grand scale. I have been following Hermann Weyl’s individual path and found the more literary passages in his works explicit enough to link them to the war, resp. to the postwar period. His being in tune with the historical events surely helps to explain the students’ petition. On the other hand, Weyl seems himself only half convinced that probabilism is really a corollary of his “medium of free becoming”, i.e., of the holistic continuum based on choice sequences. ([92], p. 122, footnote) And he does not seem to take up this hypothesis again in later articles. So Weyl may well be an individual case matching Forman’s old thesis. But I find it more remarkable that holism and probabilism are tentatively linked by Weyl via the notion of choice sequence.

At any rate, Weyl’s holistic continuum was a fairly ephemeral phenomenon. In fact, Weyl’s partisanship for the intuitionism lasted less than 10 years, and even during this time, his mathematical research outside of the foundational articles shows hardly any sign of intuitionist practice. Also, in the second half of the 1920s he tried to steer a mediating course between Brouwer and Hilbert. The reasons for Weyl’s final abandonment of Brouwer’s cause are not clear and deserve further historical investigation. Mancosu [54], pp. 80–81, discusses this completely from the point of view of the relationship with Hilbert. Epple [24] has suggested that intuitionism itself simply did not manage to live up to the high standards of proof that it called for. Remembering the frequent interaction of holistic ideas with the ambient cultural and political climate, and the fact that for Hermann Weyl, unlike other holistic scientists, taking the Nazi turn in the 1930s was never an option, for both political and personal reasons, the course of general history may also have contributed to the brief duration of Weyl’s holistic mathematics.

5. Generic Points

Still in the holistic vein and trying to address as large a part of the mathematical community as possible, still treating foundational issues at the end and suggesting a mediation between Brouwer’s and Hilbert’s programmes, Hermann Weyl

published in 1924 an article “Marginal notes on main problems of mathematics” ([92], pp. 433–452) which revisited, in a new and transparent presentation, a few problems that, according to Weyl, interest “all those who deserve to be called mathematicians, in essentially the same way.” Solomon Lefschetz from Princeton reacted to this project in a letter of 30 November 1926 (HS 91:659 in the Archives of ETH Zürich): “. . . For any sincere mathematical or scientific worker it is a very difficult and heartsearching question. What about the young who are coming up? There is a great need to unify mathematics and cast off to the wind all unnecessary parts leaving only a skeleton that an average mathematician may more or less absorb. Methods that are extremely special should be avoided. Thus if I live long enough I shall endeavor to bring the theory of Algebraic Surfaces under the fold of Analysis and An.[alysis] Situs as indicated in Ch. 4 of my Monograph [52]. The structure built by Castelnuovo, Enriques, Severi is no doubt magnificent but tremendously special and requires a terrible ‘entraînement’. It is significant that since 1909 little has been done in that direction even in Italy. I think a parallel edifice can be built up within the grasp of an average analyst.”

So Lefschetz was ready to rewrite Algebraic Geometry, or more precisely the major Italian work in Algebraic Geometry, i.e., above all the classification of algebraic surfaces, in his topologico-analytical approach. Lefschetz felt that in this way Algebraic Geometry could be reconnected to the hard core of mathematics. Note that such a rewriting would not amount to an algebraization of the Italian body of knowledge. About ten years later, when the founding fathers of Bourbaki started working towards their encyclopedic project, a few established sub-disciplines of mathematics, specifically probability theory and algebraic geometry, were still often thought (if not by the Bourbakists themselves) to be not amenable to insertion into a project like the *Eléments*, built on axiomatics starting with (logic and) set theory. The “terrible *entraînement*” needed to penetrate work of the Italian school, as Lefschetz had felt, was thought to be due to some specific intuition employed in this discipline, which would make it not reducible to logic and set theory. ([86], p. 555) As is well-known, the sub-discipline of Algebraic Geometry was in fact completely rewritten and remodelled, essentially between 1925 and 1950, by various mathematicians, and not within the Bourbaki project although Bourbaki members did play an important role, notably Weil [85]. To conclude my talk I would like to briefly discuss ways to describe this major rewriting historically. Before doing so, however, let us make sure that what we are talking about really makes sense.

First of all, that a mathematical sub-discipline of Algebraic Geometry with its own history did indeed exist, say around 1930, is documented in particular

- by a string of reports which took stock of the domain: Brill & Noether (1892–93) [11], Castelnuovo & Enriques (1914) [12], Emmy Noether (1919) [57], Snyder *et. al.* (1928/34) [78], Berzolari (1933) [9], Commessati (1932) [14], Geppert (1931) [28];

- by a string of monographs which highlight both the field and a wide range of interactions between different approaches; examples include: Schubert (1879) [70], Picard & Simart (1897–1906) [62], Bertini (1907) [8], Hensel & Landsberg (1902) [42], Severi (1908/1921) [71], Zeuthen (1914) [98], Enriques & Chisini (1915–1924) [22], Lefschetz (1924) [52], Jung (1925) [47], Severi (1926) [73], Coolidge (1931) [13];
- by ongoing production as evidenced for instance in the first volumes of *Zentralblatt* (founded in 1931). Various subsections have to be surveyed here in order to gather all the aspects of the domain we would like to trace, also in anticipation of the later rewriting: in the first place those for algebraic geometry, algebraic surfaces, algebraic curves, birational transformations; and then increasingly also sections on the theory of fields and rings. The rewriting that has taken place since can also be judged from the fact that certain authors stood out in the early thirties as particularly prolific in the bibliographical record whom the memory of the community has not conserved according to the number of their publications, Lucien Godeaux for example;
- by Hilbert’s fifteenth problem: rigorous foundation of Schubert’s calculus of enumerative geometry. Not only was this a problem in the domain of Algebraic Geometry, but it was a foundational problem, of which Severi for instance had admitted in 1912 that it was “something more than just a scruple about exaggerated rigour.” [72] In that same paper, Severi reformulated the problem in terms of algebraic correspondences which considerably enhanced its link with ongoing work in the field.

Furthermore, speaking of the *Italian school of algebraic geometry* also makes good historical sense because, after a strong initial contribution by Alfred Clebsch, Max Noether, as well as Alexander v. Brill and Paul Gordan, the main development—important foreign influence notwithstanding, for instance by Emile Picard—did lie in the hands of Italian mathematicians such as—apart from the three names mentioned by Lefschetz—Eugenio Bertini, Pasquale del Pezzo, Corrado Segre, Beppo Levi, Ruggiero Torelli, Carlo Rosati. These Italian mathematicians formed a social web and often published in not very international Italian journals. ([76], pp. 100–104) At least until the early 1930s, Italy was the place for many to go and learn Algebraic Geometry. Finally, by the 1930s, there was one uncontested leader governing the school: Francesco Severi after his fascist turn, and finally director of the newly founded *Istituto Nazionale di Alta Matematica* inaugurated on 15 April 1940. ([34], *passim* and in particular p. 272)

So who was attacking, or approaching, from where and how this international sub-discipline, and in particular its Italian branch, with a view to rewriting it? Lefschetz’s monograph [52] already contained such a partial topological rewriting, concerning algebraic surfaces and correspondences on curves. This lead was followed by Oscar Zariski’s papers on the fundamental group mostly

from the 1920s, and by Bartel L. van der Waerden's topological solution of Hilbert's 15th problem from 1929, which used intersections in the homology ring of the ambient variety. ([67], pp. 260–264)

But arguably the first attempt at an explicit *refoundation* of Algebraic Geometry grew out of Emmy Noether's work on the ideal theory of rings, and was published by van der Waerden in 1926 where we read the lines that were presumably written by Emmy Noether herself: "The rigorous foundation of the theory of algebraic varieties in n -dimensional spaces can only be given in terms of ideal theory, because the definition of an algebraic variety itself leads immediately to polynomial ideals. Indeed, a variety is called algebraic, if it is given by algebraic equations in the n coordinates, and the left hand sides of all equations that follow from the given ones form a polynomial ideal. However, this foundation can be formulated more simply than it has been done so far, without the help of elimination theory, on the sole basis of field theory and of the general theory of ideals in ring domains." ([67], p. 251) From this resulted a new notion of point on an (affine, say) algebraic variety which van der Waerden called *allgemeine Nullstelle*, i.e., a general zero (of a set of algebraic equations).

Here is in essence van der Waerden's simple observation (for a more complete analysis of this paper, see [67]): If K is a field and $\Omega = K(\xi_1, \dots, \xi_n)$ a finitely generated extension of it, then all polynomials f in $R = K[x_1, \dots, x_n]$ such that $f(\xi_1, \dots, \xi_n) = 0$ form a prime ideal \wp in R , and Ω is isomorphic via $x_i \mapsto \xi_i$ to the integral domain R/\wp . Conversely, given a prime ideal \wp in R , then there exists an extension field $\Omega = K(\xi_1, \dots, \xi_n)$ of finite type such that \wp consists precisely of the polynomials f in R such that $f(\xi_1, \dots, \xi_n) = 0$; indeed, it suffices to take $\xi_i = x_i \pmod{\wp}$ in R/\wp . Such a system (ξ_1, \dots, ξ_n) in an extension field of finite type of K is called a general zero of the ideal \wp , or a *general point* of the variety in affine n -space over K defined by the prime ideal \wp . Even though all this looks extremely elementary today, the definition, together with the notion of specialization, i.e., van der Waerden's *relationstreue Spezialisierung*, is one of the central notions of the algebraic rewriting of algebraic geometry in the 1930s and 1940s. Proofs of theorems in the rewritten algebraic geometry typically involve choosing general points of all varieties with which one has to work.

Significantly, van der Waerden when defining these general points also offered a bridge linking them to the traditional terminology of algebraic geometers saying that the general point just defined "... agrees with the meaning that the words general and special have in geometry. Indeed, by a general point of a variety, one usually means, even if this is not always clearly explained, a point which satisfies no special equation, except those equations which are met at every point. For a specific point of M , this is of course impossible to fulfil, and so one has to consider points that depend on sufficiently many parameters, i.e., points that lie in a space $C_n(\Omega)$ [affine n -space], where Ω is a transcendental extension of K . But requiring of a point of $C_n(\Omega)$ that it be a zero of all those and only those polynomials of $K[x_1, \dots, x_n]$ that vanish at all

points of the variety M yields precisely our definition of a general point of the variety M ." In other words, van der Waerden claimed that he was really only rewriting in modern algebraic language what Italian geometers for instance had meant. He also said that the traditional literature was not particularly clear on this.

Traditional Algebraic Geometry had been particularly rich in all sorts of points: apart from just plain points, there were infinitely near points of various orders, intersection points of varying order, virtual double points, etc., and there were what the Italians called *punti generici*, a word that A. Weil in [85] imported into English as "generic point", but with the precise mathematical meaning of van der Waerden's general zero. The question that arises from our last quote from van der Waerden is how well the algebraic rewriting captures what is being rewritten. Let us look at a correspondence between the two actors who would finally impose the new Algebraic Geometry by the end of the 1940s, Oscar Zariski and André Weil (from the Zariski papers in the Harvard Archives). Both of them were using van der Waerden's general points, but Zariski called them 'general', Weil 'generic'.

On 25 March 1952, Weil writes to Zariski:

... I wonder whether it is not too late to persuade you to reconsider the use of the words 'general' and 'generic.' Any unnecessary discrepancy between our terminologies is bound to accentuate the all too prevalent impression that there is a sharp cleavage between your work and mine, which is simply not true. When I selected 'generic', I certainly was not unaware of the fact that 'generale' is quite as good Italian as 'generico'. But I don't think that the Italians ever gave a sharp definition for either word; they just used them loosely. I adopted 'generic' because it is a less common word than 'general', both in French and in English, and therefore seems to lend itself better to a strictly technical meaning. One does not need two words, I contend: some points are (in my sense) 'generic', relatively to a given field, i.e., 'general' in your sense; but no point is generic in your sense. If I understand you right (from your remarks in your Congress lecture), what you mean when you say that a property P holds 'at a generic point' seems to me to be much better expressed by saying that P holds on an open set (in your topology), or (as Seidenberg does) by saying that P holds almost everywhere. I doubt very much whether the Italians ever differentiated sharply between the two concepts. As you have seen them at much closer quarters than I ever did, I am willing to take your word as to what they thought about this or that; but this is psychology, not mathematics; and I do not think that it need bother us. What is far more important is not to create unnecessary difficulties to young people who are now trying to learn algebraic geometry from your work and from mine. ... Maybe you will ask why I don't adopt the sim-

ple remedy of changing over to your terminology. Now: a) if I had found ‘general’ in common use in a well-defined technical sense, I should certainly not have tried to change it; this not being the case, I decided upon ‘generic’ for the reason indicated above, which is not a very strong one, I admit; b) . . . c) having, for the punishment of my sins, written and published a book, I am far more committed to my terminology than you who are yet to publish yours and therefore still enjoy far greater freedom in such matters.

Zariski’s answer to Weil is dated 29 March 1952:

I hope that you will not hold it against me if I say that you have not convinced me on the evidence in re general versus generic. I claim that from the work of the Italians it appears quite clearly (and objectively, not just as a matter of psychological interpretation) what they meant by the term ‘generic.’ Next I claim that, without reading a single line of the Italian papers but just using the fact that in the Italian school the ground field and the coordinate field were identical, namely the field of complex numbers, one must conclude with the corollary that their generic point could not possibly be the same thing as the ‘allgemeiner Punkt’ of van der Waerden. Finally, it is not quite true that no point is generic in my sense. I agree that no point is generic (in my sense) in itself, just as no point is generic (in your sense) in itself. Incidentally, I notice that also outside of algebraic geometry (for instance in function theory) mathematicians begin to use the term generic, and obviously not in your sense. . . .

Here Zariski could have gone to his bookshelf and quoted from [22], p. 139: “The notion of a *generic* ‘point’ or ‘element’ of a variety, i.e., the distinction between properties that pertain *in general* to the points of a variety and properties that only pertain to *exceptional* points, now takes on a precise meaning for all algebraic varieties. A property is said to pertain in general to the points of a variety V_n , of dimension n , if the points of V_n not satisfying it form—inside V_n —a variety of less than n dimensions.” We see that the authors of this quote tacitly assume that the exceptional points form a subvariety; this point will be raised incidentally in Weil’s answer of 15 April 1952:

Dear Zariski, I have no remarks of mathematical interest to make at the moment, but I want to express my renewed doubts about what the Italians are supposed to have meant by ‘generic’. Your arguments, purporting to show that they meant it in your sense, would indeed be decisive if they had been logical thinkers in such matters; but in that case they would have defined the word, and there would be no controversy. It is a plain fact (as again emphasized, and quite rightly, by Chevalley in a review of some article by Severi [74] in

Math. Rev. last year, I think) that the Italians were of the opinion that every proper subset of a variety which is defined by algebraic geometric means is a union of subvarieties; this belief alone accounts for their obstinate contention that they knew all about the Chow coordinates, when in fact the main theorem to be proved there (viz., that there is an algebraic set, every point of which is the Chow point of some cycle of the given dimension) is entirely missing from their work. This clearly means that they were *essentially unable* to distinguish between your sense of the word ‘generic’ and mine. What they would do, of course, is to *prove* that a generic point in my sense has a certain property, and to *conclude* that a generic point in your sense has that property. Presumably I have been paying more attention to their proofs, and you to their statements, so that we may well both be right. Also, the argument based on the fact that their ground field and coordinate field were identical (viz., the complex numbers) would be valid only if they had thought clearly on these subjects. Not only with them, but in the greater part of classical mathematics, a ‘variable’ is essentially a transcendental element over the field of complex numbers, even though it is never defined that way but usually as ‘an arbitrary complex number’; it follows that classical mathematics, including of course Picard and the Italians, is full of contradictions which cannot be disentangled unless one reinterprets the word ‘variable’ as I have just said. As those people were no fools, one must conclude that they had some obscure notion of “a transcendental element over complex numbers” but lacked the algebraic language to express it. . . .

Do you want me to tell you who is right? That’s easy: both are right (says the historian). The substantial difference between them is their relationship to Italian Algebraic Geometry.

Oscar Zariski—born in 1899 as Ascher Zaritski in the small town of Kobrin, then Russia, as of 1921 Poland, today Belarus—managed to go to Italy to study in 1921 and was trained in Rome, then the world center for Algebraic Geometry. Lefschetz had been visiting there before Zariski’s arrival; Severi transferred to Rome in 1922. Zariski got his doctorate with Castelnuovo in 1924 and worked also for the philosophically enclined Enriques, preparing for instance an Italian translation of Dedekind’s foundational writings, in particular of [17], with extensive commentary. Since he was not a naturalized Italian, university positions were closed to him. After two postdoc years in Rome on stipends of the Rockefeller Foundation, Castelnuovo obtained through Lefschetz for Zariski to go to the US, at first to Baltimore. Not surprisingly in view of Lefschetz’s letter quoted above, Zariski published in 1928 a paper “On a theorem of Severi” [93] where he criticized a proof that Severi had given in 1913, and proposed a topological approach instead. He took the measure of his former masters on a much bigger scale in his 1935 *Ergebnisse* volume on Algebraic Surfaces [94]

where the typical comments one finds are of the following sort. p. 18: “It is important, however, to bear in mind that in the theory of singularities the details of the proofs acquire a special importance and make all the difference between theorems which are rigorously proved and those which are only rendered highly plausible.” p. 19: “In regard to the “accidental” singularities introduced by the quadratic and the monoidal transformations and in regard to the manner in which they should be eliminated by birational transformations, Levi’s proof is not sufficiently explicit.” p. 20/21: “. . . What matters, however, and is essential for the application which Severi makes of this lemma is that . . . Hence the above formula is not correct. Since the composition indices are not diminished by projection, we can only write . . .”, and so on.

During his stay at the Institute for Advanced Study, Princeton, in 1935–1936 he came into contact with modern algebra and in particular read Wolfgang Krull’s works. Building on this, he managed between 1937 and 1947 what he called himself an *arithmetization* of Algebraic Geometry. One of the basic ideas was to define points on what he called—alluding to Dedekind and Weber from a vastly more general situation—the arithmetic “Riemann surface” attached to a polynomial ring (or a ring of formal power series) by looking at valuations with general value groups (generalizing in this way the discrete rank one valuations associated to any of Dedekind’s and Weber’s points). In this way he managed in particular to build “an arithmetic theory parallel to the geometric theory of infinitely near points” on a smooth algebraic surface. ([95], p. 14) Other big achievements of this period include: the definition of the *normalization* of a projective variety, the resolution of singularities of surfaces (two different proofs) and threefolds, and the Zariski topology. ([95], [96], [59], [77], [76])

Zariski brought with him from Rome the training and the central problems in Algebraic Geometry. When the experience of his *Ergebnisse* volume suggested the necessity to rewrite a good deal of this corpus of knowledge, he was open and creative enough to use completely different methods—those which Krull had defined as properly “arithmetic” in his personal terminology ([51], p. 746, footnote 2)—but he would never betray the language in which he had first discovered the world of Algebraic Geometry. If a generic point was a complex point in general position back in Rome, it would still be so in the US. And when he introduced the concept of a normal projective variety—which would subsequently give him the process of normalization and thereby a completely new desingularization procedure, practically impossible to reconstruct in the language of Italian Algebraic Geometry—Zariski chose this very word ‘normal’ in analogy to a traditional terminology, as if this could smoothen the transition. ([95], p. 112; cf. [77]) This shows that the casting of a rewriting is influenced by *allegiances* (in the broadest sense). Zariski’s allegiance was with his Rome education. When he was helping his Dutch colleague Kloosterman to organize the symposium on Algebraic Geometry at the Amsterdam ICM in 1954, he commented on a preliminary list of invited speakers: “There are several names I would add to your list. I am particularly worried by the omission of the name of

Severi. I think that Severi deserves a place of honor in any gathering of algebraic geometers as long as he is able and willing to attend such a gathering. We must try to avoid hurting the feelings of a man who has done so much for algebraic geometry. He is still mentally alert, despite his age, and his participation can only have a stimulating effect. I think he should be invited to participate.” (Zariski to Kloosterman, 15 January 1954) But to be sure, as we have seen, the allegiance to his Rome education had long ceased to constrict Zariski’s methods; it concerned a tradition, not a working environment. In Italy during the 1930s it was impossible to openly criticize Severi; ever since he had arrived in the US, Zariski enjoyed and used the freedom from this sort of allegiance and constraint.

André Weil’s allegiances are less easy to detect and to describe. He had also been to Rome on a Rockefeller stipend, but only briefly. For the Amsterdam ICM, André Weil would negotiate with Kloosterman a special session, within the Algebraic Geometry Symposium, on equivalence relations for algebraic cycles. This was first prompted by the announcement of Segre’s ICM lecture (Weil to Zariski, 19 January 1954), but would finally result in a direct showdown between Severi and Weil on the subject. The ensuing voluminous correspondence between Severi, van der Waerden and Pierre Samuel (preserved in the van der Waerden papers at ETH Zürich) was finally condensed in an article printed in the ICM Proceedings, but Severi would carry his grief about Weil’s attack in the discussion after his talk for years; see [5].

Weil was probably the most widely read of the Bourbaki members at the time; Bourbaki’s historical endnotes [6] were his idea and many of them supplied by him. The argument developed in his second letter quoted above is perfectly compatible with the philosophy of these notes; the rewritten, the modern mathematical notion looks for subsumable elements in older texts. And the older literature does indeed speak routinely about moving points on a variety which are mapped somewhere etc. An element which is transcendental over the ground field can model this. It is not that Zariski is the more careful historian of the two; he just refuses to let new terminology interfere with ways of formulating to which he is attached. Weil had no such specific allegiance with the Italian school. For him this was one of several corpusses of texts from the recent history of mathematics with which he had gained a certain familiarity. He undoubtedly had an allegiance to the group of collective individualists ([32]) Bourbaki of which he had been a cofounder. This can be seen inside the history of the rewriting of Algebraic Geometry by following Weil’s and Claude Chevalley’s respective contributions to it; Chevalley is also mentioned in the above letter. On a less personal note, Weil’s allegiance to Bourbaki is reflected in the format of his book [85]. It was one of the first unmistakably bourbakist books that appeared, even though it was not part of the *Eléments de mathématique*.

The word allegiance is unsuitable to describe the relationship between Weil and Zariski. But the evolution of this relationship and the way in which their

two individual projects grew—without ever merging completely—into something that was finally perceived as one rewriting of Algebraic Geometry can be followed thanks to their correspondence.

There are different ways to tell the story of a rewriting. Part of the work is of course to follow the mathematical details of published papers and available correspondence carefully. I have been doing this for several years now concerning the rewriting of Algebraic Geometry in the 1930s and 1940s. One could think that this would do the job, all the more so as the number of rewriters in that period is not big, less than ten, and their works—even when crossed with quoted literature and with the considerable resilience of the Italian school, esp. through Severi's amazing production in the 1930s—are in principle surveyable, and since the rewriting took place under the motto of new rigour, the new methods, notions and objects brought into play are relatively easy to recognize and to describe mathematically. But working on this, one notices that the history of the phenomenon in the large cannot be captured in this fashion. A better historical account on the scale of this whole rewriting of Algebraic Geometry emerges by describing allegiances. This notion allows to treat factors like methodological preferences due to established values, personal respect or the relationship between teacher and student, academic power, political agenda, and others all at once. The picture obtained in this way is something like a graph, with a small number of actors with surprisingly few coalitions among them, but definite power flows along the various edges.

In [67] for instance, I have followed van der Waerden's seemingly erratic course in his long and rich series of articles on Algebraic Geometry. It falls into place in terms of his allegiances: He first became interested in Algebraic Geometry through a lecture by Hendrik de Vries at the University of Amsterdam on Schubert calculus. In Göttingen, he was part of the group around Emmy Noether (and Emil Artin in Hamburg); his first paper, where his general points are defined, was written in that situation, in particular it was written from outside of the Algebraic Geometry community. After meeting Severi in 1932, he drastically reduced the level of algebraic abstraction in his papers and used geometric intersection constructions which were due to Severi. But he could never take advantage of the friendly course he was steering with respect to his influential Italian colleague because his enemies in the Nazi administration made it impossible for him to travel abroad. For the same reason, he had to be careful when dealing with Hasse because of the latter's political influence in Germany until 1945; I have documented [67], p. 274, a case where, against Hasse's wish, van der Waerden did not publish a 1941 proof he had done in response to a query from Hasse; but after the war he used this proof to criticize Hasse's notion of point in the very article which he had helped Hasse to complete.

Helmut Hasse's more active interest in Algebraic Geometry goes back to Deuring's 1936 programme for proving the analogue of the Riemann Hypothesis for all algebraic curves over finite fields. His contact with Severi was more

political than mathematical and started late, in the Spring of 1937. The projected axis of collaboration between the German algebraists and the Italian geometers, which they wanted to be parallel to the political axis between the two fascist states, hardly got off the ground. ([68])

The triangle van der Waerden — Severi — Hasse thus appears to have functioned in a way which effectively hampered the constitution of a new joint European research practice in Algebraic Geometry, in spite of the substantial string of papers *Zur Algebraischen Geometrie* and the excellent textbook [84] which van der Waerden published from his splendid isolation in Leipzig. The allegiances in the triangle and the political agenda which enforced them already before the war emerge even more clearly if one compares them to the absence of similar constraints which the ex-Europeans in the US—Zariski, Weil, Chevalley—were enjoying during the war. And during the last years of the war, there were hardly any actively competing rewriters in Europe. In this precise sense, the new Algebraic Geometry which would set the standard of the sub-discipline in the 1950s and until Grothendieck's second rewriting, was a product of the second World War, or more exactly of the World Wars, considering 1914–1945 as a single period of world history, marked by global warfare.

References

- [1] P.S. Alexandrov, Pages from an autobiography. *Russian Mathematical Surveys* 35 (3) (1980), pp. 315–359.
- [2] M.F. Atiyah, William Vallance Douglas Hodge. *Bulletin of the London Mathematical Society* 9 (1977), 99–118.
- [3] W.W. Rouse Ball, *A Short Account of the History of Mathematics*. London (MacMillan & Co) 1908.
- [4] H. Bergson, *Œuvres*. (Textes annotés par A. Robinet). Edition du centenaire. Paris (PUF) 1959.
- [5] G. Bolondi, C. Petrini, Lo scambio di lettere tra Francesco Severi e André Weil. *Lettera Matematica PRISTEM* 52 (2004).
- [6] N. Bourbaki, *Elements of the history of mathematics* (transl. J. Meldrum). Berlin, Heidelberg, etc. (Springer) 1994.
- [7] P. Bourdieu, Quelques propriétés des champs. Exposé ENS novembre 1976. In *Questions de sociologie*, Paris (Les Editions de Minuit) 2002; pp. 113–120.
- [8] E. Bertini, *Introduzione alla geometria proiettiva degli isperspazi con appendice sulle curve algebriche e loro singolarità*. Pisa (Spoerri) 1907.
- [9] L. Berzolari, *Algebraische Transformationen und Korrespondenzen*. Enzyklopädie der mathematischen Wissenschaften III C 11. 1933.
- [10] F. Brechenmacher, La controverse de 1874 entre Camille Jordan et Leopold Kronecker, *Revue d'histoire des mathématiques* 13 (2007), 187–257.

- [11] A. Brill, M. Noether, Die Entwicklung der Theorie der algebraischen Functionen in älterer und neuerer Zeit, Bericht erstattet der Deutschen Mathematiker-Vereinigung. *Jahresbericht der Deutschen Mathematiker-Vereinigung* 3 (1892–93), 107–566.
- [12] G. Castelnuovo, F. Enriques, *Die algebraischen Flächen vom Gesichtspunkte der birationalen Transformationen*. Enzyklopädie der mathematischen Wissenschaften III/2, 6b. 1914.
- [13] A. Coolidge, *A treatise on algebraic plane curves*. Oxford (Clarendon) 1931.
- [14] A. Comessatti, Reelle Fragen in der algebraischen Geometrie. *Jahresbericht der Deutschen Mathematiker-Vereinigung* 41 (1932), 107–135.
- [15] D. Corfield, *Towards a Philosophy of Real Mathematics*. Cambridge (CUP) 2003.
- [16] L. Corry, *Modern Algebra and the Rise of Mathematical Structures*. Basel, Boston, Berlin (Birkhäuser) 1996. 2nd ed. 2004.
- [17] R. Dedekind, *Stetigkeit und irrationale Zahlen*. Braunschweig (Vieweg) 1872. Repr. in [19], pp. 315–334.
- [18] R. Dedekind, *Gesammelte mathematische Werke*. Vol. 1. Braunschweig (Vieweg) 1930.
- [19] R. Dedekind, *Gesammelte mathematische Werke*. Vol. 3. Braunschweig (Vieweg) 1932.
- [20] R. Dedekind, H. Weber, Theorie der algebraischen Functionen einer Veränderlichen, *Journal für die reine und angewandte Mathematik* 92 (1882), 181–290. Repr. in [18], pp. 238–350.
- [21] M. Deuring, Arithmetische Theorie der Korrespondenzen algebraischer Funktionenkörper. *Journal für die reine und angewandte Mathematik* 177 (1937), 161–191; 183 (1940), 25–36.
- [22] F. Enriques, O. Chisini, *Lezioni sulla teoria geometrica delle equazioni e delle funzioni algebriche*, vol. I. Bologna (Zanichelli) 1915.
- [23] M. Epple, *Die Entstehung der Knotentheorie. Kontexte und Konstruktionen einer modernen mathematischen Theorie*. Braunschweig, Wiesbaden (Vieweg) 1999.
- [24] M. Epple, Did Brouwer’s intuitionistic analysis satisfy its own epistemological standards? In *Proof Theory. History and Philosophical Significance*. (V.F. Hendricks, S.A. Pedersen, K.F. Jorgensen, eds.) Boston, London (Dordrecht); pp. 153–178.
- [25] S. Feferman, *In the Light of Logic*. Oxford (OUP) 1998.
- [26] P. Forman, Weimar culture, causality, and quantum theory, 1918–1927: Adaptation by German physicists and mathematicians to a hostile intellectual environment. In *Historical studies in the physical sciences* (R. McCormach, ed.), Vol. 3. Philadelphia (University of Pennsylvania Press) 1971.
- [27] G. Frei, U. Stambach, *Hermann Weyl und die Mathematik an der ETH Zürich 1913–1930*. Basel (Birkhäuser) 1992.
- [28] H. Geppert, Die Klassifikation der algebraischen Flächen. *Jahresbericht der Deutschen Mathematiker-Vereinigung* 41 (1931), 18–39.

- [29] L. Boi, L. Giacardi, R. Tazzioli, *La découverte de la géométrie non euclidienne sur la pseudosphère. Les lettres d'Eugenio Beltrami à Jules Hoüel (1868–1881)*. Paris (Blanchard) 1998.
- [30] C. Goldstein, *Un théorème de Fermat et ses lecteurs*. Saint-Denis (Presses Universitaires de Vincennes) 1995.
- [31] C. Goldstein, N. Schappacher, J. Schwermer (eds.), *The Shaping of Arithmetic after C.F. Gauss's Disquisitiones Arithmeticae*. Berlin, Heidelberg, New York (Springer) 2007.
- [32] C. Goldstein, La théorie des nombres en France dans l'entre-deux-guerres: De quelques effets de la première guerre mondiale. *Revue d'histoire des sciences* 62–1 (2009), 143–175.
- [33] J. Gray, *Plato's Ghost. The Modernist Transformation of Mathematics*. Princeton, Oxford (Princeton University Press) 2008.
- [34] A. Guerraggio, P. Nastasi, *Italian Mathematicians Between the Two World Wars*. Basel etc. (Birkhäuser) 2005.
- [35] M. Hagner, M.D. Laubichler (eds.), *Der Hochsitz des Wissens. Das Allgemeine als wissenschaftlicher Wert*. Zürich, Berlin (diaphanes) 2006.
- [36] A. Harrington, *Reenchanting Science. Holism in German Culture from Wilhelm II to Hitler*. Princeton University Press 1996.
- [37] H. Hasse, *Mathematische Abhandlungen*. Vol. II. Berlin (de Gruyter) 1975.
- [38] R. Haubrich. *Zur Entstehung der algebraischen Zahlentheorie Richard Dedekinds*. Dissertation Göttingen 1992.
- [39] F. Hausdorff, *Gesammelte Werke*. Vol. II. Berlin, Heidelberg, New York (Springer) 2002.
- [40] F. Hausdorff, *Gesammelte Werke*. Vol. V. Berlin, Heidelberg, New York (Springer) 2006.
- [41] H. Heilbronn, E.H. Linfoot, On the imaginary quadratic corpora of class-number one. *Quarterly Journal of Mathematics, Oxford* 5 (1934), 293–301.
- [42] K. Hensel, G. Landsberg, *Theorie der algebraischen Funktionen einer Variablen und ihre Anwendung auf algebraische Kurven und abelsche Integrale*. Leipzig (Teubner) 1902.
- [43] M. Hallet, U. Majer (eds.), *David Hilbert's Lectures on the Foundations of Geometry 1891–1902*. Berlin, Heidelberg, New York (Springer) 2004.
- [44] D. Hilbert, *Gesammelte Abhandlungen*. Vol. I. Berlin, Heidelberg, etc. (Springer) 1932 (2nd ed. 1970).
- [45] D. Hilbert, *Gesammelte Abhandlungen*. Vol. III. Berlin, Heidelberg, etc. (Springer) 1935 (2nd ed. 1970).
- [46] W.V.D. Hodge, On multiple integrals attached to an algebraic variety. *Journal of the London Mathematical Society* 5 (1930), 283–290.
- [47] H. Jung, *Algebraische Flächen*. Hannover (Hellwig) 1925.
- [48] E. Kähler, Geometria aritmetica. *Annali di Matematica pura ed applicata* Ser. IV, Tom. XLV (1958). 399 pp.

- [49] A. Kolmogoroff. *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin (Springer) 1933.
- [50] L. Kronecker. *Werke* (K. Hensel, ed.). Vol. 2. Leipzig (Teubner) 1897.
- [51] W. Krull. *Gesammelte Abhandlungen / Collected Papers*. Vol. 1. Berlin (de Gruyter) 1999.
- [52] S. Lefschetz. *L'Analysis situs et la géométrie algébrique*. Paris (Gauthier-Villars) 1924.
- [53] R. Lipschitz, *Briefwechsel mit Cantor, Dedekind, Helmholtz, Kronecker, Weierstrass* (W. Scharlau, ed.). Braunschweig, Wiesbaden (Vieweg) 1986.
- [54] P. Mancosu, *From Brouwer to Hilbert. The Debate on the Foundations of Mathematics in the 1920s*. Oxford (OUP) 1998.
- [55] H. Mehrrens, *Moderne — Sprache — Mathematik. Eine Geschichte des Streits um die Grundlagen der Disziplin und des Subjekts formaler Systeme*. Frankfurt a.M. (Suhrkamp) 1990.
- [56] R. von Mises, Grundlagen der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* 5 (1919), 52–99.
- [57] E. Noether, Die arithmetische Theorie der algebraischen Funktionenkörper einer Veränderlichen in ihrer Beziehung zu den übrigen Theorien und zu der Zahlkörpertheorie. *Jahresbericht der Deutschen Mathematiker-Vereinigung* 38 (1919), 182–203. Repr. (with incomplete title and the final footnote missing) in [58], pp. 271–292.
- [58] E. Noether, *Gesammelte Abhandlungen / Collected Papers* (N. Jacobson, ed.). Berlin, Heidelberg, etc. (Springer) 1983.
- [59] C. Parikh, *The Unreal Life of Oscar Zariski*. Boston etc. (Academic Press) 1991.
- [60] B. Petri, N. Schappacher, On Arithmetization. In [31], pp. 343–374.
- [61] J. von Plato, *Creating Modern Probability. Its Mathematics, Physics and Philosophy in Historical Perspective*. Cambridge (CUP) 1994.
- [62] E. Picard, D. Simart, *Théorie des fonctions algébriques de deux variables indépendantes*. 2 vols. Paris (Gauthier-Villars) 1897 / 1906.
- [63] H.-J. Rheinberger, *Experimentalsysteme und epistemische Dinge. Eine Geschichte der Proteinsynthese im Reagenzglas*. Göttingen (Wallstein) 2002.
- [64] P. Roquette, Class field theory in characteristic p , its origin and development. In *Class Field Theory — Its Centenary and Prospect* (K. Miyake, ed.). Tokyo (Mathematical Society Japan) 2001; pp. 549–631.
- [65] N. Schappacher, Politisches in der Mathematik: Versuch einer Spurensicherung. *Mathematische Semesterberichte* 50 (2003), 1–27.
- [66] N. Schappacher, David Hilbert, Report on algebraic number fields (“Zahlbericht”). In *Landmark writings in Western Mathematics* (I. Grattan-Guinness, ed.). Amsterdam, Boston, etc. (Elsevier) 2005; pp. 700–709.
- [67] N. Schappacher, A historical sketch of van der Waerden’s work on Algebraic Geometry 1926–1946. In *Episodes in the History of Algebra* (J. Gray, K. Parshall, eds.), Providence, RI (AMS/LMS) 2007; pp. 245–283.

- [68] N. Schappacher, Seventy years ago: The Bourbaki Congress at El Escorial and other mathematical (non)events of 1936. *Mathematical Intelligencer*, Special Issue ICM Madrid 2006, 8–15. Spanish translation: *La gaceta de la RSME* 11, no. 4 (2008), 721–735.
- [69] N. Schappacher, Bemerkungen zu Wittgensteins Philosophie der Mathematik in ihrer Beziehung zur Mathematikgeschichte. In “*Ein Netz von Normen*”. *Wittgenstein und die Mathematik* (M. Kross, ed.). Berlin (Parerga) 2008; pp. 199–234.
- [70] H.C.H. Schubert, *Kalkül der abzählenden Geometrie*. Reprint of the original 1879 edition with an introduction by Steven L. Kleiman. Berlin, Heidelberg, New York (Springer) 1979.
- [71] F. Severi, *Lezioni di geometria algebrica: geometria sopra una curva: superficie di Riemann; integrali abeliani*. Padova (Draghi) 1908.
- [72] F. Severi, Sul principio della conservazione del numero. *Rendiconti del Circolo Matematico di Palermo* 33 (1912), 313–327.
- [73] F. Severi, *Trattato di geometria algebrica. Volume I, parte I: Geometria delle serie lineari*. Bologna (Zanichelli) 1926.
- [74] F. Severi, La géométrie algébrique italienne. Sa rigueur, ses méthodes, ses problèmes. In *Colloque de géométrie algébrique, Liège 1949*. Liège (Georges Thone) / Paris (Masson) 1950; pp. 9–55.
- [75] R. Siegmund-Schultze, Probability in 1919/1920. The von Mises — Pólya — Controversy. *Archive for History of Exact Sciences* 60 (2006), 431–515.
- [76] S. Slembek, *Weiterentwicklung oder Umbruch? Zu Oscar Zariskis Arithmetisierung der algebraischen Geometrie*. Ph.D. thesis Strasbourg / Mainz 2002.
- [77] S. Slembek, On the Arithmetization of Algebraic Geometry. In *Episodes in the History of Algebra* (J. Gray, K. Parshall, eds.), Providence, RI (AMS/LMS) 2007; pp. 285–300.
- [78] V. Snyder et al., *Selected Topics in Algebraic Geometry. Report of the Committee on Rational Transformations*. Bulletin of the National Research Council, Washington, 63 (1928), 96 (1934).
- [79] E. Steinitz. Algebraische Theorie der Körper. *Journal für die reine und angewandte Mathematik* 137 (1910), 167–309.
- [80] T. Takenouchi, On the relatively abelian corpora with respect to the corpus defined by a primitive cube root of unity. *Journal of the College of Science, Imperial University Tokyo* 37, no. 5 (1916), 70 pp.
- [81] P. Ullmann, *Mathematik — Moderne — Ideologie. Eine kritische Studie zur Legitimität und Praxis der modernen Mathematik*. Konstanz (UVK) 2008.
- [82] H. Weber, Die allgemeinen Grundlagen der Galois’schen Gleichungstheorie. *Mathematische Annalen* 43 (1983), 521–549.
- [83] H. Weber, *Lehrbuch der Algebra*. Vol. III. Braunschweig (Vieweg) 1908.
- [84] B.L. van der Waerden, *Einführung in die Algebraische Geometrie*. Berlin (Springer) 1939.
- [85] A. Weil, *Foundations of Algebraic Geometry*. American Mathematical Society 1946.

-
- [86] A. Weil, *Œuvres scientifiques / Collected Papers*. Vol. I (1926–1951). New York, Heidelberg, Berlin (Springer) 1979.
- [87] A. Weil, *Œuvres scientifiques / Collected Papers*. Vol. II (1951–1964). New York, Heidelberg, Berlin (Springer) 1979.
- [88] H. Weyl, *Die Idee der Riemannschen Fläche*. Leipzig, Berlin (Teubner) 1913.
- [89] H. Weyl, *Raum, Zeit, Materie: Vorlesungen über allgemeine Relativitätstheorie*. Berlin (Springer) 1918.
- [90] H. Weyl, *Das Kontinuum. Kritische Untersuchung über die Grundlagen der Analysis*. Leipzig (Veit & Co) 1918.
- [91] H. Weyl, *Gesammelte Abhandlungen* (K. Chandrasekharan, ed.). Vol. I. Berlin, Heidelberg, New York (Springer) 1968.
- [92] H. Weyl, *Gesammelte Abhandlungen* (K. Chandrasekharan, ed.). Vol. II. Berlin, Heidelberg, New York (Springer) 1968.
- [93] O. Zariski, On a theorem of Severi. *American Journal of Mathematics* 50 (1928), 87–92.
- [94] O. Zariski, *Algebraic Surfaces*. Berlin (Springer) 1935.
- [95] O. Zariski, *Collected Papers*. Vol. I (H. Hironaka, D. Mumford, eds.). Cambridge, MA (MIT Press) 1972.
- [96] O. Zariski, *Collected Papers*. Vol. II (M. Artin, D. Mumford, eds.). Cambridge, MA (MIT Press) 1973.
- [97] O. Zariski, P. Samuel, *Commutative Algebra*. Vol. II. Princeton (Van Nostrand) 1960.
- [98] H.G. Zeuthen, *Lehrbuch der abzählenden Methoden der Geometrie*. Leipzig (Teubner) 1914.

This page is intentionally left blank

Author Index*

(Volumes II, III, and IV)

- Adler, Jill, **IV** 3213
Anantharaman, Nalini, **III** 1839
Arnaud, Marie-Claude, **III** 1653
Auroux, Denis, **II** 917
- Baake, Ellen, **IV** 3037
Balmer, Paul, **II** 85
Belkale, Prakash, **II** 405
Benjamini, Itai, **IV** 2177
Benson, David J., **II** 113
Bernard, Patrick, **III** 1680
Billera, Louis J., **IV** 2389
Borodin, Alexei, **IV** 2188
Bose, Arup, **IV** 2203
Breuil, Christophe, **II** 203
Brydges, David, **IV** 2232
Buff, Xavier, **III** 1701
Bürgisser, Peter, **IV** 2609
- Chen, Shuxing, **III** 1884
Cheng, Chong-Qing, **III** 1714
Chéritat, Arnaud, **III** 1701
Cockburn, Bernardo, **IV** 2749
Cohn, Henry, **IV** 2416
Contreras, Gonzalo, **III** 1729
Costello, Kevin, **II** 942
Csörnyei, Marianna, **III** 1379
- Dancer, E. N., **III** 1901
De Lellis, Camillo, **III** 1910
del Pino, Manuel, **III** 1934
Delbaen, Freddy, **IV** 3054
den Hollander, Frank, **IV** 2258
Dencker, Nils, **III** 1958
- Dwork, Cynthia, **IV** 2634
- Einsiedler, Manfred, **III** 1740
Erschler, Anna, **II** 681
Eskin, Alex, **III** 1185
Evans, Steven N., **IV** 2275
- Fernández, Isabel, **II** 830
Fomin, Sergey, **II** 125
Frankowska, Héléne, **IV** 2915
Fu, Jixiang, **II** 705
Fusco, Nicola, **III** 1985
- Gabai, David, **II** 960
Gaboriau, Damien, **III** 1501
Goldman, William M., **II** 717
Gordon, Iain G., **III** 1209
Greenberg, Ralph, **II** 231
Grodal, Jesper, **II** 973
Guruswami, Venkatesan, **IV** 2648
Guth, Larry, **II** 745
- Hacon, Christopher D., **II** 427, 513
Hamenstädt, Ursula, **II** 1002
Heath-Brown, D.R., **II** 249
Hertz, Federico Rodriguez, **III** 1760
Hutchings, Michael, **II** 1022
Huybrechts, Daniel, **II** 450
- Its, Alexander R., **III** 1395
Ivanov, Sergei, **II** 769
Iwata, Satoru, **IV** 2943
Izumi, Masaki, **III** 1528
- Kaledin, D., **II** 461

*Names of invited speakers only are shown in the Index.

- Kapustin, Anton, **III** 2021
Karpenko, Nikita A., **II** 146
Kedlaya, Kiran Sridhara, **II** 258
Khare, Chandrashekhar, **II** 280
Khot, Subhash, **IV** 2676
Kisin, Mark, **II** 294
Kjeldsen, Tinne Hoff, **IV** 3233
Koskela, Pekka, **III** 1411
Kuijlaars, Arno B.J., **III** 1417
Kumar, Shrawan, **III** 1226
Kunisch, Karl, **IV** 3061
Kupiainen, Antti, **III** 2044
- Lackenby, Marc, **II** 1042
Lando, Sergei K., **IV** 2444
Lapid, Erez M., **III** 1262
Leclerc, Bernard, **IV** 2471
Liu, Chiu-Chu Melissa, **II** 497
Losev, Ivan, **III** 1281
Lück, Wolfgang, **II** 1071
Lurie, Jacob, **II** 1099
- Ma, Xiaonan, **II** 785
Maini, Philip K., **IV** 3091
Marcolli, Matilde, **III** 2057
Markowich, Peter A., **IV** 2776
Marques, Fernando Codá, **II** 811
Martin, Gaven J., **III** 1433
Mastropietro, Vieri, **III** 2078
McKay, Brendan D., **IV** 2489
McKernan, James, **II** 427
McKernan, James, **II** 513
Mira, Pablo, **II** 830
Mirzakhani, Maryam, **II** 1126
Moore, Justin Tatch, **II** 3
Morel, Sophie, **II** 312
- Nabutovsky, Alexander, **II** 862
Nadirashvili, Nikolai, **III** 2001
Naor, Assaf, **III** 1549
Nazarov, Fedor, **III** 1450
Nešetřil, J., **IV** 2502
Nesterov, Yurii, **IV** 2964
Neuhauser, Claudia, **IV** 2297
Nies, André, **II** 30
- Nochetto, Ricardo H., **IV**, 2805
- Oh, Hee, **III** 1308
- Pacard, Frank, **II** 882
Park, Jongil, **II** 1146
Păun, Mihai, **II** 540
Peterzil, Ya'acov, **II** 58
- Quastel, Jeremy, **IV** 2310
- Rains, Eric M., **IV** 2530
Reichstein, Zinovy, **II** 162
Riordan, Oliver, **IV** 2555
Rudelson, Mark, **III** 1576
- Saito, Shuji, **II** 558
Saito, Takeshi, **II** 335
Sarig, Omri M., **III** 1777
Schappacher, Norbert, **IV** 3258
Schreyer, Frank-Olaf, **II** 586
Schütte, Christof, **IV** 3105
Seregin, Gregory A., **III** 2105
Shah, Nimish A., **III** 1332
Shao, Qi-Man, **IV** 2325
Shapiro, Alexander, **IV** 2979
Shen, Zuowei, **IV** 2834
Shlyakhtenko, Dimitri, **III** 1603
Slade, Gordon, **IV** 2232
Sodin, Mikhail, **III** 1450
Soundararajan, K., **II** 357
Spielman, Daniel A., **IV** 2698
Spohn, Herbert, **III** 2128
Srinivas, Vasudevan, **II** 603
Starchenko, Sergei, **II** 58
Stipsicz, András I., **II** 1159
Stroppel, Catharina, **III** 1344
Sudakov, Benny, **IV** 2579
Suresh, V., **II** 189
- Thomas, Richard P., **II** 624
Toro, Tatiana, **III** 1485
Touzi, Nizar, **IV** 3132
Turaev, Dmitry, **III** 1804
- Vadhan, Salil, **IV** 2723

-
- Vaes, Stefaan, **III** 1624
van de Geer, Sara, **IV** 2351
van der Vaart, Aad, **IV** 2370
Venkataramana, T. N., **III** 1366
Venkatesh, Akshay, **II** 383
Vershynin, Roman, **III** 1576
- Weismantel, Robert, **IV** 2996
Welschinger, Jean-Yves, **II** 652
Wendland, Katrin, **III** 2144
- Wheeler, Mary F., **IV** 2864
Wilkinson, Amie, **III** 1816
Wintenberger, Jean-Pierre, **II** 280
- Xu, Jinchao, **IV** 2886
Xu, Zongben, **IV** 3151
- Yamaguchi, Takao, **II** 899
- Zhang, Xu, **IV** 3008
Zhou, Xun Yu, **IV** 3185



World Scientific

www.worldscientific.com

7920 hc

ISBN-13 978-981-4324-34-2

ISBN-10 981-4324-34-5



9 789814 324342

ISBN-13 978-981-4324-30-4 (hard)

ISBN-10 981-4324-30-2 (paper)



9 789814 324304