

# Proceedings of the **International Congress of Mathematicians**

Rio de Janeiro 2018

VOLUME IV  
Invited Lectures

**Boyan Sirakov**  
**Paulo Ney de Souza**  
**Marcelo Viana**  
Editors



Proceedings of the

# **International Congress of Mathematicians**

Rio de Janeiro 2018



Editors

Boyan Sirakov, PUC – Rio de Janeiro

Paulo Ney de Souza, University of California, Berkeley

Marcelo Viana, IMPA – Rio de Janeiro

Technical Editors: Books in Bytes

Proceedings of the International Congress of Mathematicians

August 1 – 9, 2018, Rio de Janeiro, Brazil

Copyright © 2018 by Sociedade Brasileira de Matemática and International Mathematical Union.

Printed in Singapore

All rights reserved. No part of the material protected by the copyright herein may be reproduced or transmitted in any form or by any means, electronic or mechanical, including, but not limited to photocopying, recording, or by any information storage and retrieval system, without express permission from the copyright owner.

**Published and Distributed by**

World Scientific Publishing Co Pte Ltd  
5 Toh Tuck Link  
Singapore 596224

Tel: 65-6466-5775  
Fax: 65-6467-7667  
[www.wspc.com](http://www.wspc.com)  
[sales@wspc.com](mailto:sales@wspc.com)

ISBN: 978-981-3272-93-4 (volume print)

ISBN: 978-981-3272-87-3 (set print)

ISBN: 978-981-3272-88-0 (set ebook)

# Proceedings of the **International Congress of Mathematicians**

Rio de Janeiro 2018

VOLUME IV  
Invited Lectures

**Boyan Sirakov**  
**Paulo Ney de Souza**  
**Marcelo Viana**  
Editors



# Contents

## 1 Logic and Foundations

<b>Matthias Aschenbrenner, Lou van den Dries and Joris van der Hoeven</b>	
On numbers, germs, and transseries	19
<b>Stephen Jackson</b>	
Towards a theory of definable sets	43
<b>Jochen Koenigsmann</b>	
Decidability in local and global fields	63
<b>Ulrich Kohlenbach</b>	
Proof-theoretic methods in nonlinear analysis	79
<b>Maryanthe Malliaris</b>	
Model theory and ultraproducts	101

## 2 Algebra

<b>Christof Geiß</b>	
Quivers with relations for symmetrizable Cartan matrices and algebraic Lie theory	117
<b>Osamu Iyama</b>	
Tilting Cohen–Macaulay representations	143
<b>Moritz Kerz</b>	
On negative algebraic $K$ -groups	181
<b>Sonia Natale</b>	
On the classification of fusion categories	191
<b>Ivan Panin</b>	
On Grothendieck–Serre conjecture concerning principal bundles	219
<b>Pham Huu Tiep</b>	
Representations of finite groups and applications	241

### 3 Number Theory

#### **Fabrizio Andreatta, Adrian Iovita and Vincent Pilloni**

$p$ -adic variation of automorphic sheaves 267

#### **Yves André**

Perfectoid spaces and the homological conjectures 295

#### **Laurent Fargues**

La courbe 309

#### **Kaisa Matomäki and Maksym Radziwiłł**

Multiplicative functions in short intervals, and correlations of  
multiplicative functions 339

#### **James Maynard**

Gaps between primes 363

#### **Ritabrata Munshi**

The subconvexity problem for  $L$ -functions 381

#### **Georgios Pappas**

Arithmetic models for Shimura varieties 395

#### **Bjorn Poonen**

Heuristics for the arithmetic of elliptic curves 417

#### **Jack Thorne**

Potential automorphy of  $\widehat{G}$ -local systems 433

#### **Jacob Tsimerman**

Functional transcendence and arithmetic applications 453

#### **Maryna Viazovska**

Sharp sphere packings 473

#### **Miguel N. Walsh**

Characteristic subsets and the polynomial method 485

#### **Wei Zhang**

Periods, cycles, and  $L$ -functions: A relative trace formula approach 505

## 4 Algebraic and Complex Geometry

### **Dan Abramovich**

- Resolution of singularities of complex algebraic varieties and their families 541

### **Carolina Araujo**

- Positivity and algebraic integrability of holomorphic foliations 565

### **Caucher Birkar**

- Birational geometry of algebraic varieties 583

### **Sébastien Boucksom**

- Variational and non-Archimedean aspects of the Yau–Tian–Donaldson conjecture 609

### **Serge Cantat**

- Automorphisms and dynamics: A list of open problems 637

### **Lucia Caporaso**

- Recursive combinatorial aspects of compactified moduli spaces 653

### **Jungkai A. Chen and Meng Chen**

- On explicit aspect of pluricanonical maps of projective varieties 671

### **Paul Hacking and Sean Keel**

- Mirror symmetry and cluster algebras 689

### **JongHae Keum**

- Algebraic surfaces with minimal Betti numbers 717

### **Wojciech Kucharz and Krzysztof Kurdyka**

- From continuous rational to regulous functions 737

### **András Némethi**

- Pairs of invariants of surface singularities 767

### **Mihnea Popa**

- $\mathcal{D}$ -modules in birational geometry 799

### **Chenyang Xu**

- Interaction between singularity theory and the minimal model program 825

## 5 Geometry

### Nicolas Bergeron

Hodge theory and cycle theory of locally symmetric spaces 849

### Bo Berndtsson

Complex Brunn–Minkowski theory and positivity of vector bundles 877

### Mahan Mj

Cannon–Thurston maps 903

### Emmy Murphy

(No article provided for publication)

### Denis V. Osin

Groups acting acylindrically on hyperbolic spaces 937

### Pedro A. S. Salomão and Umberto L. Hryniewicz

Global surfaces of section for Reeb flows in dimension three and beyond 959

### Ivan Smith

Stability conditions in symplectic topology 987

### Song Sun

Degenerations and moduli spaces in Kähler geometry 1011

### Anna Wienhard

An invitation to higher Teichmüller theory 1031

## 6 Topology

### Arthur Bartels

$K$ -theory and actions on Euclidean retracts 1059

### Tobias Ekholm

Knot contact homology and open Gromov–Witten theory 1081

### Koji Fujiwara

Constructing group actions on quasi-trees 1105

<b>Fanny Kassel</b>	
Geometric structures and representations of discrete groups	1133
<b>Robert Lipshitz and Sucharit Sarkar</b>	
Spatial refinements and Khovanov homology	1171
<b>Ciprian Manolescu</b>	
Homology cobordism and triangulations	1193
<b>John Pardon</b>	
(No article provided for publication)	
<b>Alan W. Reid</b>	
Profinite rigidity	1211
<b>Bernardo Uribe</b>	
The evenness conjecture in equivariant unitary bordism	1235
<b>Thomas Willwacher</b>	
Little disks operads and Feynman diagrams	1259
<b>7 Lie Theory and Generalizations</b>	
<b>Tomoyuki Arakawa</b>	
Representation theory of W-algebras and Higgs branch conjecture	1281
<b>Michael Finkelberg</b>	
Double affine Grassmannians and Coulomb branches of $3d \mathcal{N} = 4$ quiver gauge theories	1301
<b>Vyacheslav Futorny</b>	
Representations of Galois algebras	1321
<b>Tsachik Gelander</b>	
A view on invariant random subgroups and lattices	1339
<b>Xuhua He</b>	
Some results on affine Deligne–Lusztig varieties	1363
<b>Dipendra Prasad</b>	
Ext-analogues of Branching laws	1385



**Olivier G. Schiffmann**

Kac polynomials and Lie algebras associated to quivers and curves 1411

**Akshay Venkatesh**

(No article provided for publication)

**Eva Viehmann**

Moduli spaces of local  $\mathbf{G}$ -shtukas 1443

**Zhiwei Yun**

Hitchin type moduli stacks in automorphic representation theory 1465

**8 Analysis and Operator Algebras****Spiros A. Argyros and Richard G. Haydon**

Bourgain–Delbaen  $\mathcal{L}_\infty$ -spaces, the scalar-plus-compact property and related problems 1495

**Christopher J. Bishop**

Harmonic measure: Algorithms and applications 1529

**Ciprian Demeter**

Decouplings and applications 1557

**Tien-Cuong Dinh**

Pluripotential theory and complex dynamics in higher dimension 1579

**Ruy Exel and Benjamin Steinberg**

The inverse hull of 0-left cancellative semigroups 1601

**Mouhamed Moustapha Fall**

Constant nonlocal mean curvatures surfaces and related problems 1631

**Adrian Ioana**

Rigidity for von Neumann algebras 1657

**William B. Johnson**

Some 20+ year old problems about Banach spaces and operators on them 1691

**Svitlana Mayboroda**

The effect of disorder and irregularities on solutions to boundary value problems  
and spectra of differential operators 1709

**András Máthé**

Measurable equidecompositions 1731

**Stefanie Petermichl**

Higher order commutators and multi-parameter BMO 1751

**Alexei Poltoratski**

Toeplitz methods in completeness and spectral problems 1771

**Andreas Thom**

Finitary approximations of groups and their applications 1797

**Wilhelm Winter**

Structure of nuclear  $C^*$ -algebras: From quasidiagonality to classification and  
back again 1819

**9 Dynamical Systems and Ordinary Differential Equations****Jairo Bochi**

Ergodic optimization of Birkhoff averages and Lyapunov exponents 1843

**Lewis P. Bowen**

A brief introduction to sofic entropy theory 1865

**Laura DeMarco**

Critical orbits and arithmetic equidistribution 1885

**Lorenzo J. Díaz**

Nonhyperbolic ergodic measures 1905

**Bassam Fayad and Raphaël Krikorian**

Some questions around quasi-periodic dynamics 1927

**Sébastien Gouëzel**

Subadditive cocycles and horofunctions 1951

<b>Michael Hochman</b>	
Dimension theory of self-similar sets and measures	1967
<b>Konstantin Khanin</b>	
Renormalization and rigidity	1991
<b>Andres Korozecki and Meysam Nassiri</b>	
Boundary dynamics for surface homeomorphisms	2013
<b>Martin Möller</b>	
Geometry of Teichmüller curves	2035
<b>Andrés Navas</b>	
Group actions on 1-manifolds: A list of very concrete open questions	2053
<b>Rafael Potrie</b>	
Robust dynamics, invariant structures and topological classification	2081
<b>Feliks Przytycki</b>	
Thermodynamic formalism methods in one-dimensional real and complex dynamics	2105
<b>Jiangong You</b>	
Quantitative almost reducibility and its applications	2131
<b>10 Partial Differential Equations</b>	
<b>Jacob Bedrossian, Yu Deng and Nader Masmoudi</b>	
The Orr mechanism: Stability/Instability of the Couette flow for the 2D Euler dynamic	2155
<b>Didier Bresch and Pierre-Emmanuel Jabin</b>	
Quantitative estimates for Advective Equation with degenerate anelastic constraint	2185
<b>Diego Córdoba</b>	
Interface dynamics for incompressible fluids: Splash and Splat singularities	2211
<b>Guido De Philippis and Filip Rindler</b>	
On the structure of measures constrained by linear PDEs	2233

**Jean-Marc Delort**

Long time existence results for solutions of water waves equations 2259

**Jean Dolbeault, Maria J. Esteban and Michael Loss**

Symmetry and symmetry breaking: Rigidity and flows in elliptic PDEs 2279

**Yoshikazu Giga**

On large time behavior of growth by birth and spread 2305

**Massimiliano Gubinelli**

A panorama of singular SPDEs 2329

**Colin Guillarmou**

Analytic tools for the study of flows and inverse problems 2357

**Alexander Kiselev**

Small scales and singularity formation in fluid dynamics 2381

**Alexander Logunov and Eugenia Malinnikova**

Quantitative propagation of smallness for solutions of elliptic equations 2409

**Mohamed Majdoub and Slim Tayachi**

Well-posedness, global existence and decay estimates for the heat equation with general power-exponential nonlinearities 2431

**Yvan Martel**

Interaction of solitons from the PDE point of view 2457

**Clément Mouhot**

De Giorgi–Nash–Moser and Hörmander theories: New interplays 2485

**Stéphane Nonnenmacher**

Resonances in hyperbolic dynamics 2513

**Helena J. Nussenzweig Lopes and Milton C. Lopes Filho**

Fluids, walls and vanishing viscosity 2537

**11 Mathematical Physics****Jørgen Ellegaard Andersen and Rinat Kashaev**

The Teichmüller TQFT 2559

**Alexander Belavin**

Special geometry on Calabi–Yau moduli spaces and  $Q$ -invariant  
Milnor rings 2585

**Philippe Di Francesco**

Integrable combinatorics 2599

**Yasuyuki Kawahigashi**

Conformal field theory, vertex operator algebras and operator algebras 2615

**Claudio Landim**

Variational formulae for the capacity induced by second-order elliptic  
differential operators 2635

**Carlangelo Liverani**

Transport in partially hyperbolic fast-slow systems 2661

**Benjamin Schlein**

Bogoliubov excitation spectrum for Bose–Einstein condensates 2687

**Mariya Shcherbina and Tatyana Shcherbina**

Transfer operator approach to 1d random band matrices 2705

**Yuji Tachikawa**

On ‘categories’ of quantum field theories 2727

**Fabio Toninelli**

$(2 + 1)$ -dimensional interface dynamics: Mixing time, hydrodynamic  
limit and anisotropic KPZ growth 2751

**Simone Warzel**

(No article provided for publication)

**12 Probability and Statistics****Paul Bourgade**

Random band matrices 2777

**Peter Bühlmann**

Invariance in heterogeneous, large-scale and high-dimensional data 2803

**Dmitry Chelkak**

Planar Ising model at criticality: State-of-the-art and perspectives 2819

**Hugo Duminil-Copin**

Sixty years of percolation 2847

**Noureddine El Karoui**

Random matrices and high-dimensional statistics: Beyond covariance matrices 2875

**Josselin Garnier**

Multiscale analysis of wave propagation in random media 2895

**Vladimir Koltchinskii**

Asymptotic efficiency in high-dimensional covariance estimation 2921

**Can M. Le, Elizaveta Levina and Roman Vershynin**

Concentration of random graphs and application to community detection 2943

**Jason Miller**

Liouville quantum gravity as a metric space and a scaling limit 2963

**Andrea Montanari**

Mean field asymptotics in high-dimensional statistics: From exact results to efficient algorithms 2991

**Byeong U. Park**

Nonparametric additive regression 3013

**Allan Sly**

(No article provided for publication)

**Jonathan E. Taylor**

A selective survey of selective inference 3037

**Bálint Tóth**

Diffusive and super-diffusive limits for random walks and diffusions with long memory 3057

## 13 Combinatorics

**József Balogh, Robert Morris and Wojciech Samotij**

The method of hypergraph containers 3077

**June Huh**

Combinatorial applications of the Hodge–Riemann relations 3111

**Peter Keevash**

Hypergraph matchings and designs 3131

**Richard Kenyon**

Limit shapes and their analytic parameterizations 3155

**Igor Pak**

Complexity problems in enumerative combinatorics 3171

**Alexander Postnikov**

Positive Grassmannian and polyhedral subdivisions 3199

**Balázs Szegedy**

From graph limits to higher order Fourier analysis 3231

**Gábor Tardos**

Extremal theory of ordered graphs 3253

**Nicholas Wormald**

Asymptotic enumeration of graphs with given degree sequence 3263

## 14 Mathematical Aspects of Computer Science

**Andris Ambainis**

Understanding quantum algorithms via query complexity 3283

**Alexandr Andoni, Piotr Indyk and Ilya Razenshteyn**

Approximate nearest neighbor search in high dimensions 3305

**László Babai**

Group, graphs, algorithms: The Graph Isomorphism Problem 3337

**Yael Tauman Kalai**

Delegating computation via no-signaling strategies 3355

**Neeraj Kayal**

(No article provided for publication)

**Aleksander Mądry**

Gradients and flows: Continuous optimization approaches to the  
Maximum Flow Problem 3379

**Prasad Raghavendra, Tselil Schramm and David Steurer**

High dimensional estimation via Sum-of-Squares Proofs 3407

**Benjamin Rossman**

Lower bounds for subgraph isomorphism 3443

**Virginia Vassilevska Williams**

On some fine-grained questions in algorithms and complexity 3465

**15 Numerical Analysis and Scientific Computing****Raimund Bürger, Julio Careaga, Stefan Diehl, Camilo Mejías and Ricardo Ruiz Baier**

Convection-diffusion-reaction and transport-flow problems motivated by  
models of sedimentation: Some recent advances 3507

**Manuel J. Castro, Marc de la Asunción, Enrique D. Fernández Nieto, José M. Gallardo, José M. González Vida, Jorge Macías, Tomás Morales, Sergio Ortega and Carlos Parés**

A review on high order well-balanced path-conservative finite volume  
schemes for geophysical flows 3533

**Qiang Du**

An invitation to nonlocal modeling, analysis and computation 3559

**Michael B. Giles**

An introduction to multilevel Monte Carlo methods 3589

**Kai Jiang and Pingwen Zhang**

Numerical mathematics of quasicrystals 3609



**Shi Jin**

- Mathematical analysis and numerical methods for multiscale kinetic equations with uncertainties 3629

**Siddhartha Mishra**

- On the convergence of numerical schemes for hyperbolic systems of conservation laws 3659

**Tao Tang**

- On effective numerical methods for phase-field models 3687

**Anna-Karin Tornberg**

- FFT based spectral Ewald methods as an alternative to fast multipole methods 3709

**Barbara Wohlmuth**

- (No article provided for publication)

**16 Control Theory and Optimization****Coralia Cartis, Nicholas I. M. Gould and Philippe L. Toint**

- Worst-case evaluation complexity and optimality of second-order methods for nonconvex smooth optimization 3729

**Matti Lassas**

- Inverse problems for linear and non-linear hyperbolic equations 3769

**Jean B. Lasserre**

- The moment-SOS hierarchy 3791

**Claudia Sagastizábal**

- A  $\mathcal{V}\mathcal{U}$ -point of view of nonsmooth optimization 3815

**Rekha R. Thomas**

- Spectrahedral lifts of convex sets 3837

**Emmanuel Trélat**

- Optimal shape and location of sensors or actuators in PDE models 3861

## 17 Mathematics in Science and Technology

### **Andrea L. Bertozzi**

Graphical models in machine learning, networks and uncertainty quantification 3883

### **Mark van der Boor, Sem C. Borst, Johan S. H. van Leeuwen and Debankur Mukherjee**

Scalable load balancing in networked systems: Universality properties and stochastic coupling methods 3911

### **Pierre Degond**

Mathematical models of collective dynamics and self-organization 3943

### **Selim Esedoğlu**

Algorithms for motion of networks by weighted mean curvature 3965

### **Richard D. James**

Symmetry, invariance and the structure of matter 3985

### **Amit Singer**

Mathematics for cryo-electron microscopy 4013

## 18 Mathematics Education and Popularization of Mathematics

### **Marianna Bosch Casabò**

Study and research paths: A model for inquiry 4033

### **Luis Radford**

On theories in mathematics education and their conceptual differences 4055

## 19 History of Mathematics

### **Jan von Plato**

In search of the sources of incompleteness 4075

### **Tatiana Roque**

IMPA's coming of age in a context of international reconfiguration of mathematics 4093

**David E. Rowe**

On Franco–German relations in mathematics, 1870–1920

4113

# RANDOM BAND MATRICES

PAUL BOURGADE

## Abstract

We survey recent mathematical results about the spectrum of random band matrices. We start by exposing the Erdős–Schlein–Yau dynamic approach, its application to Wigner matrices, and extension to other mean-field models. We then introduce random band matrices and the problem of their Anderson transition. We finally expose a method to obtain delocalization and universality in some sparse regimes, highlighting the role of quantum unique ergodicity.

## Contents

<b>1</b>	<b>Introduction</b>	<b>2777</b>
<b>2</b>	<b>Mean-field random matrices</b>	<b>2780</b>
<b>3</b>	<b>Random band matrices and the Anderson transition</b>	<b>2787</b>
<b>4</b>	<b>Quantum unique ergodicity and universality</b>	<b>2793</b>

## 1 Introduction

This note explains the interplay between eigenvectors and eigenvalues statistics in random matrix theory, when the considered models are not of mean-field type, meaning that the interaction is short range and geometric constraints enter in the definition of the model.

If the range or strength of the interaction is small enough, it is expected that eigenvalues statistics will fall into the *Poisson universality class*, intimately related to the notion of independence. Another class emerged in the past fifty years for many correlated systems,

---

This work is supported by the NSF grant DMS#1513587.

MSC2010: 15B52.

Keywords: band matrices, delocalization, quantum unique ergodicity, Gaussian free field.

initially from calculations on random linear operators. This *random matrix universality class* was proposed by Wigner [1957], first as a model for stable energy levels of typical heavy nuclei. The models he introduced have since been understood to connect to integrable systems, growth models, analytic number theory and multivariate statistics (see e.g. Deift [2017]).

Ongoing efforts to understand universality classes are essentially of two types. First, *integrability* consists in finding possibly new statistics for a few, rigid models, with methods including combinatorics and representation theory. Second, *universality* means enlarging the range of models with random matrix statistics, through probabilistic methods. For example, the Gaussian random matrix ensembles are mean-field integrable models, from which local spectral statistics can be established for the more general Wigner matrices, by comparison, as explained in Section 1. For random operators with shorter range, no integrable models are known, presenting a major difficulty in understanding whether their spectral statistics will fall in the Poisson or random matrix class.

In Wigner's original theory, the eigenvectors play no role. However, their statistics are essential in view of a famous dichotomy of spectral behaviors, widely studied since Anderson's tight binding model P. Anderson [1958]:

- (i) Poisson spectral statistics usually occur together with localized eigenstates,
- (ii) random matrix eigenvalue distributions should coincide with delocalization of eigenstates.

The existence of the localized phase has been established for the Anderson model in any dimension Fröhlich and Spencer [1983], but delocalization has remained elusive for all operators relevant in physics. An important question consists in proving extended states and GOE local statistics for one such model<sup>1</sup>, giving theoretical evidence for conduction in solids. How localization implies Poisson statistics is well understood, at least for the Anderson model Minami [1996]. In this note, we explain the proof of a strong notion of delocalization (quantum unique ergodicity), and how it implies random matrix spectral statistics, for the  $1d$  random band matrix (RBM) model.

In this model, vertices are elements of  $\Lambda = \llbracket 1, N \rrbracket^d$  ( $d = 1, 2, 3$ ) and  $H = (H_{ij})_{i,j \in \Lambda}$  have centered real entries, independent up to the symmetry  $H_{ij} = H_{ji}$ . The band width  $W < N/2$  means

$$(1-1) \quad H_{ij} = 0 \text{ if } |i - j| > W,$$

where  $|\cdot|$  is the periodic  $L^1$  distance on  $\Lambda$ , and all non-trivial  $H_{ij}$ 's have a variance  $\sigma_{ij}^2$  with the same order of magnitude, normalized by  $\sum_j \sigma_{ij}^2 = 1$  for any  $i \in \Lambda$ . Mean-field

---

<sup>1</sup>GOE eigenvalues statistics appear in Trotter's tridiagonal model Trotter [1984], which is clearly local, but the entries need varying variance adjusted to a specific profile.

models correspond to  $W = N/2$ . When  $W \rightarrow \infty$ , the empirical spectral measure of  $H$  converges to the semicircle distribution  $d\rho_{\text{sc}}(x) = \frac{1}{2\pi}(4 - x^2)^{1/2}dx$ .

It has been conjectured that the random band matrix model exhibits the localization-delocalization (and Poisson-GOE) transition at some critical band width  $W_c(N)$  for eigenvalues in the bulk of the spectrum  $|E| < 2 - \kappa$ . The localized regime supposedly occurs for  $W \ll W_c$  and delocalization for  $W \gg W_c$ , where

$$(1-2) \quad W_c = \begin{cases} N^{1/2} & \text{for } d = 1, \\ (\log N)^{1/2} & \text{for } d = 2, \\ O(1) & \text{for } d = 3. \end{cases}$$

This transition corresponds to localization length  $\ell \approx W^2$  in dimension 1,  $\ell \approx e^{W^2}$  in dimension 2.

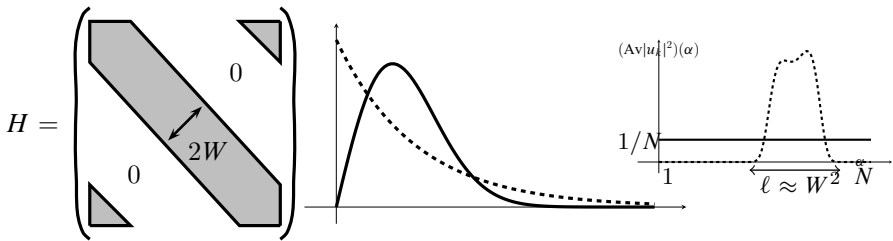


Figure 1: Conjectural behavior of the RBM model for  $d = 1$ . For any eigenvalue  $|\lambda_k| < 2 - \kappa$ , the rescaled gap  $N\rho_{\text{sc}}(\lambda_k)(\lambda_{k+1} - \lambda_k)$  converges to an exponential random variable for  $W \ll N^{1/2}$ , and the Gaudin GOE distribution for  $W \gg N^{1/2}$ . The associated eigenvector  $u_k$  is localized on  $\ell \approx W^2$  sites for  $W \ll N^{1/2}$ , it is flat for  $W \gg N^{1/2}$ . Here  $(\text{Av} f)(\alpha) = (2n)^{-1} \sum_{|i-\alpha| < n} f(i)$  where  $1 \ll n \ll W^2$  is some averaging scale.

This review first explains universality techniques for mean-field models. We then state recent progress in proving the existence of the delocalized phase for the random band matrix model for  $d = 1$ , explaining how quantum unique ergodicity is proved by dynamics. We finally explain, at the heuristic level, a connection between quantum unique ergodicity for band matrices and the Gaussian free field, our main goal being to convince the reader that the transition exponents in (1-2) are natural.

For the sake of conciseness, we only consider the orthogonal symmetry class corresponding to random symmetric matrices with real entries. Analogous results hold in the complex Hermitian class.

## 2 Mean-field random matrices

**2.1 Integrable model.** The Gaussian orthogonal ensemble (GOE) consists in the probability density

$$(2-1) \quad \frac{1}{Z_N} e^{-\frac{N}{4} \text{Tr}(H^2)}$$

with respect to the Lebesgue measure on the set of  $N \times N$  symmetric matrices. This corresponds to all entries being Gaussian and independent up to the symmetry condition, with off-diagonal entries  $H_{ij} \sim N^{-1/2} \mathcal{N}(0, 1)$ , and diagonal entries  $H_{ii} \sim (N/2)^{-1/2} \mathcal{N}(0, 1)$ .

Our normalization is chosen so that the eigenvalues  $\lambda_1 \leq \dots \leq \lambda_N$  (with associated eigenvectors  $u_1, \dots, u_N$ ) have a converging empirical measure:  $\frac{1}{N} \sum_{k=1}^N \delta_{\lambda_i} \rightarrow d\rho_{\text{sc}}$  almost surely. A more detailed description of the spectrum holds at the microscopic scale, in the bulk and at the edge: there exists a translation invariant point process  $\chi_1$  [Mehta and Gaudin \[1960\]](#) and a distribution  $\text{TW}_1$  (for [Tracy and Widom \[1994\]](#)) such that

$$(2-2) \quad \sum_{k=1}^N \delta_{N\rho_{\text{sc}}(E)(\lambda_k - E)} \rightarrow \chi_1,$$

$$(2-3) \quad N^{2/3}(\lambda_N - 2) \rightarrow \text{TW}_1,$$

in distribution. Note that  $\chi_1$  is independent of  $E \in (-2 + \kappa, 2 - \kappa)$ .

Concerning the eigenvectors, for any  $\mathcal{O} \in O(N)$ , from (2-1) the distributions of  $\mathcal{O}^t H \mathcal{O}$  and  $H$  are the same, so that the eigenbasis  $\mathbf{u} = (u_1, \dots, u_N)$  of  $H$  is Haar-distributed (modulo a sign choice) on  $O(N)$ :  $\mathcal{O}\mathbf{u}$  has same distribution as  $\mathbf{u}$ . In particular, any  $u_k$  is uniform of the sphere  $\mathcal{S}^{(N-1)}$ , and has the same distribution as  $\mathcal{N}/\|\mathcal{N}\|_2$  where  $\mathcal{N}$  is a centered Gaussian vector with covariance  $\text{Id}_N$ . This implies that for any deterministic sequences of indices  $k_N \in \llbracket 1, N \rrbracket$  and unit vectors  $\mathbf{q}_N \in \mathcal{S}^{(N-1)}$  (abbreviated  $k, \mathbf{q}$ ), the limiting *Borel-Lévy law* holds:

$$(2-4) \quad N^{1/2} \langle u_k, \mathbf{q} \rangle \rightarrow \mathcal{N}(0, 1)$$

in distribution. This microscopic behavior can be extended to several projections being jointly Gaussian.

The fact that eigenvectors are extended can be formulated with less precision, and quantified in different manners. For example, for the GOE model, for any small  $\varepsilon > 0$  and large  $D > 0$ , we have

$$(2-5) \quad \mathbb{P} \left( \|u_k\|_\infty \geq \frac{N^\varepsilon}{\sqrt{N}} \right) \leq N^{-D},$$

which we refer to as *delocalization* (the above  $N^\varepsilon$  can also be replaced by some logarithmic power).

Delocalization does not imply that the eigenvectors are flat in the sense of Figure 1, as  $u_k$  could be supported on a small fraction of  $\llbracket 1, N \rrbracket$ . A strong notion of flat eigenstates was introduced by Rudnick and Sarnak [1994] for Riemannian manifolds: for any negatively curved and compact  $\mathfrak{M}$  with volume measure  $\mu$ ,

$$(2-6) \quad \int_A |\psi_k(x)|^2 \mu(dx) \xrightarrow{k \rightarrow \infty} \int_A \mu(dx),$$

for any  $A \subset \mathfrak{M}$ . Here  $\psi_k$  is an eigenfunction (associated to the eigenvalue  $\lambda_k$ ) of the Laplace-Beltrami operator,  $0 \leq \lambda_1 \leq \dots \leq \lambda_k \leq \dots$  and  $\|\psi_k\|_{L^2(\mu)} = 1$ . This *quantum unique ergodicity* (QUE) notion strengthens the quantum ergodicity from Shnirel'man [1974], Colin de Verdière [1985], and Zelditch [1987], defined by an additional averaging on  $k$  and proved for a wide class of manifolds and deterministic regular graphs Anantharaman and Le Masson [2015] (see also Brooks and Lindenstrauss [2013]). QUE was rigorously proved for arithmetic surfaces, Lindenstrauss [2006], Holowsinsky [2010], and Holowsinsky and Soundararajan [2010]. We will consider a probabilistic version of QUE at a local scale, for eigenvalues in the bulk of the spectrum ( $\kappa \geq 0$  is small and fixed). By simple properties of the uniform measure on the unit sphere it is clear that the following version holds for the GOE: for any given (small)  $\varepsilon > 0$  and (large)  $D > 0$ , for  $N \geq N_0(\varepsilon, D)$ , for any deterministic sequences  $k_N \in \llbracket \kappa N, (1 - \kappa)N \rrbracket$  and  $I_N \subset \llbracket 1, N \rrbracket$  (abbreviated  $k, I$ ),  $|I_N| > n$ , we have

$$(2-7) \quad \mathbb{P} \left( \left| \sum_{\alpha \in I} (u_k(\alpha))^2 - \frac{1}{N} \right| > \frac{N^\varepsilon |I|^{1/2}}{N} \right) \leq N^{-D}.$$

We now consider the properties (2-2), (2-3) (2-4), (2-5), (3-3) for the following general model.

**Definition 2.1** (Generalized Wigner matrices). *A sequence  $H_N$  (abbreviated  $H$ ) of real symmetric centered random matrices is a generalized Wigner matrix if there exists  $C, c > 0$  such that  $\sigma_{ij}^2 := \text{Var}(H_{ij})$  satisfies*

$$(2-8) \quad c \leq N \sigma_{ij}^2 \leq C \text{ for all } N, i, j \text{ and } \sum_j \sigma_{ij}^2 = 1 \text{ for all } i.$$

We also assume subgaussian decay of the distribution of  $\sqrt{N} H_{ij}$ , uniformly in  $i, j, N$ , for convenience (this could be replaced by a finite high moment assumption).



**2.2 Eigenvalues universality.** The second constraint in (2-8) imposes the macroscopic behavior of the limiting spectral measure:  $\frac{1}{N} \sum_{k=1}^N \delta_{\lambda_i} \rightarrow d\rho_{\text{sc}}$  for all generalized Wigner matrices. This convergence to the semicircle distribution was strengthened up to optimal polynomial scale, thanks to an advanced diagrammatic analysis of the resolvent of  $H$ .

**Theorem 2.2** (Rigidity of the spectrum Erdős, Yau, and Yin [2012b]). *Let  $H$  be a generalized Wigner matrix as in Definition 2.1. Define  $\hat{k} = \min(k, N+1-k)$  and  $\gamma_k$  implicitly by  $\int_{-\gamma_k}^{\gamma_k} d\rho_{\text{sc}} = \frac{k}{N}$ . Then for any  $\varepsilon > 0$ ,  $D > 0$  there exists  $N_0$  such that for  $N > N_0$  we have*

$$(2-9) \quad \mathbb{P} \left( |\lambda_k - \gamma_k| > N^{-\frac{2}{3}+\varepsilon} (\hat{k})^{-\frac{1}{3}} \right) \leq N^{-D}.$$

Given the above scale of fluctuations, a natural problem consists in the limiting distribution. In particular, the (Wigner-Dyson-Mehta) conjecture states that (2-2) holds for random matrices way beyond the integrable GOE class. It has been proved in a series of works in the past years, with important new techniques based on the Harish-Chandra-Itzykson-Zuber integral Johansson [2001] (in the special case of Hermitian symmetry class), the dynamic interpolation through Dyson Brownian motion Erdős, Schlein, and Yau [2011] and the Lindeberg exchange principle Tao and Vu [2011]. The initial universality statements for general classes required an averaging over the energy level  $E$  Erdős, Schlein, and Yau [2011] or the first four moments of the matrix entries to match the Gaussian ones Tao and Vu [2011].

We aim at explaining the dynamic method which was applied in a remarkable variety of settings. For example, GOE local eigenvalues statistics holds for generalized Wigner matrices.

**Theorem 2.3** (Fixed energy universality Bourgade, Erdős, Yau, and Yin [2016]). *The convergence (2-2) holds for generalized Wigner matrices.*

The key idea for the proof, from Erdős, Schlein, and Yau [2011], is interpolation through matrix Dyson Brownian motion (or its Ornstein Uhlenbeck version)

$$(2-10) \quad dH_t = \frac{1}{\sqrt{N}} dB_t - \frac{1}{2} H_t dt$$

with initial condition  $H_0 = H$ , where  $(B_{ij})_{i < j}$  and  $(B_{ii}/\sqrt{2})_i$  are independent standard Brownian motions. The GOE measure (2-1) is the equilibrium for these dynamics. The proof proceeds in two steps, in which the dynamics  $(H_t)_{t \geq 0}$  is analyzed through complementary viewpoints. One relies on the repulsive eigenvalues dynamics, the other on the matrix structure. Both steps require some a priori knowledge on eigenvalues density, such as Theorem 2.2.

*First step: relaxation.* For any  $t \geq N^{-1+\varepsilon}$ , (2-2) holds:  $\sum_{k=1}^N \delta_{N\rho_{\text{sc}}(E)(\lambda_k(t)-E)} \rightarrow \chi_1$ , where we denote  $\lambda_1(t) \leq \dots \leq \lambda_k(t)$  the eigenvalues of  $H_t$ . The proof relies on the Dyson Brownian motion for the eigenvalues dynamics [Dyson \[1962\]](#), given by

$$(2-11) \quad d\lambda_k(t) = \frac{d\tilde{B}_k(t)}{\sqrt{N}} + \left( \frac{1}{N} \sum_{\ell \neq k} \frac{1}{\lambda_k(t) - \lambda_\ell(t)} - \frac{1}{2} \lambda_k(t) \right) dt$$

where the  $\tilde{B}_k/\sqrt{2}$ 's are standard Brownian motions. Consider the dynamics with a different initial condition  $x_1(0) \leq \dots \leq x_N(0)$  given by the eigenvalues of a GOE matrix, and by taking the difference between these two equations we observe that  $\delta_\ell(t) := e^{t/2}(x_\ell(t) - \lambda_\ell(t))$  satisfy an integral equation of parabolic type [Bourgade, Erdős, Yau, and Yin \[2016\]](#), namely

$$(2-12) \quad \partial_t \delta_\ell(t) = \sum_{k \neq \ell} b_{k\ell}(t)(\delta_k(t) - \delta_\ell(t)), \quad b_{k\ell}(t) = \frac{1}{N(x_\ell(t) - x_k(t))(\lambda_\ell(t) - \lambda_k(t))}.$$

From [Theorem 2.2](#), in the bulk of the spectrum we expect that  $b_{k\ell}(t) \approx N/(k - \ell)^2$ , so that Hölder regularity holds for  $t \gg N^{-1}$ :  $\delta_k(t) = \delta_{k+1}(t)(1 + o(1))$ , meaning  $\lambda_{k+1}(t) - \lambda_k(t) = y_{k+1}(t) - y_k(t) + o(N^{-1})$ . Gaps between the  $\lambda_k$ 's and  $x_k$ 's therefore become identical, hence equal to the GOE gaps as the law of  $y_{k+1}(t) - y_k(t)$  is invariant in time. In fact, an equation similar to (2-12) previously appeared in the first proof of GOE gap statistics for generalized Wigner matrices [Erdős and Yau \[2015\]](#), emerging from a Helffer-Sjöstrand representation instead of a probabilistic coupling. [Theorem 2.3](#) requires a much more precise analysis of (2-12) [Bourgade, Erdős, Yau, and Yin \[2016\]](#) and [Landon, Sosoë, and Yau \[2016\]](#), but the conceptual picture is clear from the above probabilistic coupling of eigenvalues.

Relaxation after a short time can also be understood by functional inequalities for relative entropy [Erdős, Schlein, and Yau \[2011\]](#) and [Erdős, Yau, and Yin \[2012a\]](#), a robust method which also gives GOE statistics when averaging over the energy level  $E$ . In the special case of the Hermitian symmetry class, relaxation also follows from explicit formulas for the eigenvalues density at time  $t$  [Johansson \[2001\]](#), [Erdős, Péché, Ramírez, Schlein, and Yau \[2010\]](#), and [Tao and Vu \[2011\]](#).

*Second step: density.* For any  $t \leq N^{-\frac{1}{2}-\varepsilon}$ ,

$$\sum_{k=1}^N \delta_{N\rho_{\text{sc}}(E)(\lambda_k(t)-E)} \quad \text{and} \quad \sum_{k=1}^N \delta_{N\rho_{\text{sc}}(E)(\lambda_k(0)-E)}$$

have the same distribution at leading order. This step can be proved by a simple Itô lemma based on the matrix evolution [Bourgade and Yau \[2017\]](#), which takes a particularly simple

form for Wigner matrices (i.e.  $\sigma_{ij}^2 = N^{-1} + N^{-1}\mathbb{1}_{i=j}$ ). It essentially states that for any smooth function  $F(H)$  we have

$$(2-13) \quad \mathbb{E}F(H_t) - \mathbb{E}F(H_0) = O(tN^{1/2})$$

$$\sup_{i \leq j, 0 \leq s \leq t} \mathbb{E} \left( (N^{3/2}|H_{ij}(s)|^3 + \sqrt{N}|H_{ij}(s)|) |\partial_{ij}^3 F(H_s)| \right)$$

where  $\partial_{ij} = \partial_{H_{ij}}$ . In particular, if  $F$  is stable in the sense that  $\partial_{ij}^3 F = O(N^\varepsilon)$  with high probability (this is known for functions encoding the microscopic behavior thanks to the a-priori rigidity estimates from [Theorem 2.2](#)), then invariance of local statistics holds up to time  $N^{-\frac{1}{2}-\varepsilon}$ .

Invariance of local spectral statistics in matrix neighborhoods has also been proved by other methods, for example by a reverse heat flow when the entries have a smooth enough density [Erdős, Schlein, and Yau \[2011\]](#), or the Lindeberg exchange principle [Tao and Vu \[2011\]](#) for matrices with moments of the entries coinciding up to fourth moment.

**2.3 Eigenvectors universality.** Eigenvalues rigidity (2-9) was an important estimate for the proof of [Theorem 2.3](#). Similarly, to understand the eigenvectors distribution, one needs to first identify their natural fluctuation scale. By analysis of the resolvent of  $H$ , the following was first proved when  $\mathbf{q}$  is an element from the canonical basis [Erdős, Schlein, and Yau \[2009\]](#) and [Erdős, Yau, and Yin \[2012b\]](#), and extended to any direction.

**Theorem 2.4** (Isotropic delocalization [Knowles and Yin \[2013b\]](#) and [Bloemendal, Erdős, Knowles, Yau, and Yin \[2014\]](#)). *For any sequence of generalized Wigner matrices,  $\varepsilon, D > 0$ , there exists  $N_0(\varepsilon, D)$  such that for any  $N \geq N_0$  deterministic  $k$  and unit vector  $\mathbf{q}$ ,*

$$\mathbb{P} \left( \langle u_k, \mathbf{q} \rangle \geq N^{-\frac{1}{2}+\varepsilon} \right) \leq N^{-D}.$$

The more precise fluctuations (2-4) were proved by the Lindeberg exchange principle in [Knowles and Yin \[2013a\]](#) and [Tao and Vu \[2012\]](#), under the assumption of the first four (resp. two) moments of  $H$  matching the Gaussian ones for eigenvectors associated to the spectral bulk (resp. edge). This Lévy-Borel law holds without these moment matching assumptions, and some form of quantum unique ergodicity comes with it.

**Theorem 2.5** (Eigenvectors universality and weak QUE [Bourgade and Yau \[2017\]](#)). *For any sequence of generalized Wigner matrices, and any deterministic  $k$  and unit vector  $\mathbf{q}$ , the convergence (2-4) is true.*

*Moreover, for any  $\varepsilon > 0$  there exists  $D > 0$  such that (3-3) holds.*

The above statement is a *weak* form of QUE, holding for some small  $D = D(\varepsilon)$  although it should be true for any large  $D > 0$ . Section 3 will show a strong form of QUE for some band matrices.

The proof of [Theorem 2.5](#) follows the dynamic idea already described for eigenvalues, by considering the evolution of eigenvectors through (2-10). The *density* step is similar: with (2-13) one can show that the distribution of  $\sqrt{N}\langle u_k(t), \mathbf{q} \rangle$  is invariant up to time  $t \leq N^{-\frac{1}{2}-\varepsilon}$ . The *relaxation step* is significantly different from the coupling argument described previously. The eigenvectors dynamics are given by

$$du_k = \frac{1}{\sqrt{N}} \sum_{\ell \neq k} \frac{d\tilde{B}_{k\ell}}{\lambda_k - \lambda_\ell} u_\ell - \frac{1}{2N} \sum_{\ell \neq k} \frac{dt}{(\lambda_k - \lambda_\ell)^2} u_k,$$

where the  $\tilde{B}_{k\ell}$ 's are independent standard Brownian motions, and most importantly independent from the  $\tilde{B}_k$ 's from (2-11). This eigenvector flow was computed in the context of Brownian motion on ellipsoids [Norris, Rogers, and Williams \[1986\]](#), real Wishart processes [Bru \[1989\]](#), and for GOE/GUE in [G. W. Anderson, Guionnet, and Zeitouni \[2010\]](#).

Due to its complicated structure and high dimension, this eigenvector flow had not been analyzed. Surprisingly, these dynamics can be reduced to a multi-particle random walk in a dynamic random environment. More precisely, a configuration  $\eta$  consists in  $d$  points of  $\llbracket 1, N \rrbracket$ , with possible repetition. The number of particles at site  $x$  is  $\eta_x$ . A configuration obtained by moving a particle from  $i$  to  $j$  is denoted  $\eta^{ij}$ . The main observation from [Bourgade and Yau \[2017\]](#) is as follows. First denote  $z_k = \sqrt{N}\langle \mathbf{q}, u_k \rangle$ , which is random and time dependent. Then associate to a configuration  $\eta$  with  $j_k$  points at  $i_k$ , the renormalized moments observables (the  $\mathcal{N}_{i_k}$  are independent Gaussians) conditionally to the eigenvalues path,

$$(2-14) \quad f_{t,\lambda}(\eta) = \mathbb{E} \left( \prod_k z_{i_k}^{2j_k} \mid \lambda \right) / \mathbb{E} \left( \prod_k \mathcal{N}_{i_k}^{2j_k} \right).$$

Then  $f_{t,\lambda}$  satisfies the parabolic partial differential equation

$$(2-15) \quad \partial_t f_{t,\lambda}(\eta) = \mathcal{B}(t) f_{t,\lambda}(\eta)$$

where

$$\mathcal{B}(t) f(\eta) = \frac{1}{N} \sum_{i \neq j} 2\eta_i (1 + 2\eta_j) \frac{f(\eta^{ij}) - f(\eta)}{(\lambda_i(t) - \lambda_j(t))^2}.$$

As shown in the above drawing, the generator  $\mathcal{B}(t)$  corresponds to a random walk on the space of configurations  $\eta$ , with time-dependent rates given by the eigenvalues dynamics. This equation is parabolic and by the scale argument explained for (2-12),  $f_{t,\lambda}$  becomes locally constant (in fact, equal to 1 by normalization constraint) for  $t \geq N^{-1+\varepsilon}$ . This Hölder regularity is proved by a maximum principle.

**2.4 Other models.** The described dynamic approach applies beyond generalized Wigner matrices. We do not attempt to give a complete list of applications of this method. Below are a few results.

- (i) Wigner-type matrices designate variations of Wigner matrices with non centered  $H_{ii}$ 's [Lee, Schnell, Stetler, and Yau \[2015\]](#), or the normalization constraint in (2-8) not satisfied (the limiting spectral measure may differ from semicircular) [Ajanki, Erdős, and Kruger \[2017\]](#), or the  $H_{ij}$ 's non-centered and correlated [Ajanki, Erdős, and Kruger \[2018\]](#) and [Erdős, Kruger, and Schroder \[2017\]](#). In all cases, GOE bulk statistics is known.
- (ii) For small mean-field perturbations of diagonal matrices (the Rosenzweig-Porter model), GOE statistics [Landon, Sosoe, and Yau \[2016\]](#) occur with localization [Benigni \[2017\]](#) and [von Soosten and Warzel \[2017\]](#). We refer to [Facoetti, Vivo, and Biroli \[2016\]](#) for the physical meaning of this unusual regime.
- (iii) Random graphs also have bulk or edge GOE statistics when the connectivity grows fast enough with  $N$ , as proved for example for the Erdős–Renyi [Erdős, Knowles, Yau, and Yin \[2012\]](#), [Landon, Huang, and Yau \[2015\]](#), [Huang, Landon, and Yau \[2017\]](#), and [Lee and Schnell \[2015\]](#) and uniform  $d$ -regular models [Bauerschmidt, Huang, Knowles, and Yau \[2017\]](#). Eigenvectors statistics are also known to coincide with the GOE for such graphs [Bourgade, Huang, and Yau \[2017\]](#).
- (iv) The convolution model  $D_1 + U^* D_2 U$ , where  $D_1, D_2$  are diagonal and  $U$  is uniform on  $O(N)$ , appears in free probability theory. Its empirical spectral measure is understood up to the optimal scale [Bao, Erdős, and Schnell \[2017\]](#), and GOE bulk statistics were proved in [Che and Landon \[2017\]](#).
- (v) For  $\beta$ -ensembles, the external potential does not impact local statistics, a fact first shown when  $\beta = 1, 2, 4$  (the classical invariant ensembles) by asymptotics of orthogonal polynomials [Bleher and Its \[1999\]](#), [Deift \[1999\]](#), [Deift and Gioev \[2009\]](#), [Lubinsky \[2009\]](#), and [L. Pastur and M. Shcherbina \[1997\]](#). The dynamics approach extended this result to any  $\beta$  [Bourgade, Erdős, and Yau \[2014a,b\]](#). Other methods

based on sparse models [Krishnapur, Rider, and Virág \[2016\]](#) and transport maps [Bekerman, Figalli, and Guionnet \[2015\]](#) and [M. Shcherbina \[2014\]](#) were also applied to either  $\beta$ -ensembles or multimatrix models [Figalli and Guionnet \[2016\]](#).

- (vi) Extremal statistics. The smallest gaps in the spectrum of Gaussian ensembles and Wigner matrices have the same law [Bourgade \[2018\]](#), when the matrix entries are smooth. The relaxation step (2-12) was quantified with an optimal error so that the smallest spacing scale ( $N^{-4/3}$  in the GUE case [Arous and Bourgade \[2011\]](#)) can be perceived.

The above models are mean-field, a constraint inherent to the dynamic proof strategy. Indeed, the density step requires the matrix entries to fluctuate substantially: lemmas of type (2-13) need a constant variance of the entries along the dynamics (2-10).

### 3 Random band matrices and the Anderson transition

In the Wigner random matrix model, the entries, which represent the quantum transition rates between two quantum states, are all of comparable size. More realistic models involve geometric structure, as typical quantum transitions only occur between nearby states. In this section we briefly review key results for Anderson and band matrix models.

**3.1 Brief and partial history of the random Schrödinger operator.** Anderson's random Schrödinger operator [P. Anderson \[1958\]](#) on  $\mathbb{Z}^d$  describes a system with spatial structure. It is of type

$$(3-1) \quad H_{\text{RS}} = \Delta + \lambda V$$

where  $\Delta$  is the discrete Laplacian and the random variables  $V(x)$ ,  $x \in \mathbb{Z}^d$ , are i.i.d and centered with variance 1. The parameter  $\lambda > 0$  measures the strength of the disorder. The spectrum of  $H_{\text{RS}}$  is supported on  $[-2d, 2d] + \lambda \text{supp}(\mu)$  where  $\mu$  is the distribution of  $V(0)$

Amongst the many mathematical contributions to the analysis of this model, Anderson's initial motivation (localization, hence the suppression of electron transport due to disorder) was proved rigorously by [Fröhlich and Spencer \[1983\]](#) by a multiscale analysis: localization holds for strong disorder or at energies where the density of states  $\rho(E)$  is very small (localization for a related one-dimensional model was previously proved by [Gold-sheid, S. A. Molchanov, and L. Pastur \[1977\]](#)). An alternative derivation was given in [Aizenman and S. Molchanov \[1993\]](#), who introduced a fractional moment method. From the scaling theory of localization [Abraham, P. W. Anderson, Licciardello, and Ramakrishnan \[1979\]](#), extended states supposedly occur in dimensions  $d \geq 3$  for  $\lambda$  small enough, while eigenstates are only marginally localized for  $d = 2$ .

Unfortunately, there has been no progress in establishing the delocalized regime for the random Schrödinger operator on  $\mathbb{Z}^d$ . The existence of absolutely continuous spectrum (related to extended states) in the presence of substantial disorder is only known when  $\mathbb{Z}^d$  is replaced by homogeneous trees [Klein \[1994\]](#).

These results and conjecture were initially for the Anderson model in infinite volume. If we denote  $H_{\text{RS}}^N$  the operator (3-1) restricted to the box  $[-N/2, N/2]^d$  with periodic boundary conditions, its spectrum still lies on a compact set and one expects that the bulk eigenvalues in the microscopic scaling (i.e. multiplied by  $N^d$ ) converge to either Poisson or GOE statistics ( $H_{\text{RS}}^N$  corresponds to GOE rather than GUE because it is a real symmetric matrix). Minami proved Poisson spectral statistics from exponential decay of the resolvent [Minami \[1996\]](#), in cases where localization in infinite volume is known. For  $H_{\text{RS}}^N$ , not only is the existence of delocalized states in dimension three open, but also there is no clear understanding about how extended states imply GOE spectral statistics.

**3.2 Random band matrices: analogies, conjectures, heuristics.** The band matrix model we will consider was essentially already defined around (1-1). In addition, in the following we will assume subgaussian decay of the distribution of  $W^{\frac{d}{2}} H_{ij}$ , uniformly in  $i, j, N$ , for convenience (this could be replaced by a finite high moment assumption).

Although random band matrices and the random Schrödinger operator (3-1) are different, they are both local (their matrix elements  $H_{ij}$  vanish when  $|i - j|$  is large). The models are expected to have the same properties when

$$(3-2) \quad \lambda \approx \frac{1}{W}.$$

For example, eigenvectors for the Anderson model in one dimension are proved to decay exponentially fast with a localization length proportional to  $\lambda^2$ , in agreement with the analogy (3-2) and the [Equation \(1-2\)](#) when  $d = 1$ . For  $d = 2$ , it is conjectured that all states are localized with a localization length of order  $\exp(W^2)$  for band matrices,  $\exp(\lambda^{-2})$  for the Anderson model, again coherently with (3-2) and (1-2). For some mathematical justification of the analogy (3-2) from the point of view of perturbation theory, we refer to [Spencer \[2012, Appendix 4.11\]](#).

The origins of [Equation \(1-2\)](#) first lie on numerical evidence, at least for  $d = 1$ . In [Casati, Molinari, and Izrailev \[1990Apr\]](#) it was observed, based on computer simulations, that the bulk eigenvalue statistics and eigenvector localization length of  $1d$  random band matrices are essentially a function of  $W^2/N$ , with the sharp transition as described before (1-2). Fyodorov and Mirlin gave the first theoretical explanation for this transition [Fyodorov and Mirlin \[1991\]](#). They considered a slightly different ensemble with complex Gaussian entries decaying exponentially fast at distance greater than  $W$  from the diagonal.

Based on a non-rigorous supersymmetric approach [Efetov \[1997\]](#), they approximate relevant random matrix statistics with expectations for a  $\sigma$ -model approximation, from which a saddle point method gives the localization/delocalization transition for  $W \approx \sqrt{N}$ . Their work also gives an estimate on the localization length  $\ell$ , anywhere in the spectrum [Fyodorov and Mirlin \[1991\]](#), equation (19): it is expected that at energy level  $E$  (remember our normalization  $\sum_j \sigma_{ij}^2 = 1$  for  $H$  so that the equilibrium measure is  $\rho_{sc}$ ),

$$\ell \approx W^2(4 - E^2).$$

Finally, heuristics for localization/delocalization transition exponents follow from the conductance fluctuations theory developed by [Thouless \[1977\]](#), based on scaling arguments. For a discussion of mathematical aspects of the Thouless criterion, see [Spencer \[2012, 2010\]](#), and [Wang \[1992, Section III\]](#) for some rigorous scaling theory of localization. This criterion was introduced in the context of Anderson localisation, and was applied in [Sodin \[2010, 2014\]](#) to  $1d$  band matrices, including at the edge of the spectrum, in agreement with the prediction from [Fyodorov and Mirlin \[1991\]](#). A different heuristic argument for (1-2) is given in Section 3, for any dimension in the bulk of the spectrum.

**3.3 Results.** The density of states ( $\mathbb{E} (N^{-1} \sum_k \delta_{\lambda_k})$ ) of properly scaled random band matrices in dimension 1 converges to the semicircular distribution for any  $W \rightarrow \infty$ , as proved in [Bogachev, S. A. Molchanov, and L. A. Pastur \[1991\]](#). This convergence was then strengthened and fluctuations around the semicircular law were studied in [Guionnet \[2002\]](#), [G. W. Anderson and Zeitouni \[2006\]](#), [Jana, Saha, and Soshnikov \[2016\]](#), and [Li and Soshnikov \[2013\]](#) by the method of moments, at the macroscopic scale.

Interesting transitions extending the microscopic one (1-2) are supposed to occur at mesoscopic scales  $\eta$ , giving a full phase diagram in  $(\eta, W)$ . The work [Erdős and Knowles \[2015\]](#) rigorously analyzed parts of this diagram by studying linear statistics in some mesoscopic range and in any dimension, also by a moment-based approach.

The microscopic scale transitions (1-2) are harder to understand, but recent progress allowed to prove the existence of localization and delocalization for some polynomial scales in  $W$ . These results are essentially of four different types: (i) the localization side for general models, (ii) localization and delocalization for specific Gaussian models, (iii) delocalization for general models. Finally, (iv) the edge statistics are fully understood by the method of moments. Unless otherwise stated, all results below are restricted to  $d = 1$ .

(i) *Localization for general models.* A seminal result in the analysis of random band matrices is the following estimate on the localization scale. For simplicity one can assume that the entries of  $H$  are i.i.d. Gaussian, but the method from [Schenker \[2009\]](#) allows to treat more general distributions.



**Theorem 3.1** (The localization regime for band matrices [Schenker \[2009\]](#)). *Let  $\mu > 8$ . There exists  $\tau > 0$  such that for large enough  $N$ , for any  $\alpha, \beta \in \llbracket 1, N \rrbracket$  one has*

$$\mathbb{E} \left( \sup_{1 \leq k \leq N} |u_k(\alpha) u_k(\beta)| \right) \leq W^\tau e^{-\frac{|\alpha - \beta|}{W^\mu}}.$$

Localization therefore holds simultaneously for all eigenvectors when  $W \ll N^{1/8}$ , which was improved to  $W \ll N^{1/7}$  in [Peled, Schenker, Shamir, and Sodin \[2017\]](#) for some specific Gaussian model described below.

(ii) *Gaussian models with specific variance profile and supersymmetry.* For some Gaussian band matrices, the supersymmetry (SUSY) technique gives a purely analytic technique towards spectral properties. This approach has first been developed by physicists [Efetov \[1997\]](#). A rigorous supersymmetry method started with the expected density of states on arbitrarily short scales for a  $3d$  band matrix ensemble [Disertori, Pinson, and Spencer \[2002\]](#), extended to  $2d$  in [Disertori and Lager \[2017\]](#) (see [Spencer \[2012\]](#) for much more about the mathematical aspects of SUSY). More recently, the work [T. Shcherbina \[2014b\]](#) proved local GUE local statistics for  $W \geq cN$ , and delocalization was obtained in a strong sense for individual eigenvectors, when  $W \gg N^{6/7}$  and the first four moments of the matrix entries match the Gaussian ones [Bao and Erdős \[2017\]](#). These recent rigorous results assume complex entries and hold for  $|E| < \sqrt{2}$ , for a block-band structure of the matrix with a specific variance profile.

We briefly illustrate the SUSY method for moments of the characteristic polynomial: remarkably, this is currently the only observable for which the transition at  $W \approx \sqrt{N}$  was proved. Consider a matrix  $H$  whose entries are complex centered Gaussian variables such that

$$\mathbb{E}(H_{ij} H_{\ell k}) = \mathbb{1}_{i=k, j=\ell} J_{ij} \text{ where } J_{ij} = (-W^2 \Delta + 1)_{ij}^{-1},$$

and  $\Delta$  is the discrete Laplacian on  $\llbracket 1, N \rrbracket$  with periodic boundary condition. The variance  $J_{ij}$  is exponentially small for  $|i - j| > W^{1+\varepsilon}$ , so that  $H$  can be considered a random band matrix with band width  $W$ . Define

$$F_2(E_1, E_2) = \mathbb{E} (\det(E_1 - H) \det(E_2 - H)), \quad D_2 = F_2(E, E).$$

**Theorem 3.2** (Transition at the level of characteristic polynomials [T. Shcherbina \[2014a\]](#) and [M. Shcherbina and T. Shcherbina \[2017\]](#)). *For any  $E \in (-2, 2)$  and  $\varepsilon > 0$ , we have*

$$\lim_{N \rightarrow \infty} (D_2)^{-1} F_2 \left( E + \frac{x}{N \rho_{\text{sc}}(E)}, E - \frac{x}{N \rho_{\text{sc}}(E)} \right) = \begin{cases} \frac{\sin(2\pi x)}{2\pi x} & \text{if } N^\varepsilon < W < N^{\frac{1}{2}-\varepsilon} \\ 1 & \text{if } N^{\frac{1}{2}+\varepsilon} < W < N \end{cases}.$$

Unfortunately, currently the local eigenvalues statistics cannot be identified from moments of characteristic polynomials: they require ratios which are more difficult to analyze by the SUSY method.

We briefly mention the key steps of the proof of [Theorem 3.2](#). First, an integral representation for  $F_2$  is obtained by integration over Grassmann variables. These variables give convenient formulas for the product of characteristic polynomials: they allow to express the determinant as Gaussian-type integral. Integrate over the Grassmann variables then gives an integral representation (in complex variables) of the moments of interest. More precisely, the Gaussian representation for  $F_2 \left( E + \frac{x}{N\rho_{\text{sc}}(E)}, E - \frac{x}{N\rho_{\text{sc}}(E)} \right)$ , from [T. Shcherbina \[2014a\]](#), is

$$\frac{1}{(2\pi)^N} \frac{1}{\det J^2} \int e^{-\frac{W^2}{2} \sum_{j=-n+1}^n \text{Tr}(X_j - X_{j-1})^2 - \frac{1}{2} \sum_{j=-n}^n \text{Tr}(X_j + \frac{i\Delta_E}{2} + i \frac{\Delta_x}{N\rho_{\text{sc}}(E)})^2} \prod_{j=-n}^n \det(X_j - i\Delta_E/2) dX_j,$$

where  $N = 2n + 1$ ,  $\Delta_E = \text{diag}(E, E)$ ,  $\Delta_x = \text{diag}(x, -x)$ , and  $dX_j$  is the Lebesgue measure on  $2 \times 2$  Hermitian matrices. This form of the correlation of characteristic polynomial is then analyzed by steepest descent. Analogues of the above representation hold in any dimension, where the matrices  $X_j, X_k$ , are coupled in a quadratic way when  $k$  and  $j$  are neighbors in  $\mathbb{Z}^d$ , similarly to the Gaussian free field.

Finally, based on their integral representations, it is expected that random band matrices behave like  $\sigma$ -models, which are used by physicists to understand complicated statistical mechanics systems. We refer to the recent work [M. Shcherbina and T. Shcherbina \[2018\]](#) for rigorous results in this direction.

(iii) *Delocalization for general models.* Back to general models with no specific distribution of the entries (except sufficient decay of the distribution, for example subgaussian), the first delocalization results for random band matrices relied on a difficult analysis of their resolvent.

For example, the Green's function was controlled down to the scale  $W^{-1}$  in [Erdős, Yau, and Yin \[2012a\]](#), implying that the localization length of all eigenvectors is at least  $W$ . Analysis of the resolvent also gives full delocalization for most eigenvectors, for  $W$  large enough. In the theorem below, We say that an eigenvector  $u_k$  is subexponentially localized at scale  $\ell$  if there exists  $\varepsilon > 0$ ,  $I \subset \llbracket 1, N \rrbracket$ ,  $|I| \leq \ell$ , such that  $\sum_{\alpha \notin I} |u_k(\alpha)|^2 < e^{-N^\varepsilon}$ .

**Theorem 3.3** (Delocalized regime on average [Erdős, Knowles, Yau, and Yin \[2013\]](#)). *Assume  $W \gg N^{4/5}$  and  $\ell \ll N$ . Then the fraction of eigenvectors subexponentially localized on scale  $\ell$  vanishes as  $N \rightarrow \infty$ , with large probability.*

This result when  $W \geq N^{6/7}$  was previously obtained in [Erdős and Knowles \[2011\]](#), and similar statements were proved in higher dimension.

Delocalization was recently proved without averaging, together with eigenvalues statistics and flatness of individual eigenvectors. The main new ingredient is that quantum unique ergodicity is a convenient delocalization notion, proved by dynamics.

To simplify the statement below, assume that  $H$  is a Gaussian-divisible, in the sense that  $\sqrt{W}H_{ij}$  is the sum of two independent random variables,  $X + \mathcal{N}(0, c)$ , where  $c$  is an arbitrary small constant (the result holds for more general entries).

**Theorem 3.4** (Delocalized regime [Bourgade, Yau, and Yin \[2018\]](#)). *Assume  $W \gg N^{3/4+a}$  for some  $a > 0$ . Let  $\kappa > 0$  be fixed.*

- (a) *For any  $E \in (-2 + \kappa, 2 - \kappa)$  the eigenvalues statistics at energy level  $E$  converge to the GOE, as in (2-2).*
- (b) *The bulk eigenvectors are delocalized: for any (small)  $\varepsilon > 0$ , (large)  $D > 0$ , for  $N \geq N_0(\varepsilon, D, \kappa)$  and  $k \in [\kappa N, (1 - \kappa)N]$ , we have*

$$\mathbb{P} \left( \|u_k\|_\infty > N^{-\frac{1}{2}+\varepsilon} \right) < N^{-D}.$$

- (c) *The bulk eigenvectors are flat on any scale greater than  $W$ . More precisely, for any given (small)  $\varepsilon > 0$  and (large)  $D > 0$ , for  $N \geq N_0(\varepsilon, D)$ , for any deterministic  $k \in [\kappa N, (1 - \kappa)N]$  and interval  $I \subset [1, N]$ ,  $|I_N| > W$ , we have*

$$(3-3) \quad \mathbb{P} \left( \left| \sum_{\alpha \in I} (u_k(\alpha)^2 - \frac{1}{N}) \right| > N^{-a+\varepsilon} \frac{|I|}{N} \right) \leq N^{-D}.$$

A strong form of QUE similar to (c) holds for random  $d$ -regular graphs [Bauerschmidt, Huang, and Yau \[2016\]](#), the proof relying on exchangeability. For models with geometric constraints, other ideas are explained in the next section.

[Theorem 3.4](#) relies on a mean-field reduction strategy initiated in [Bourgade, Erdős, Yau, and Yin \[2017\]](#), and an extension of the dynamics (2-15) to observables much more general than (2-14), as explained in the next section. New ingredients compared to 3.4 are (a) quantum unique ergodicity for mean-field models after Gaussian perturbation, in a strong sense, (b) estimates on the resolvent of the band matrix at the (almost macroscopic) scale  $N^{-\varepsilon}$ .

The current main limitation of the method to approach the transition  $W_c = N^{1/2}$  comes from (b). These resolvent estimates are obtained by intricate diagrammatics developed in a series of previous works including [Erdős, Knowles, Yau, and Yin \[2013\]](#), and currently only proved for  $W \gg N^{3/4}$ .

(iv) *Edge statistics.* The transition in eigenvalues statistics is understood at the edge of the spectrum: the prediction from the Thouless criterion was made rigorous by a subtle method of moments. This was proved under the assumption that  $\sqrt{2W}(H_{ij})_{i \leq j}$  are  $\pm 1$  independent centered Bernoulli random variables, but the method applies to more general distributions. Remember that we order the eigenvalues  $\lambda_1 \leq \dots \leq \lambda_N$ .

**Theorem 3.5** (Transition at the edge of the spectrum [Sodin \[2010\]](#)). *Let  $\varepsilon > 0$ . If  $W \geq N^{\frac{5}{6} + \varepsilon}$ , then (2-3) holds. If  $W \leq N^{\frac{5}{6} - \varepsilon}$ , (2-3) does not hold.*

For eigenvectors (including at the edge of the spectrum), localization cannot hold on less than  $W/\log(N)$  entries as proved in [Benaych-Georges and P\'ech\'e \[2014\]](#), also by the method of moments.

## 4 Quantum unique ergodicity and universality

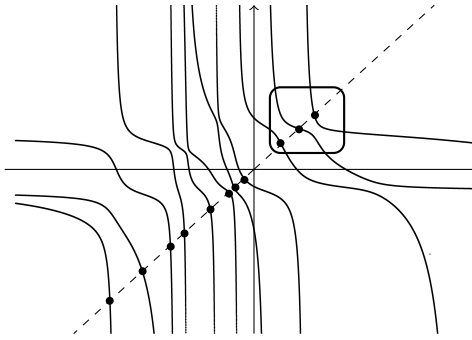
For non mean-field models, eigenvalues and eigenvectors interplay extensively, and their statistics should be understood jointly. Localization (decay of Green's function) is a useful a priori estimate in the proof of Poisson statistics for the Anderson model [Minami \[1996\]](#), and in a similar way we explain below why quantum unique ergodicity implies GOE statistics.

**4.1 Mean-field reduction.** The method introduced in [Bourgade, Erdős, Yau, and Yin \[2017\]](#) for GOE statistics of band matrices proceeds as follows. We decompose the  $1d$  band matrix from (1-1) and its eigenvectors as

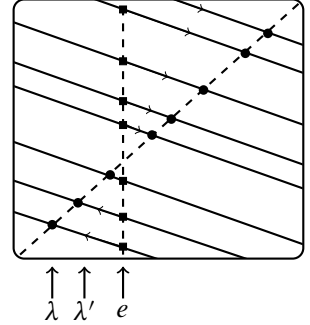
$$H = \begin{pmatrix} A & B^* \\ B & D \end{pmatrix}, \quad \mathbf{u}_j := \begin{pmatrix} \mathbf{w}_j \\ \mathbf{p}_j \end{pmatrix},$$

where  $A$  is a  $W \times W$  matrix. From the eigenvector equation  $H\mathbf{u}_j = \lambda_j\mathbf{u}_j$  we have  $(A - B^* \frac{1}{D - \lambda_j} B)\mathbf{w}_j = \lambda_j\mathbf{w}_j$ . The matrix elements of  $A$  do not vanish and thus the above eigenvalue problem features a mean-field random matrix (of smaller size). Hence one can consider the eigenvector equation  $Q_e \mathbf{w}_k(e) = \xi_k(e) \mathbf{w}_k(e)$  where

$$(4-1) \quad Q_e = A - B^*(D - e)^{-1}B,$$



(a) A simulation of eigenvalues of  $Q_e = A - B^*(D - e)^{-1}B$ , i.e. functions  $e \mapsto \xi_j(e)$ . Here  $N = 12$  and  $W = 3$ . The  $\lambda_i$ 's are the abscissa of the intersections with the diagonal.



(b) Zoom into the framed region of Figure (a), for large  $N, W$ : the curves  $\xi_j$  are almost parallel, with slope about  $1 - N/W$ . The eigenvalues of  $A - B^*(D - e)^{-1}B$  and those of  $H$  are related by a projection to the diagonal followed by a projection to the horizontal axis.

Figure 2: The idea of mean-field reduction: universality of gaps between eigenvalues for fixed  $e$  implies universality on the diagonal through parallel projection. For  $e$  fixed, we label the curves by  $\xi_k(e)$ .

and  $\lambda_k(e)$ ,  $\mathbf{w}_k(e)$  are eigenvalues and normalized eigenvectors. As illustrated below, the slopes of the functions  $e \mapsto \lambda_k(e)$  seem to be locally equal and concentrated:

$$\frac{d}{de} \lambda_k(e) \approx 1 - \frac{1}{\sum_{\alpha=1}^W w_k(\alpha)^2} (1 + o(1)) \approx 1 - \frac{N}{W},$$

which holds for  $e$  close to  $\lambda_k$ . The first equality is a simple perturbation formula<sup>2</sup>, and the second is true provided QUE for  $\mathbf{u}_k$  holds, in the sense of Equation (3-3) for example.

The GOE local spectral statistics holds for  $Q_e$  in the sense (2-2) (it is a mean-field matrix so results from Landon, Sosoe, and Yau [2016] apply), hence it also holds for  $H$  by parallel projection: GOE local spectral statistics follow from QUE.

This reduces the problem to QUE for band matrices, which is proved by the same mean-field reduction strategy: on the one hand, by choosing different overlapping blocks  $A$  along the diagonal, QUE for  $H$  follows from QUE for  $Q_e$  by a simple patching procedure (see section 3.3 for more details); on the other hand, QUE for mean-field models is

<sup>2</sup>The perturbation formula gives a slightly different equation, replacing  $\mathbf{w}_k$  by the eigenvector of a small perturbation of  $H$ , but we omit this technicality.

known thanks to a strengthening of the eigenvector moment flow method [Bourgade and Yau \[2017\]](#) and [Bourgade, Huang, and Yau \[2017\]](#), explained below.

**4.2 The eigenvector moment flow.** Obtaining quantum unique ergodicity from the regularity of [Equation \(2-15\)](#) (the eigenvector moment flow) is easy:  $\sqrt{N}\langle \mathbf{q}, u_k \rangle$  has limiting Gaussian moments for any  $\mathbf{q}$ , hence the entries of  $u_k$  are asymptotically independent Gaussian and the following variant of (3-3) holds for  $Q_e$  by Markov's inequality ( $w_k$  is rescaled to a unit vector): there exists  $\varepsilon > 0$  such that for any deterministic  $1 \leq k \leq W$  and  $I \subset \llbracket 1, W \rrbracket$ , for any  $\delta > 0$  we have

$$(4-2) \quad \mathbb{P} \left( \left| \sum_{i \in I} |w_k(\alpha)|^2 - \frac{|I|}{N} \right| \geq \delta \right) \leq N^{-\varepsilon} / \delta^2.$$

The main problem with this approach is that the obtained QUE is weak: one would like to replace the above  $\varepsilon$  with any large  $D > 0$ , as in (3-3). For this, it was shown in [Bourgade, Yau, and Yin \[2018\]](#) that much more general observables than (2-14) also satisfy the eigenvector moment flow parabolic [Equation \(2-15\)](#).

These new tractable observables are described as follows (we now switch to matrices of dimension  $N$  and spectrum  $\lambda_1 \leq \dots \leq \lambda_N$ , eigenvectors  $u_1, \dots, u_N$ , for comparison with (2-14)). Let  $I \subset \llbracket 1, N \rrbracket$  be given,  $(\mathbf{q}_\alpha)_{\alpha \in I}$  be any family of fixed vectors, and  $C_0 \in \mathbb{R}$ . Define

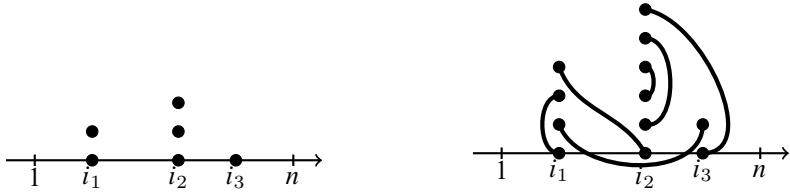
$$\begin{aligned} p_{ij} &= \sum_{\alpha \in I} \langle u_i, \mathbf{q}_\alpha \rangle \langle u_j, \mathbf{q}_\alpha \rangle \quad i \neq j \in \llbracket 1, n \rrbracket, \\ p_{ii} &= \sum_{\alpha \in I} \langle u_i, \mathbf{q}_\alpha \rangle^2 - C_0, \quad i \in \llbracket 1, n \rrbracket, \end{aligned}$$

When the  $\mathbf{q}_\alpha$ 's are elements of the canonical basis and  $C_0 = |I|/N$ , this reduces to

$$p_{ij} = \sum_{\alpha \in I} u_i(\alpha) u_j(\alpha), \quad (i \neq j) \quad p_{ii} = \sum_{\alpha \in I} u_i(\alpha)^2 - \frac{|I|}{N}, \quad i \in \llbracket 1, N \rrbracket,$$

and therefore the  $p_{ij}$ 's becomes natural partial overlaps measuring quantum unique ergodicity. For any given configuration  $\eta$  as given before (2-14), consider the set of vertices  $\mathcal{V}_\eta = \{(i, a) : 1 \leq i \leq n, 1 \leq a \leq 2\eta_i\}$ . Let  $\mathcal{G}_\eta$  be the set of perfect matchings of the complete graph on  $\mathcal{V}_\eta$ , i.e. this is the set of graphs  $G$  with vertices  $\mathcal{V}_\eta$  and edges  $\mathcal{E}(G) \subset \{\{v_1, v_2\} : v_1 \in \mathcal{V}_\eta, v_2 \in \mathcal{V}_\eta, v_1 \neq v_2\}$  being a partition of  $\mathcal{V}_\eta$ . For any given edge  $e = \{(i_1, a_1), (i_2, a_2)\}$ , we define  $p(e) = p_{i_1, i_2}$ ,  $P(G) = \prod_{e \in \mathcal{E}(G)} p(e)$  and

$$(4-3) \quad \widetilde{f}_{\lambda, t}(\eta) = \frac{1}{\mathfrak{M}(\eta)} \mathbb{E} \left( \sum_{G \in \mathcal{G}_\eta} P(G) \mid \lambda \right), \quad \mathfrak{M}(\eta) = \prod_{i=1}^n (2\eta_i)!!,$$



(a) A configuration  $\eta$  with  $\mathfrak{N}(\eta) = 6$ ,  
 $\eta_{i_1} = 2, \eta_{i_2} = 3, \eta_{i_3} = 1$ .

(b) A perfect matching  $G \in \mathfrak{G}_\eta$ . Here,  
 $P(G) = p_{i_1 i_1} p_{i_1 i_2} p_{i_2 i_2}^2 p_{i_2 i_3} p_{i_3 i_3} p_{i_3 i_1}$ .

where  $(2m)!! = \prod_{k \leq 2m, k \text{ odd}} k$ . The following lemma is a key combinatorial fact.

**Lemma 4.1.** *The above function  $\tilde{f}$  satisfies the eigenvector moment flow Equation (2-15).*

This new class of observables (4-3) widely generalizes (2-14) and directly encodes the  $L^2$  mass of eigenvectors, contrary to (2-14). Together with the above lemma, one can derive a new strong estimate, (2-14) with some small  $\varepsilon > 0$  replaced by any  $D > 0$ . The mean-field reduction strategy can now be applied in an efficient way: union bounds are costless thanks to the new small  $N^{-D}$  error term.

For  $d = 2, 3$ , the described mean-field reduction together with the strong version of the eigenvector moment flow should apply to give delocalization in some polynomial regime, such as  $W \gg N^{99/100}$ . However, this is far from the conjectures from (1-2). To approach these transitions, one needs to take more into account the geometry of  $\mathbb{Z}^d$ .

**4.3 Quantum unique ergodicity and the Gaussian free field.** At the heuristic level, the QUE method suggests the transition values  $W_c$  from (1-2). More precisely, consider a given eigenvector  $u = u_k$  associated to a bulk eigenvalue  $\lambda_k$ . For notational convenience, assume the model's band width is  $2W$  instead of  $W$ .

For  $\mathbb{1} = (1, \dots, 1) \in \mathbb{Z}^d$ , define  $\mathcal{W} = \llbracket 1, N \rrbracket^d \cap (2W\mathbb{Z}^d + W\mathbb{1})$ . For any  $w \in \mathcal{W}$ , let  $\mathcal{C}_w = \{\alpha \in \mathbb{Z}^d : \|w - \alpha\|_\infty \leq W\}$  be the cell of side length  $2W$  around  $w$ .

Let  $X_w = \sum_{\alpha \in \mathcal{C}_w} u(\alpha)^2$ . Consider a set  $I$ ,  $|I| = 2^d$ , such that the cells  $(\mathcal{C}_w)_{w \in I}$  form an hypercube  $\mathcal{H}$  of size  $(4W)^d$ . Assume one can apply the strong QUE statement (3-3) to a Schur complement  $Q_e$  of type (4-1) where  $A$  is now chosen to be the  $(4W)^d \times (4W)^d$  mean-field matrix indexed by the vertices from  $\mathcal{H}$ . We would obtain, for any two adjacent cells  $\mathcal{C}_w, \mathcal{C}_v$  with  $w, v \in I$ ,

$$(4-4) \quad \sum_{\alpha \in \mathcal{C}_w} u(\alpha)^2 = \sum_{\alpha \in \mathcal{C}_v} u(\alpha)^2 + O\left(N^\varepsilon \frac{W^{d/2}}{N^d}\right)$$

with overwhelming probability. By patching these estimates over successive adjacent cells, this gives

$$\sum_{\alpha \in \mathcal{C}_w} u(\alpha)^2 = \left(\frac{W}{N}\right)^d + O\left(N^\varepsilon \frac{W^{d/2}}{N^d}\right) \times \left(\frac{N}{W}\right)^d,$$

and therefore the leading order of  $\sum_{\alpha \in \mathcal{C}_w} u(\alpha)^2$  is identified (i.e. QUE holds) for  $W \gg N^{\frac{2}{3}}$ . This criterion, independent of the dimension  $d$ , is more restrictive than (1-2) and omits the important fact that the error term in (4-4) has a random sign.

One may assume that such error terms are asymptotically jointly Gaussian and independent for different pairs of adjacent cells (or at least for such pairs sufficiently far away). We consider the graph with vertices  $\mathcal{W}$  and edges the set of pairs  $(v, w)$  such that  $\mathcal{C}_v$  and  $\mathcal{C}_w$  are adjacent cells. A good model for  $(X_w)_{w \in \mathcal{W}}$  therefore is a Gaussian vector such that the increments  $X_v - X_w$  are independent, with distribution  $\mathcal{N}(0, W^d/N^{2d})$  when  $(v, w)$  is an edge, and conditioned to (1)  $\sum (X_{v_{i+1}} - X_{v_i}) = 0$  for any closed path  $v_1, v_2, \dots, v_j, v_1$  in the graph, (2)  $X_{v_0} = (W/N)^d$  to fix the ambiguity about definition of  $X$  modulo a constant. This model is simply the Gaussian free field, with density for  $(X_v)_v$  proportional to

$$e^{-\frac{N^{2d}}{2W^d} \sum_{v \sim w} (x_v - x_w)^2}.$$

As is well known, the Gaussian free field  $(Y_v)_v$  on  $\llbracket 1, n \rrbracket^d$  with density  $e^{-\frac{1}{2} \sum_{v \sim w} (y_v - y_w)^2}$  conditioned to  $Y_{v_0} = 0$  has the following typical fluctuation scale, for any deterministic  $v$  chosen at macroscopic distance from  $v_0$  (see e.g. [Biskup \[2017\]](#)):

$$(4-5) \quad \text{Var}(Y_v)^{1/2} \approx \begin{cases} n^{1/2} & \text{for } d = 1, \\ (\log n)^{1/2} & \text{for } d = 2, \\ O(1) & \text{for } d = 3. \end{cases}$$

We expect that quantum unique ergodicity (and GOE statistics by the mean-field reduction) holds when  $\text{Var}(X_v)^{1/2} \ll \mathbb{E}(X_v)$ . With  $n = N/W$ , this means

$$\frac{W^{d/2}}{N^d} \text{Var}(Y_v)^{1/2} \ll \frac{W^d}{N^d}$$

i.e.  $W \gg N^{1/2}$  for  $d = 1$ ,  $(\log N)^{1/2}$  for  $d = 2$ ,  $O(1)$  for  $d = 3$ .

**Acknowledgments.** The author's knowledge on this topic comes from collaborations with Laszlo Erdős, Horng-Tzer Yau, and Jun Yin. This note reports on joint progress with these authors.



## References

- E. Abraham, P. W. Anderson, D. C. Licciardello, and T. V. Ramakrishnan (1979). “Scaling theory of localization: absence of quantum diffusion in two dimensions”. *Phys. Rev. Lett.* 42, p. 673 (cit. on p. [2787](#)).
- M. Aizenman and S. Molchanov (1993). “Localization at large disorder and at extreme energies: an elementary derivation”. *Commun. Math. Phys.* 157, pp. 245–278 (cit. on p. [2787](#)).
- O. Ajanki, L. Erdős, and T. Kruger (2017). “Universality for general Wigner-type matrices”. *Probability Theory and Related Fields* 169 (3–4), pp. 667–727 (cit. on p. [2786](#)).
- (2018). “Stability of the matrix Dyson equation and random matrices with correlations”. *Probability Theory and Related Fields*, pp. 1–81 (cit. on p. [2786](#)).
- N. Anantharaman and E. Le Masson (2015). “Quantum ergodicity on large regular graphs”. *Duke Math. J.* 164.4, pp. 723–765 (cit. on p. [2781](#)).
- G. W. Anderson, A. Guionnet, and O. Zeitouni (2010). *An introduction to random matrices*. Vol. 118. Cambridge Studies in Advanced Mathematics. Cambridge University Press, pp. xiv+492 (cit. on p. [2785](#)).
- G. W. Anderson and O. Zeitouni (2006). “A CLT for a band matrix model”. *Probab. Theory Related Fields* 134.2, pp. 283–338 (cit. on p. [2789](#)).
- P. Anderson (1958). “Absences of diffusion in certain random lattices”. *Phys. Rev.* Pp. 1492–1505 (cit. on pp. [2778](#), [2787](#)).
- G. Ben Arous and P. Bourgade (2011). “Extreme gaps in the eigenvalues of random matrices”. *Annals of Probability* (cit. on p. [2787](#)).
- Z. Bao and L. Erdős (2017). “Delocalization for a class of random block band matrices”. *Probab. Theory Related Fields* 167.3-4, pp. 673–776 (cit. on p. [2790](#)).
- Z. Bao, L. Erdős, and K. Schnelli (2017). “Local law of addition of random matrices on optimal scale”. *Comm. Math. Phys.* 349.3, pp. 947–990 (cit. on p. [2786](#)).
- R. Bauerschmidt, J. Huang, A. Knowles, and H.-T. Yau (2017). “Bulk eigenvalue statistics for random regular graphs”. *Ann. Probab.* 45.6A, pp. 3626–3663 (cit. on p. [2786](#)).
- R. Bauerschmidt, J. Huang, and H.-T. Yau (2016). “Local Kesten-McKay law for random regular graphs”. *prepublication* (cit. on p. [2792](#)).
- F. Bekerman, A. Figalli, and A. Guionnet (2015). “Transport maps for  $\beta$ -matrix models and universality”. *Comm. Math. Phys.* 338.2, pp. 589–619 (cit. on p. [2787](#)).
- F. Benaych-Georges and S. Péché (2014). “Largest eigenvalues and eigenvectors of band or sparse random matrices”. *Electron. Commun. Probab.* 19, no. 4, 9 (cit. on p. [2793](#)).
- L. Benigni (2017). “Eigenvectors distribution and quantum unique ergodicity for deformed Wigner matrices”. *prepublication* (cit. on p. [2786](#)).
- M. Biskup (2017). “Extrema of the two-dimensional Discrete Gaussian Free Field” (cit. on p. [2797](#)).

- P. Bleher and A. Its (1999). “Semiclassical asymptotics of orthogonal polynomials, Riemann-Hilbert problem, and universality in the matrix model”. *Ann. of Math.* 150, pp. 185–266 (cit. on p. [2786](#)).
- A. Bloemendal, L. Erdős, A. Knowles, H.-T. Yau, and J. Yin (2014). “Isotropic local laws for sample covariance and generalized Wigner matrices”. *Elect. J. Probab.* 19 (cit. on p. [2784](#)).
- L. V. Bogachev, S. A. Molchanov, and L. A. Pastur (1991). “On the density of states of random band matrices”. Russian. *Mat. Zametki* 50.6, pp. 31–42, 157 (cit. on p. [2789](#)).
- P. Bourgade (2018). “Extreme gaps between eigenvalues of Wigner matrices”. *prepublication* (cit. on p. [2787](#)).
- P. Bourgade, L. Erdős, and H.-T. Yau (2014a). “Edge universality for beta ensembles”. *Communications in Mathematical Physics* 332.1, pp. 261–353 (cit. on p. [2786](#)).
- (2014b). “Universality of general  $\beta$ -ensembles”. *Duke Math. J.* 163.6, pp. 1127–1190 (cit. on p. [2786](#)).
- P. Bourgade, L. Erdős, H.-T. Yau, and J. Yin (2016). “Fixed energy universality for generalized Wigner matrices”. *Comm. Pure Appl. Math.* 69.10, pp. 1815–1881 (cit. on pp. [2782](#), [2783](#)).
- (2017). “Universality for a class of random band matrices”. *Advances in Theoretical and Mathematical Physics* 21 (3), pp. 739–800 (cit. on pp. [2792](#), [2793](#)).
- P. Bourgade, J. Huang, and H.-T. Yau (2017). “Eigenvector statistics of sparse random matrices”. *Electron. J. Probab.* 22, Paper No. 64, 38 (cit. on pp. [2786](#), [2795](#)).
- P. Bourgade and H.-T. Yau (2017). “The eigenvector moment flow and local quantum unique ergodicity”. *Comm. Math. Phys.* 350.1, pp. 231–278 (cit. on pp. [2783–2785](#), [2795](#)).
- P. Bourgade, H.-T. Yau, and J. Yin (2018). “Random band matrices in the delocalized phase I: Quantum unique ergodicity and universality”. *prepublication* (cit. on pp. [2792](#), [2795](#)).
- S. Brooks and E. Lindenstrauss (2013). “Non-localization of eigenfunctions on large regular graphs”. *Israel J. Math.* 193.1, pp. 1–14 (cit. on p. [2781](#)).
- M.-F. Bru (1989). “Diffusions of perturbed principal component analysis”. *J. Multivariate Anal.* 29.1, pp. 127–136 (cit. on p. [2785](#)).
- G. Casati, L. Molinari, and F. Izrailev (1990Apr). “Scaling properties of band random matrices”. *Phys. Rev. Lett.* 64, pp. 1851–1854 (cit. on p. [2788](#)).
- Z. Che and B. Landon (2017). “Local spectral statistics of the addition of random matrices”. *prepublication* (cit. on p. [2786](#)).
- Y. Colin de Verdière (1985). “Ergodicité et fonctions propres du laplacien”. French, with English summary. *Comm. Math. Phys.* 102.3, pp. 497–502 (cit. on p. [2781](#)).

- P. Deift (1999). *Orthogonal polynomials and random matrices: a Riemann–Hilbert approach*. Vol. 3. Courant Lecture Notes in Mathematics. New York University Courant Institute of Mathematical Sciences, pp. viii+273 (cit. on p. [2786](#)).
- (2017). “Some open problems in random matrix theory and the theory of integrable systems. II”. *SIGMA Symmetry Integrability Geom. Methods Appl.* 13, Paper No. 016, 23 (cit. on p. [2778](#)).
- P. Deift and D. Gioev (2009). *Random matrix theory: invariant ensembles and universality*. Vol. 18. Courant Lecture Notes in Mathematics. Courant Institute of Mathematical Sciences, New York; Amer. Math. Soc., Providence, RI, pp. x+217 (cit. on p. [2786](#)).
- M. Disertori and M. Lager (2017). “Density of states for random band matrices in two dimensions”. *Ann. Henri Poincaré* 18.7, pp. 2367–2413 (cit. on p. [2790](#)).
- M. Disertori, L. Pinson, and T. Spencer (2002). “Density of states for random band matrices”. *Commun. Math. Phys.* 232, pp. 83–124 (cit. on p. [2790](#)).
- F. Dyson (1962). “A Brownian-motion model for the eigenvalues of a random matrix”. *J. Math. Phys.* 3, pp. 1191–1198 (cit. on p. [2783](#)).
- K. Efetov (1997). “Supersymmetry in disorder and chaos”. *Cambridge University Press* (cit. on pp. [2789](#), [2790](#)).
- L. Erdős and A. Knowles (2011). “Quantum Diffusion and Delocalization for Band Matrices with General Distribution”. *Ann. Inst. H. Poincaré* 12.7, pp. 1227–1319 (cit. on p. [2792](#)).
- (2015). “The Altshuler-Shklovskii formulas for random band matrices I: the unimodular case”. *Comm. Math. Phys.* 333.3, pp. 1365–1416 (cit. on p. [2789](#)).
- L. Erdős, A. Knowles, H.-T. Yau, and J. Yin (2012). “Spectral statistics of Erdős–Rényi graphs II: eigenvalue spacing and the extreme eigenvalues”. *Comm. Math. Phys.* 314, pp. 587–640 (cit. on p. [2786](#)).
- (2013). “Delocalization and diffusion profile for random band matrices”. *Comm. Math. Phys.* 323.1, pp. 367–416 (cit. on pp. [2792](#), [2793](#)).
- L. Erdős, T. Kruger, and D. Schröder (2017). “Random Matrices with Slow Correlation Decay”. *prepublication* (cit. on p. [2786](#)).
- L. Erdős, S. Péché, J. A. Ramírez, B. Schlein, and H.-T. Yau (2010). “Bulk universality for Wigner matrices”. *Comm. Pure Appl. Math.* 63.7, pp. 895–925 (cit. on p. [2783](#)).
- L. Erdős, B. Schlein, and H.-T. Yau (2009). “Local semicircle law and complete delocalization for Wigner random matrices”. *Commun. Math. Phys.* 287, pp. 641–655 (cit. on p. [2784](#)).
- (2011). “Universality of random matrices and local relaxation flow”. *Inv. Math.* 187.1, pp. 75–119 (cit. on pp. [2782](#)–[2784](#)).
- L. Erdős and H.-T. Yau (2015). “Gap universality of generalized Wigner and beta ensembles”. *J. Eur. Math. Soc.* 17, pp. 1927–2036 (cit. on p. [2783](#)).

- L. Erdős, H.-T. Yau, and J. Yin (2012a). “Bulk universality for generalized Wigner matrices”. *Probab. Theory Related Fields* 154.1-2, pp. 341–407 (cit. on pp. 2783, 2791).
- (2012b). “Rigidity of eigenvalues of generalized Wigner matrices”. *Adv. Math.* 229.3, pp. 1435–1515 (cit. on pp. 2782, 2784).
- D. Facoetti, P. Vivo, and G. Biroli (2016). “From non-ergodic eigenvectors to local resolvent statistics and back: A random matrix perspective”. *EPL (Europhysics Letters)* 115.4 (cit. on p. 2786).
- A. Figalli and A. Guionnet (2016). “Universality in several-matrix models via approximate transportmaps”. *Acta Math.* 217.1, pp. 81–176 (cit. on p. 2787).
- J. Fröhlich and T. Spencer (1983). “Absence of diffusion in the Anderson tight binding model for large disorder or low energy”. *Commun. Math. Phys.* 88, pp. 151–184 (cit. on pp. 2778, 2787).
- Y.V. Fyodorov and A.D. Mirlin (1991). “Scaling properties of localization in random band-matrices: a sigma-model approach”. *Phys. Rev. Lett.* 67 (cit. on pp. 2788, 2789).
- I. Goldsheid, S. A. Molchanov, and L. Pastur (1977). “A random homogeneous Schrödinger operator has a pure point spectrum”. *Funkcional. Anal. i Prilozhen.* 11 (cit. on p. 2787).
- A. Guionnet (2002). “Large deviations upper bounds and central limit theorems for non-commutative functionals of Gaussian large random matrices”. *Annales de l’Institut Henri Poincaré (B)* 38 (3), pp. 341–384 (cit. on p. 2789).
- R. Holowinsky (2010). “Sieving for mass equidistribution”. *Ann. of Math. (2)* 172.2, pp. 1499–1516 (cit. on p. 2781).
- R. Holowinsky and K. Soundararajan (2010). “Mass equidistribution for Hecke eigenforms”. *Ann. of Math. (2)* 172.2, pp. 1517–1528 (cit. on p. 2781).
- J. Huang, B. Landon, and H.-T. Yau (2017). “Transition from Tracy-Widom to Gaussian fluctuations of extremal eigenvalues of sparse Erdős-Rényi graphs”. *prepublication* (cit. on p. 2786).
- I. Jana, K. Saha, and A. Soshnikov (2016). “Fluctuations of linear eigenvalue statistics of random bandmatrices”. *Theory Probab. Appl.* 60.3, pp. 407–443 (cit. on p. 2789).
- K. Johansson (2001). “Universality of the local spacing distribution in certain ensembles of Hermitian Wigner matrices”. *Comm. Math. Phys.* 215.3, pp. 683–705 (cit. on pp. 2782, 2783).
- A. Klein (1994). “Absolutely continuous spectrum in the Anderson model on the Bethe-lattice”. *Math. Res. Lett.* 1.4, pp. 399–407 (cit. on p. 2788).
- A. Knowles and J. Yin (2013a). “Eigenvector distribution of Wigner matrices”. *Probab. Theory Related Fields* 155.3-4, pp. 543–582 (cit. on p. 2784).
- (2013b). “The isotropic semicircle law and deformation of Wigner matrices”. *Comm. Pure Appl. Math.* 66, pp. 1663–1750 (cit. on p. 2784).

- M. Krishnapur, B. Rider, and B. Virág (2016). “Universality of the stochastic Airy operator”. *Comm. Pure Appl. Math.* 69.1, pp. 145–199 (cit. on p. [2787](#)).
- B. Landon, J. Huang, and H.-T. Yau (2015). “Bulk Universality of Sparse Random Matrices”. *J. Math. Phys.* 56.12 (cit. on p. [2786](#)).
- B. Landon, P. Sosoe, and H.-T. Yau (2016). “Fixed energy universality for Dyson Brownian motion”. *prepublication* (cit. on pp. [2783](#), [2786](#), [2794](#)).
- J.-O. Lee and K. Schnelli (2015). “Local law and Tracy-Widom limit for sparse random matrices”. *to appear in Probab. Theory Related Fields* (cit. on p. [2786](#)).
- J.-O. Lee, K. Schnelli, B. Stetler, and H.-T. Yau (2015). “Bulk universality for deformed Wigner matrices”. *to appear in Annals of Probability* (cit. on p. [2786](#)).
- L. Li and A. Soshnikov (2013). “Central limit theorem for linear statistics of eigenvalues of bandrandom matrices”. *Random Matrices Theory Appl.* 2.4, pp. 1350009, 50 (cit. on p. [2789](#)).
- E. Lindenstrauss (2006). “Invariant measures and arithmetic quantum unique ergodicity”. *Ann. of Math. (2)* 163.1, pp. 165–219 (cit. on p. [2781](#)).
- D. S. Lubinsky (2009). “A new approach to universality limits involving orthogonal polynomials”. *Ann. of Math. (2)* 170.2, pp. 915–939 (cit. on p. [2786](#)).
- M. L. Mehta and M. Gaudin (1960). “On the density of Eigenvalues of a random matrix”. *Nuclear Physics* 18, pp. 420–427 (cit. on p. [2780](#)).
- N. Minami (1996). “Local fluctuation of the spectrum of a multidimensional Anderson tightbinding model”. *Commun. Math. Phys.* 177, pp. 709–725 (cit. on pp. [2778](#), [2788](#), [2793](#)).
- J. R. Norris, L. C. G. Rogers, and David Williams (1986). “Brownian motions of ellipsoids”. *Trans. Amer. Math. Soc.* 294.2, pp. 757–765 (cit. on p. [2785](#)).
- L. Pastur and M. Shcherbina (1997). “Universality of the local eigenvalue statistics for a class of unitary invariant random matrix ensembles”. *J. Stat. Phys.* 86.1-2, pp. 109–147 (cit. on p. [2786](#)).
- R. Peled, J. Schenker, M. Shamir, and S. Sodin (2017). “On the Wegner orbital model”. *International Mathematical Research Notices* (cit. on p. [2790](#)).
- Z. Rudnick and P. Sarnak (1994). “The behaviour of eigenstates of arithmetic hyperbolic manifolds”. *Comm. Math. Phys.* 161.1, pp. 195–213 (cit. on p. [2781](#)).
- J. Schenker (2009). “Eigenvector localization for random band matrices with power law-band width”. *Comm. Math. Phys.* 290, pp. 1065–1097 (cit. on pp. [2789](#), [2790](#)).
- M. Shcherbina (2014). “Change of variables as a method to study general  $\beta$ -models: Bulk universality”. *J. Math. Phys.* 55 (cit. on p. [2787](#)).
- M. Shcherbina and T. Shcherbina (2017). “Characteristic polynomials for 1D random band matrices from the localization side”. *Communications in Mathematical Physics* 351 (cit. on p. [2790](#)).

- (2018). “Universality for 1d Random Band Matrices: Sigma-Model Approximation”. *Journal of Statistical Physics*, pp. 1–38 (cit. on p. [2791](#)).
- T. Shcherbina (2014a). “On the Second Mixed Moment of the Characteristic Polynomials of 1D Band Matrices”. *Communications in Mathematical Physics* 328, pp. 45–82 (cit. on pp. [2790](#), [2791](#)).
- (2014b). “Universality of the local regime for the block band matrices with a finite number of blocks”. *J. Stat. Phys.* 155.3, pp. 466–499 (cit. on p. [2790](#)).
- A. I. Shnirel’man (1974). *Uspekhi Mat. Nauk* 29.6, pp. 181–182 (cit. on p. [2781](#)).
- S. Sodin (2010). “The spectral edge of some random band matrices”. *Ann. of Math.* 173.3, pp. 2223–2251 (cit. on pp. [2789](#), [2793](#)).
- (2014). “Several applications of the moment method in random matrix theory”. *Proceedings of the International Congress of Mathematicians* (cit. on p. [2789](#)).
- P. von Soosten and S. Warzel (2017). “Non-Ergodic Delocalization in the Rosenzweig-Porter Model”. *prepublication* (cit. on p. [2786](#)).
- T. Spencer (2010). “Random banded and sparse matrices (Chapter 23) in “Oxford Handbook of Random Matrix Theory” edited by G. Akemann, J. Baik, and P. Di Francesco” (cit. on p. [2789](#)).
- (2012). “SUSY statistical mechanics and random band matrices” (cit. on pp. [2788](#)–[2790](#)).
- T. Tao and V. Vu (2011). “Random matrices: universality of local eigenvalue statistics”. *Acta Math.* 206.1, pp. 127–204 (cit. on pp. [2782](#)–[2784](#)).
- (2012). “Random matrices: universal properties of eigenvectors”. *Random Matrices Theory Appl.* 1.1 (cit. on p. [2784](#)).
- D. J. Thouless (1977). “Maximum metallic resistance in thin wires”. *Physical Review Letters* 39.18, pp. 1167–1169 (cit. on p. [2789](#)).
- C. Tracy and H. Widom (1994). “Level Spacing Distributions and the Airy Kernel”. *Communications in Mathematical Physics* 159, pp. 151–174 (cit. on p. [2780](#)).
- H. F. Trotter (1984). “Eigenvalue distributions of large Hermitian matrices; Wigner’s semi-circle law and a theorem of Kac, Murdock, and Szegő”. *Adv. in Math.* 54.1, pp. 67–82 (cit. on p. [2778](#)).
- W.-M. Wang (1992). *On localization and density of states for the random Schroedinger-operator*. Thesis (Ph.D.)—Princeton University. ProQuest LLC, Ann Arbor, MI, p. 92 (cit. on p. [2789](#)).
- E. Wigner (1957). “Distribution of neutron resonance level spacing”. In *International conference on the neutron interactions with the nucleus (Columbia University, New York, 1957)*, *Columbia Univ. Rept. CU-175 (TID-7547)*, pp. 49–50 (cit. on p. [2778](#)).
- S. Zelditch (1987). “Uniform distribution of eigenfunctions on compact hyperbolic surfaces”. *Duke Math. J.* 55.4, pp. 919–941 (cit. on p. [2781](#)).

Received 2018-03-05.

PAUL BOURGADE  
COURANT INSTITUTE  
NEW YORK UNIVERSITY  
[bourgade@cims.nyu.edu](mailto:bourgade@cims.nyu.edu)

# INVARIANCE IN HETEROGENEOUS, LARGE-SCALE AND HIGH-DIMENSIONAL DATA

PETER BÜHLMANN

## Abstract

Statistical inference from large-scale data can benefit from sources of heterogeneity. We discuss recent progress of the mathematical formalization and theory for exploiting heterogeneity towards predictive stability and causal inference in high-dimensional models. The topic is directly motivated by a broad range of applications and we will show an illustration from molecular biology with gene knock out experiments.

## 1 Introduction

In the advent of large data acquisition we expect that heterogeneity occurs within datasets. The data are typically *not* realizations from independent and identically distributed random variables, nor from a stationary process. We either know that the data come from large-scale experimental perturbations, for example in many bio-molecular applications, or one can empirically detect heterogeneity in terms of shifts, non-stationarities or cluster-membership, for example in macroeconomics.

Rather than considering heterogeneity as a nuisance, one can exploit and use it to obtain more insights and better predictions – in folklore: “Make heterogeneity your friend rather than your enemy”. This line of thinking in the context of large-scale data seems “new” [Peters, Bühlmann, and Meinshausen \[2016\]](#): the foundations though are much older and go back to Trygve Haavelmo in 1943 [Haavelmo \[1943\]](#), who received the Nobel Prize in economics in 1989 “for his clarification of the probability theory foundations of econometrics and his analyses of simultaneous economic structures”. Haavelmo has advocated an invariance property across changing (heterogeneous) structures, technically in terms of structural equation models, and this is nowadays adopted in the framework of causality [Pearl \[2000, cf.\]](#).

---

*MSC2010:* primary 62J07; secondary 62F99, 68T99.

*Keywords:* Causality, Hidden variables, High-dimensional models, Structural equation models, Predictive stability.



We discuss here the “reverse relation” where invariance can be extracted from data, enabling predictive stability and towards inference of causal parameters. The corresponding mathematical formulation and theory involve the combination of identifiability from causal inference and techniques from high-dimensional statistical inference (the latter is due to the dimensions of many modern datasets). The methodology and mathematical problems are in close vicinity of applications: we will illustrate an ambitious example of predicting gene knock out perturbations, a fundamental task in molecular-biology for gaining insights into causal gene interactions and for prioritizing future biological experiments.

## 2 The setting

We consider regression or classification problems with  $d$ -dimensional covariates (features)  $X$  and a one-dimensional response variable of interest  $Y$ .

As a starting point, consider a linear model for  $n$  data points, being realizations from

$$Y_i = X_i^T \beta^0 + \varepsilon_i \quad (i = 1, \dots, n),$$

where the covariates  $X_i$ , the response  $Y_i$  and the errors  $\varepsilon_i$  are random with  $\mathbb{E}[\varepsilon_i | X_i] = 0$  and the  $d \times 1$  vector  $\beta^0$  denotes the unknown true regression parameter of interest. (Because this true underlying parameter is of special interest, we denote it with an additional superscript “0”). The most often used assumption is that the random variables  $X_i$  and  $\varepsilon_i$  are independent from each other and both of them independent and identically distributed (i.i.d.) across  $i = 1, \dots, n$ , and hence  $(Y_i, X_i)$ ,  $i = 1, \dots, n$ , are i.i.d. as well. We often use the short-hand notation for a linear model

$$(1) \quad \mathbf{Y} = \mathbf{X}\beta^0 + \boldsymbol{\varepsilon},$$

where  $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ ,  $\boldsymbol{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)^T$  are  $n \times 1$  vectors and  $\mathbf{X} = (X_1, \dots, X_n)^T$  is the  $n \times d$  design matrix.

Over the last 15 years, a huge amount of literature has been devoted to the problem of estimating the unknown parameter vector  $\beta^0$  in the high-dimensional sparse case where  $d \gg n$ : some of the earlier references include [Donoho \[1993\]](#), [Donoho and Johnstone \[1994\]](#), [Tibshirani \[1996\]](#), [Chen, Donoho, and Saunders \[2001\]](#), [Greenshtein and Ritov \[2004\]](#), [Bühlmann \[2006\]](#), [Meinshausen and Bühlmann \[2006\]](#), [Bunea, Tsybakov, and Wegkamp \[2007\]](#), [Zou \[2006\]](#), [Zhao and Yu \[2006\]](#), [Candès and Tao \[2007\]](#), [Bickel, Ritov, and Tsybakov \[2009\]](#), and [Koltchinskii \[2009a,b\]](#), and see also the monographs [Bühlmann and van de Geer \[2011\]](#), [Giraud \[2014\]](#), and [Hastie, Tibshirani, and Wainwright \[2015\]](#). Furthermore, estimation of the parameter  $\beta^0$  in the noiseless case is the same as compressed sensing [Donoho and Huo \[2001\]](#), [Donoho \[2006\]](#), [Candès, Romberg, and Tao \[2006\]](#), and [Candès and Tao \[2006, cf.\]](#), and this itself is a huge field by now.

We will focus on the case where an i.i.d. assumption as above for the random variables  $(Y_i, X_i)$  ( $i = 1, \dots, n$ ) does not hold. This seems particularly relevant for large-scale “big” data. In the advent of large data collection, it is often reasonable to assume that the data exhibits “heterogeneity”. Loosely speaking, we mean by this that the data come from e.g.: (i) different regimes, for example across time in applications such as economics, finance or neuroscience; (ii) from different perturbations, for example in molecular biology; (iii) from different sub-populations, for example in online advertisement or auction pricing. In an abstract sense, we generalize the linear model from (1) to

$$(2) \quad (\mathbf{Y}^e, \mathbf{X}^e) \sim F^e, \quad e \in \mathcal{E},$$

where  $\mathbf{Y}^e$  is an  $n^e \times 1$  vector,  $\mathbf{X}^e$  an  $n^e \times d$  design matrix,  $e$  denotes an environment or a sub-population from a space of environments  $\mathcal{E}$ , and  $F^e$  is the distribution depending on environment  $e$ . Typically, we assume that the environments  $e \in \mathcal{E}$  are known (observed), but see below for an example where they are unknown.

*Example: Gene knock out perturbations in yeast* [Meinshausen, Hauser, Mooij, Peters, Versteeg, and Bühlmann \[2016\]](#).

Among the approximately 6’000 genes in yeast, 1’479 have been knocked out and a phenotypic response is measured. The space  $\mathcal{E}$  corresponds to the different gene knock-out perturbations. In the extreme case, every gene knock out corresponds to a single  $e$  and the space  $\mathcal{E} = \{1, 2, \dots, 1479\}$ . One can also think to pool (some of) the different perturbations and the space  $\mathcal{E}$  is then smaller. This example is discussed further in Section 5.4.

*Example: Monetary policy in macro economics* [Pfister, Bühlmann, and Peters \[2017\]](#). The data are monthly observations of the Euro – Swiss Franc exchange rate over 18 years and nine macro economic variables such as GDP or inflation rate. The space  $\mathcal{E}$  corresponds to unknown time-dependent regimes. Although the environments  $e$  (the regimes) in  $\mathcal{E}$  are unknown, they are assumed to be present in the observed data.

Consider a space of (mostly unobserved) environments  $\mathcal{F}$ , typically  $\mathcal{F} \supset \mathcal{E}$  being much larger than the observed environments in  $\mathcal{E}$ . In the examples above,  $\mathcal{F}$  would be  $\mathcal{E}$  but in addition also including the gene perturbations or the future time-dependent regimes which are not observed in the data. We are interested in the following problems.

1. Prediction for new scenarios or environments in  $\mathcal{F}$ . When adopting a linear model, consider the following objective:

$$\beta^* = \operatorname{argmin}_{\beta} \max_{e \in \mathcal{F}} \mathbb{E}[|Y^e - (X^e)^T \beta|^2],$$

and how to infer  $\beta^*$  from data as in (2) from much fewer observed environments in  $\mathcal{E}$ . That is, we want to infer the parameter  $\beta^*$  which optimizes the worst case loss

within a class of new unobserved scenarios in  $\mathcal{F}$ . We will discuss in Section 5 (e.g. Theorem 4) some cases for  $\mathcal{F}$ .

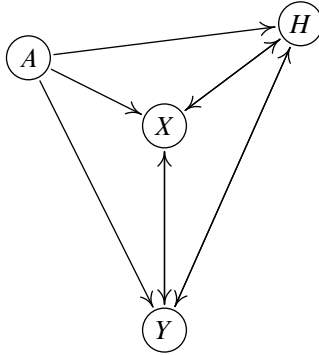
2. Predicting unseen interventions or perturbations. We want to predict  $Y^e$  when doing a perturbation on some of the covariates  $X^e$ , for  $e \in \mathcal{F}$  corresponding to a new perturbation which was not observed in the data. That is, in popular terms, a “what if I do (perturb)” question: the answer to such a question is at the fundamental basis of causal inference Pearl [2000] and Peters, Janzing, and Schölkopf [2017, cf.].
3. Finding stable structures. We also aim to find subsets of covariates  $S \subseteq \{1, \dots, d\}$  leading to (near) invariance in terms of the conditional distribution  $\mathcal{L}(Y^e | X_S^e)$  being (nearly) constant across environments  $e \in \mathcal{F}$ .

All these three points above are closely related. When it comes to inference from finite samples, the setting is usually high-dimensional since in each observed environment  $e$ , the sample size  $n^e$  is often not so large implying that the covariate dimension  $d \gg n^e$ . Therefore, the underlying mathematical developments involve high-dimensional statistical theory together with perturbation analysis in structural equation models (see (4)) for analyzing heterogeneity.

### 3 Modeling heterogeneity and perturbations

We discuss here a general model for heterogeneous data as in (2). There is a  $d$ -dimensional covariate  $X$ , a  $q$ -dimensional hidden (latent) variable  $H$ , an  $r$ -dimensional “anchor” variable  $A$  and a one-dimensional response  $Y$ . A discrete space of environments  $\mathcal{E} = \{1, \dots, m\}$  mentioned before can be described with  $r = m$  anchor variables, each of them being binary where  $A_k = 1$  means that the corresponding environment is  $e = k$  ( $k = 1, \dots, m$ ).

The involved variables  $(Y, X, H, A)^T$  is a  $(1 + d + q + r)$ -dimensional random vector and each component is corresponding to a node in a directed graph describing the “structure” among the variables. The directed graph is qualitatively as follows:



The (bi-)directed arrows correspond to the structure of a structural equation model (see below). There are structural relations among the components of  $X$ ,  $H$  and  $A$  as well but this is not visible in the displayed graph. Bi-directed arrow allow to exhibit directed cycles, i.e., feedback loops. The variables in  $H$  are hidden confounders between  $X$  and  $Y$  which makes it hard to identify effects from components  $X_j$  to  $Y$  which are not due to the hidden confounding.

A quantitative model on the graph is given by a structural equation model: a linear structural equation model is given below in (4) in its abstract form. To exemplify: the equation for the response, which is of major interest since  $Y$  is the target one wants to understand, reads

$$(3) \quad Y \leftarrow \sum_{k \in \text{pa}(Y) \cap \{X\}} \beta_k^0 X_k + \sum_{k \in \text{pa}(Y) \cap \{H\}} \delta_k H_k + \sum_{k \in \text{pa}(Y) \cap \{A\}} \alpha_k A_k + \varepsilon_Y,$$

where  $\text{pa}(Y)$  denotes the parental set of a variable  $Y$  in the directed graph,  $\{X\}$  denotes the nodes corresponding to the random variables from the components of  $X$  (and analogously for  $H$ ,  $A$ ), and  $\varepsilon_Y$  is an exogenous stochastic noise term. The directed arrow “ $\leftarrow$ ” means that the variable on the left hand side is a “direct function” or “caused” by the variables on the right hand side: it can be replaced by the expression of “equality in distribution”. The parameter  $\beta^0$  is of special interest: in the literature it is called the direct causal effect [Pearl \[2000, cf.\]](#), describing the direct effect from  $X$  to  $Y$  (see below). In fact, the causal parameter describes what happens when doing a perturbation/intervention on the  $X$ -variables, see goal 2. in Section 2, and it is intrinsically related to invariance properties with respect to perturbations [Haavelmo \[1943\]](#) and [Peters, Bühlmann, and Meinshausen \[2016\]](#): we will take up the latter point in Section 5. We note that an  $L_2$ -projection does not lead to  $\beta^0$ :  $\arg\min_{\beta} \mathbb{E}[|Y - X^T \beta|^2] \neq \beta^0$ ; thus, inferring  $\beta^0$  from data is a more complicated task than using standard regression methodology.

Having displayed above the structural equation of the response  $Y$ , all other variables have such a structural representation as well, for example

$$\begin{aligned} X_j \leftarrow & \sum_{k \in \text{pa}(X_j) \cap \{X\}} \kappa_{j,k} X_k + \sum_{k \in \text{pa}(X_j) \cap \{H\}} \gamma_{j,k} H_k \\ & + \sum_{k \in \text{pa}(X_j) \cap \{A\}} \alpha_{j,k} A_k + \xi_j Y I(Y \in \text{pa}(X_j)) + \varepsilon_j \end{aligned}$$

We can represent the model in algebraic form as

$$(4) \quad \begin{pmatrix} Y \\ X \\ H \end{pmatrix} = B \begin{pmatrix} Y \\ X \\ H \end{pmatrix} + MA + \varepsilon,$$

where  $B$  is a  $(1 + d + r) \times (1 + d + r)$  matrix,  $M$  a  $(1 + d + r) \times r$  matrix and  $\varepsilon$  a stochastic noise vector of dimension  $(1 + d + r) \times 1$ . We note that since  $A$  is an “anchor” or a source node in the graph, it is exogenous and hence it appears on the right hand side of (4) only.

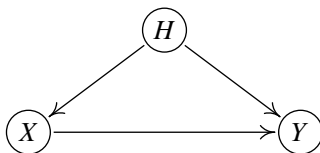
Of main interest is the equation and dynamics for the response  $Y$ : if  $(I - B)$  is invertible, see below, we can express  $Y$  and all other  $X, H$  as a function of  $A$  and  $\varepsilon$ ,

$$(5) \quad \begin{pmatrix} X \\ Y \\ H \end{pmatrix} = (I - B)^{-1}(\varepsilon + MA).$$

As mentioned above, the variable  $A$  can describe heterogeneity, and the formulation in (5) is a useful representation for the perturbation effect of shift interventions, see Section 5.2. The matrix  $(I - B)$  is invertible if the underlying structure (encoding zeroes in  $B$ ) is a directed acyclic graph; for cyclic graphs, one typically assumes an equilibrium solution of the dynamical system when conditioning on  $\varepsilon$  and  $A$ , for example requiring that the cycle-product is strictly less than one (but we do not need such an equilibrium assumption).

**3.1 Some special cases.** Some special cases highlight the generality of the model in (4).

**Hidden confounding.** This model has no “anchor” variable  $A$  but some hidden (unobserved) confounders  $H$ . The directed graph looks as follows.



The structural equation model is:

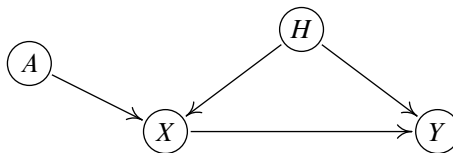
$$\begin{aligned}
 H &= (H_1, \dots, H_q)^T \text{ uncorrelated with covariance matrix } I_q \\
 X &\leftarrow \Gamma H + \varepsilon_X, \\
 (6) \quad Y &\leftarrow X^T \beta^0 + H^T \delta + \varepsilon_Y
 \end{aligned}$$

where all the components of  $\varepsilon_X, \varepsilon_Y, H$  are jointly independent,  $\Gamma$  is a  $d \times q$  matrix and  $\beta^0$  the  $d \times 1$  regression vector. There is an implicit directionality assumption saying that there are no structural directions from  $Y$  to some of the components of  $X$  (i.e.,  $Y$  is “childless”).

We will argue that despite the hidden confounders  $H$ , one can estimate the causal coefficient vector  $\beta^0$  when the setting is high-dimensional (among other conditions).

A prominent example for such a model are genome-wide association studies (GWAS). The response variable  $Y$  is often a medical or disease status, the covariates  $X$  are genetic biomarkers in terms of single nucleotide polymorphisms (SNPs) and the hidden confounders can come from various sources such as environment or unmeasured genetic profiles. The dimensionality of the SNPs is in the order of  $d = O(10^6)$  and a typical sample size is in the range of  $O(10^3)$ . It is a very high-dimensional setting with the interesting additional information about direction: if there is an association between  $X$  and  $Y$ , it must point from  $X$  to  $Y$ ; this, because the SNPs are genetic information and the medical status cannot influence the genetics (although there are some exceptions with retroviruses like HIV).

**Instrumental variables regression.** This is a very popular and well-studied model in economics. The variables in  $A$  are called instruments and the directed graph looks as follows.



In contrast to the general situation considered above, there are no directed arrows from either  $X, Y$  or  $H$  to  $A$ , and there are no bi-directed arrows.

The corresponding structural equation model is as in (4) with the constraint from the directed graph above:

$$\begin{aligned}
 H &= (H_1, \dots, H_q)^T \text{ from a distribution } F_H, \\
 A &= (A_1, \dots, A_r)^T \text{ from a distribution } F_A, \\
 X &\leftarrow \Gamma H + MA + \varepsilon_X, \\
 Y &\leftarrow X^T \beta^0 + H^T \delta + \varepsilon_Y.
 \end{aligned}
 \tag{7}$$

A necessary condition for identifiability of  $\beta^0$  is to have at least as many instruments as covariates, i.e.,  $r \geq d$ . More precisely, the condition  $\text{rank}(\mathbb{E}[(AA^T)]M^T) \geq d$ , is necessary and sufficient; and this involves also that the coefficient matrix  $M$  is not “too degenerate”.

## 4 Hidden confounding in high-dimensional settings

Consider here the model in (6) with hidden confounder variables. Such hidden confounders can also be thought as generating different regimes or environments in the data. In high-dimensional settings, we assume that the regression parameter  $\beta^0$  (with a causal interpretation) is sparse.

When using the population least squares principle, we obtain

$$\begin{aligned}
 \beta_{\text{LS}} &= \Sigma_X^{-1} \text{Cov}(Y, X) = \beta^0 + b, \\
 b &= \Sigma_X^{-1} \Gamma \delta,
 \end{aligned}$$

where  $\Sigma_X = \text{Cov}(X) = \Gamma \Gamma^T + \text{Cov}(\varepsilon_X)$ .

*Example: one hidden confounder ( $r = 1$ ) and same noise terms for  $X$ :  $\text{Cov}(\varepsilon_X) = \sigma_\varepsilon^2 I$ .*

We obtain that the bias equals

$$b = \Gamma \frac{\delta}{\Lambda_{\max}^2(\Gamma^T \Gamma) + \sigma_\varepsilon^2} = \Gamma \frac{\delta}{\|\Gamma\|_2^2 + \sigma_\varepsilon^2},$$

where  $\Lambda_{\max}^2(\Gamma^T \Gamma)$  denotes the maximal eigenvalue of  $\Gamma \Gamma^T$  (the only non-zero eigenvalue). Suppose that the number of non-zero entries in  $\Gamma$  is  $m$  and that the non-zero entries in  $\Gamma$  are not too small, i.e.,  $\|\Gamma\|_2^2 \asymp m \rightarrow \infty$  as  $d \rightarrow \infty$ . Then,

$$\|b\|_2^2 = O(m^{-1}) \text{ as } d \rightarrow \infty.$$

Thus, if the hidden variables have an effect which is sufficiently spread out, i.e.  $m$  being large, the bias of population least squares (which ignores the hidden confounders) will disappear in high dimensions.

**4.1 Sparse plus dense signals.** The analysis above can be used as follows. We can write the model by  $L_2$ -projection as

$$Y = X^T \beta_{\text{LS}} + \eta, \quad \text{Cov}(\eta, X) = 0,$$

and thus we can also write

$$(9) \quad Y = X^T (\beta^0 + b) + \eta,$$

with the bias term as above which is typically dense (under fairly natural conditions, see above).

We assume to have observed data  $(\mathbf{Y}, \mathbf{X})$  being i.i.d. realizations of  $(Y_i, X_i)$  ( $i = 1, \dots, n$ ) from (6), also involving the unobserved  $\mathbf{H}$  from i.i.d. realizations  $H_i$  ( $i = 1, \dots, n$ ). The sample version of the model in (9) is as follows:

$$\mathbf{Y} = \mathbf{X}(\beta^0 + b) + \eta, \quad \eta \text{ uncorrelated with } \mathbf{X}.$$

In the high-dimensional regime with  $d \gg n$ , we have a linear model with a signal being a composition of a sparse  $\beta^0$  and a dense  $b$ . Estimation of the sparse vector  $\beta^0$  can be done with a combination of  $\ell_1$ - and  $\ell_2$  regularization, the Lava estimator [Chernozhukov, Hansen, Liao, et al. \[2017\]](#):

$$\text{argmin}_{\beta, b} (\|\mathbf{Y} - \mathbf{X}(\beta + b)\|_2^2/n + \lambda_1 \|\beta\|_1 + \lambda_2 \|b\|_2^2).$$

Note that for  $\lambda_2 = \infty$  we obtain the  $\ell_1$ -regularized Lasso estimator, and analogously  $\lambda_1 = \infty$  corresponds to the  $\ell_2$ -regularized Ridge procedure (Tikhonov regularization). The solution of this convex optimization problem is given by:

$$(10) \quad \begin{aligned} \hat{\beta} &= \text{argmin}_{\beta} \{ \|\tilde{\mathbf{Y}} - \tilde{\mathbf{X}}\beta\|_2^2/n + \lambda_1 \|\beta\|_1 \}, \\ \hat{b} &= (\mathbf{X}^T \mathbf{X} + n\lambda_2 I)^{-1} \mathbf{X}^T (\mathbf{Y} - \mathbf{X}\hat{\beta}). \end{aligned}$$

Here,  $\tilde{\mathbf{X}}$  and  $\tilde{\mathbf{Y}}$  are given by  $K_{\lambda_2}^{1/2} \mathbf{X}$  and  $K_{\lambda_2}^{1/2} \mathbf{Y}$  respectively, where

$$K_{\lambda_2} = I - \mathbf{X}(\mathbf{X}^T \mathbf{X} + n\lambda_2 I)^{-1} \mathbf{X}^T.$$

The representation in (10) leads to some consequences and insights. First, the computation can simply be done by an  $\ell_1$ -norm regularization to a transformed problem with response  $\tilde{\mathbf{Y}}$  and covariates  $\tilde{\mathbf{X}}$ . Second, the mathematical properties of  $\hat{\beta}$  can be studied from the view point of the theory for  $\ell_1$ -norm regularization (and compressed sensing): there are two obstacles though, namely: (i) the underlying coefficient vector is sparse plus dense, and a bias term will be inherent in sparse approximation; (ii) the transformed design



matrix  $\tilde{\mathbf{X}}$  has other restricted eigenvalues [Bickel, Ritov, and Tsybakov \[2009\]](#) and [van de Geer and Bühlmann \[2009\]](#) than the original  $\mathbf{X}$  and the transformed error  $\tilde{\varepsilon} = K_{\lambda_2}^{1/2}$  is correlated and possibly inflates.

The following result holds.

**Theorem 1.** [Ćevđid, Bühlmann, and Meinshausen \[2018\]](#) *Assume Gaussian variables  $H, \varepsilon_X, \varepsilon_Y$  in (6) with mean zero. In addition, assume condition (A) described below. Denote by  $s_0 = |\text{supp}(\beta^0)|$  and assume that  $s_0 \sqrt{\log(d)/n} = o(1)$  ( $d \gg n \rightarrow \infty$ ). Then, for the Lava estimator in (10), there exist suitable values  $\lambda_1$  and  $\lambda_2$  such that with probability tending to one as  $d \gg n \rightarrow \infty$ :*

$$(11) \quad \|\hat{\beta} - \beta^0\|_1 \leq C \frac{\sigma s_0}{\Lambda_{\min}^2(\Sigma)} \sqrt{\log(d)/n},$$

where  $0 < C < \infty$  is a constant,  $\sigma^2 = \mathbb{E}|\eta|^2$  in (9) and  $\Lambda_{\min}^2(\Sigma_X)$  denotes the minimal eigenvalue of  $\Sigma_X = \text{Cov}(X) = \Gamma\Gamma^T + \text{Cov}(\varepsilon_X)$ .

The assumption (A) below ensures that the bias term is asymptotically negligible. There is a broader range of scenarios implying negligible bias, and a simple example is as follows.

**(A)** In model (6), the entries of the  $d \times q$  matrix  $\Gamma$  are i.i.d. from an absolutely continuous distribution w.r.t. Lebesgue measure,  $q < \infty$  is a fixed number, and  $\varepsilon_X$  has i.i.d. components.

The parameter  $\lambda_2$  can be chosen according to a spectral clustering property and  $\lambda_1$  then remains as the only tuning parameter as for the  $\ell_1$ -norm regularization scheme with the Lasso. The result in Theorem 1 is based on an analysis in the transformed model with  $\tilde{Y}$  and  $\tilde{X}$  in (10): we can trade-off between the behavior of the singular values of  $\tilde{X}$  and an inflation of the transformed error  $K_{\lambda_2}^{1/2}\varepsilon$ . It involves recent results about the behavior of singular or spectral values in large random matrices.

The estimation strategy with the Lava estimator in (10) and its justification in Theorem 1 for the hidden confounder model has major implications in practice, including also the broad area of genome-wide association studies where accounting for sub-population structure is important.

## 5 Predictive stability, invariance and causal regularization

We considered in Section 4 a problem where the direction of an association is known to point from  $X$  to  $Y$ . Even when having confounding structure, one can then essentially identify the causal regression effect in high-dimensional problems.

For cases where the direction of an association is not known, one needs more to identify the direction of an association and eventually the causal regression parameter  $\beta^0$  in (3). Heterogeneity and perturbations help towards identifiability of directions and causality. Instrumental variable models as in (7), originating from economics [Geary \[1949\]](#), are now popular in other fields as well. And indeed, if  $r = \dim(A) \geq d = \dim(X)$ , they typically lead to identifiability of the directed associations and causal parameter  $\beta^0$  from  $X$  to  $Y$ . The key idea relies on the fact that  $A$  and  $\varepsilon_Y$  are independent in the model (6): hence we should choose a regression parameter  $\beta$  such that

$$(12) \quad \mathbb{E}[A(Y - X^T \beta)]\|_q = 0 \text{ for some } q \geq 1.$$

For  $\beta = \beta^0$ , the equality in (12) holds, and if  $A$  is sufficiently rich, necessarily requiring that  $r \geq d$ ,  $\beta^0$  is the only solution satisfying (12).

But assuming a structure as in the instrumental variable model in (7) is often more an uncertain bet than a realistic assumption. The model in (4) or (5) relaxes this restriction substantially. In addition, we do not assume to have  $r \geq d$  (or more precisely that the number of children of  $A$  is larger or equal to  $d$ ). “Everything” is possible, including cycles, except that the “anchor”  $A$  is exogenous (meaning that it is a source node in the graph). In such a general setting, the causal parameter  $\beta^0$  in (3) is typically not identifiable.

**5.1 Causal regularization.** It seems natural in general to have a soft version of the constraint in (12). A regression method can be equipped with an additional regularization term:

$$(13) \quad \hat{\beta} = \hat{\beta}_{\gamma;q} = \operatorname{argmin}_{\beta \in \mathbb{R}^d} (\|Y - X\beta\|_2^2/n + \gamma \|A^T(Y - X\beta)/n\|_q^2 + \lambda \|\beta\|_1),$$

where  $\|\beta\|_1$  is a regularizer for high-dimensionality; typical choices are  $q = 2$  or  $q = \infty$ . Here  $A$  is the  $n \times r$  matrix of the observed variables of  $A$ . For  $\gamma = 0$ , we get the usual penalized regression estimator while for  $\gamma = \infty$  we enforce the finite-sample version of the restriction in (12). The latter restriction might not be possible to be fulfilled for any  $\beta$  and thus,  $\gamma = \infty$  might not be appropriate. The question then becomes: what are the properties of such an estimator in (13) in general? We address this in the next section.

**5.2 Predictive stability and invariance under shift interventions: the population case.** We consider the problem of predictive stability and invariance of residuals under a class of shift interventions. For example: in the instrumental variable model in (7), the residuals  $Y - X\beta^0 = H\delta + \varepsilon_Y$  are invariant under any interventions/perturbations on  $X$  which leave the structure and parameters in the structural equation model in (4) unchanged. That is: the causal parameter leads to predictions whose errors remain invariant under arbitrary perturbation scenarios on  $X$ .

We want to understand such invariance and predictive stability in the general model (4) or (5) where  $A$  are not instruments (they can point to  $X$ ,  $H$  or  $Y$ ; and feedback cycles are allowed). For this, we restrict ourselves to shift interventions. We consider shifts  $v$ , a  $(1 + d + r)$ -dimensional vector being deterministic or random, which can act on any of the variables  $Y, X, H$  (but not on  $A$ ): the shifted random variables  $(Y^v, X^v, H^v, A)$  are given as the solution of

$$(14) \quad \begin{pmatrix} Y^v \\ X^v \\ H^v \end{pmatrix} = B \begin{pmatrix} Y^v \\ X^v \\ H^v \end{pmatrix} + v + MA + \varepsilon.$$

In the random case we assume that  $v$  is independent of  $\varepsilon$  and  $A$ . A shift intervention  $v$  acts as a shift  $v_k$  on the component  $(Y, X, H)_k$  (for all  $k$  where  $v_k \neq 0$ ) and such shifts  $v_k$  are propagated through the SEM, changing the distribution of other components  $(Y, X, H)_j$  for  $j \neq k$ . In the alternative form analogous to (5) we can write

$$(15) \quad \begin{pmatrix} Y^v \\ X^v \\ H^v \end{pmatrix} = (I - B)^{-1}(v + MA + \varepsilon).$$

We now consider a class of shifts of the form

$$C_\gamma^q = \{v; v = M\delta \text{ for some } \delta \text{ with } \|\delta\|_q \leq \gamma\}.$$

Thus,  $C_\gamma^q$  includes shifts in the span of  $M$  which have at most a certain strength, measured by the  $\ell_q$ -norm of the coefficient vector  $\delta$  in the representation of the shift. We then see that the term  $v + MA$  in (14) or (15) becomes:  $v + MA = M(\delta + A)$  which intrinsically links the shift  $v$  to a perturbation of  $A$  of the form  $A + \delta$ .

The following fundamental result connects some worst case risk to causal regularization.

**Theorem 2.** *Rothenhäusler, Bühlmann, Meinshausen, and Peters [2018] Consider the model in (4) or (5) with  $(I - B)$  being invertible. Then, for any  $p, q \geq 1$  with  $p^{-1} + q^{-1} = 1$ , and for any  $b \in \mathbb{R}^d$ :*

$$\max_{v \in C_\gamma^q} \mathbb{E}[|Y^v - X^v b|^2] = \mathbb{E}[|Y - Xb|^2] + \gamma \|\mathbb{E}[A(Y - Xb)]\|_p^2.$$

Theorem 2 has important consequences on predictive stability. First of all, since the result holds for any  $b \in \mathbb{R}^d$ , we can consider the argmin on both sides of the equality:

$$(16) \quad \begin{aligned} b_{\gamma;q} &= \operatorname{argmin}_{\beta \in \mathbb{R}^d} \max_{v \in C_\gamma^q} \mathbb{E}[|Y^v - X^v b|^2] \\ &= \operatorname{argmin}_{\beta \in \mathbb{R}^d} (\mathbb{E}[|Y - X\beta|^2] + \gamma \|\mathbb{E}[A(Y - X\beta)]\|_p^2). \end{aligned}$$

Thus, the optimizer for the worst case risk within the class  $C_Y^q$  equals exactly the one from the regularized criterion for the non-shifted variables. The interpretation is as follows: the shifted variables  $(Y^v, X^v)$  can be associated to test data where some shift perturbations have occurred, while the non-shifted variables  $(Y, X)$  are the ones from the non-perturbed training data. Therefore, we obtain predictive stability and protection against some worst case shifts on new test data.

The corner-point of the regularization on the right-hand side of (16) is with  $\gamma = \infty$ :  $b_\infty$  is the parameter  $b$  in

$$I = \{b; \mathbb{E}[A(Y - X^T b)] = 0\}$$

which minimizes the squared error risk. All the elements of  $I$  lead to an interesting invariance.

**Theorem 3.** *Rothenhäusler, Bühlmann, Meinshausen, and Peters [ibid.] Consider the model in (4) or (5) with  $(I - B)$  being invertible. Then:*

$$b \in I \iff Y - Xb \stackrel{d}{=} Y^v - X^v b \text{ for all } v \text{ in } \text{span}(M).$$

The result says that enforcing the constraint  $\mathbb{E}[A(Y - X^T b)] = 0$  leads to invariance of the error terms, and protection against any shifts in  $\text{span}(M)$  is guaranteed.

**5.3 Properties of high-dimensional anchor regression for finite samples.** We consider here the finite-sample estimator in (13). We make the following assumptions for the high-dimensional setting.

**(A1)**  $A, X, Y$  are jointly Gaussian, and the minimal eigenvalue of  $\text{Cov}(X)$  satisfies  $\Lambda_{\min}^2(\text{Cov}(X)) L > 0$ ;

**(A2)**  $S_0(\gamma; q) = \text{supp}(b_{\gamma; q})$  has cardinality  $s_0(\gamma; q) = o(\sqrt{\log(d)/n})$  ( $d \gg n \rightarrow \infty$ );

The Gaussian assumption in (A1) is only for technical simplicity and extensions to sub-Gaussian distributions are possible.

**Theorem 4.** *Rothenhäusler, Bühlmann, Meinshausen, and Peters [ibid.] Assume the model in (4) or (5) and that (A1)-(A2) hold. Consider the estimator in (13). Then, with probability tending to one as  $d \gg n \rightarrow \infty$ , we have that for any  $\gamma \geq 0$ :*

$$\begin{aligned} \|\hat{\beta}_{\gamma; q} - b_{\gamma; q}\|_1 &\leq C \lambda s_0(\gamma; q) \quad (q \in \{2, \infty\}), \\ \text{for } q = 2: \lambda &\asymp \sqrt{r \max(\log(r), \log(d))/n}, \\ \text{for } q = \infty: \lambda &\asymp \sqrt{\log(r) \log(d)/n}. \end{aligned}$$

where  $0 < C < \infty$  is a constant. Furthermore, for the risk  $R(v, \beta) = \mathbb{E}[|Y^v - X^v \beta|^2]$  with shift  $v$ , we have that

$$\max_{v \in C_\gamma^q} R(v, \hat{\beta}_{\gamma;q}) \leq \max_{v \in C_\gamma^q} R(v, b_{\gamma;q}) + C g(\lambda) s_0(\gamma; q),$$

where  $g(\lambda) = \lambda^2$  for  $q = 2$  and  $g(\lambda) = \lambda$  for  $q = \infty$ , with  $\lambda$  as specified above for  $q \in \{2, \infty\}$ . Note that  $\max_{v \in C_\gamma^q} R(v, b_{\gamma;q}) = \min_\beta \max_{v \in C_\gamma^q} R(v, b_{\gamma;q})$ .

For the case with high-dimensional anchors with  $r \gg n$  we should choose  $q = \infty$ . For small values of  $r = \dim(A)$ ,  $q = 2$  seems to be a more natural choice in terms of the class  $C_\gamma^q$  for protection or predictive stability, see Theorems 2 and 3.

**5.4 Predicting single gene knock out experiments.** As described in Sections 5.1–5.3, the methodology and corresponding theory is tailored for predictive stability and prediction of new unseen perturbations. A score for measuring the effect-strength of a perturbation at covariate  $X_j$  for the response  $Y$  is given by the parameter  $(b_{\gamma;q})_j$  in (16). Note that for  $\gamma = \infty$  and in identifiable scenarios,  $(b_{\gamma=\infty})_j = \beta_j^0$  equals the direct causal effect in (3).

We summarize some findings for predicting single gene knock out experiments in yeast (*Saccharomyces cerevisiae*) [Meinshausen, Hauser, Mooij, Peters, Versteeg, and Bühlmann \[2016\]](#). The observed data is for the expressions of 6170 genes in yeast, and there are  $n_{\text{observ}} = 160$  observational data points (from the system in steady state, without any interventions, from wild-type yeast) and  $n_{\text{interv}} = 1479$  interventional data points, each of them corresponding to a single gene knock out experiment where a single strain has been deleted. The response  $Y$  is the expression of (say) gene  $k$ , and the covariates correspond to the expressions of all the genes without gene  $k$ , thus being of dimension  $d = 6170 - 1 = 6169$ . Consider this encoding into response and covariates for all  $k = 1, 2, \dots, 6170$ , that is, the expression of each gene is once the response: and thus, we can predict the effect-strength of a perturbation at each gene to another one. The model in (4) or (5) is used with two environments corresponding to  $r = 2$  binary components in  $A$ :  $A_1$  encodes the observational data environments with 160 samples,  $A_2$  is encoding all interventional data by (crudely) pooling all of them into a single environment with 1479 samples.

Holding out a random third (repeatedly three times) of the 1479 interventional samples enables us to validate the predictions. We aim to predict a response  $Y$  under an intervention at one of the covariates  $X_j$  in the hold-out data: the parameters for the prediction are trained (estimated) based on the 160 observational and two thirds of the interventional samples. Thanks to the hold-out data, we can then validate the performance of the prediction. We consider binarized outcomes: if the prediction for  $Y$  is large (in absolute value), we denote it as a predicted “positive” and otherwise as a predicted “negative”. Similarly

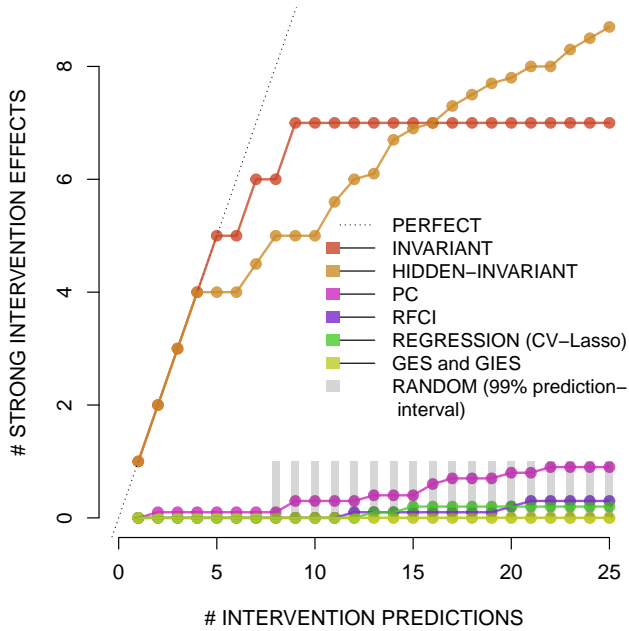


Figure 1: Prediction of single gene perturbations. x-axis: number of false positives; y-axis: number of true positives. Red (no hidden variables) and orange (including hidden variables) lines are algorithms exploiting (near) invariance similarly as described in Theorem 3. Other colored lines correspond to some competitor methods, and the gray bars indicate random guessing. The figure is taken from [Meinshausen, Hauser, Mooij, Peters, Versteeg, and Bühlmann \[2016\]](#).

for the true value in the hold-out observational data: if an intervention at a covariate has a strong effect on  $Y$ , we denote it as “true”, otherwise as false. One can then validate methods and algorithms in terms of their capacity to predict “true positives” (predicted “positive” and actually being “true”) in relation to “false positives” (predicted “positive” but actually being “false”). Figure 1 summarizes the results. The problem of correctly predicting unseen gene interventions is very ambitious: we predict only a few strong intervention effects, but the highest scoring predictions (the first 5 or 4, respectively) are all correct (i.e., “true”).

## 6 Conclusions

Large-scale data with heterogeneity from different environments or perturbations provides novel opportunities for predictive stability and causal inference. The word “causal” is ambitious and perhaps a bit philosophical: in a nutshell, its meaning is to predict the outcome of an unseen (in the data) perturbation, a policy or treatment. This fundamental prediction problem is very different from the standard one where we want to predict an outcome from roughly the same population from which we have collected data. We are now just at the beginning of new developments of methodology, algorithms and fundamental mathematical understanding for statistical inference from heterogeneous large-scale data.

## References

- P. Bickel, Y. Ritov, and A. Tsybakov (2009). “Simultaneous analysis of Lasso and Dantzig selector.” *Annals of Statistics* 37, pp. 1705–1732 (cit. on pp. 2804, 2812).
- P. Bühlmann (2006). “Boosting for high-dimensional linear models”. *Annals of Statistics* 34, pp. 559–583 (cit. on p. 2804).
- P. Bühlmann and S. van de Geer (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer (cit. on p. 2804).
- F. Bunea, A. Tsybakov, and M.H. Wegkamp (2007). “Sparsity oracle inequalities for the Lasso”. *Electronic Journal of Statistics* 1, pp. 169–194 (cit. on p. 2804).
- E. Candès and T. Tao (2007). “The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$  (with discussion)”. *Annals of Statistics* 35, pp. 2313–2404 (cit. on p. 2804).
- E.J. Candès, J.K. Romberg, and T. Tao (2006). “Stable signal recovery from incomplete and inaccurate measurements”. *Communications on Pure and Applied Mathematics* 59, pp. 1207–1223 (cit. on p. 2804).
- E.J. Candès and T. Tao (2006). “Near-optimal signal recovery from random projections: Universal encoding strategies?” *IEEE Transactions on Information Theory* 52, pp. 5406–5425 (cit. on p. 2804).
- D. Čevič, P. Bühlmann, and N. Meinshausen (2018). *Work in progress* (cit. on p. 2812).
- S.S. Chen, D.L. Donoho, and M.A. Saunders (2001). “Atomic decomposition by basis pursuit”. *SIAM review* 43, pp. 129–159 (cit. on p. 2804).
- V. Chernozhukov, C. Hansen, Y. Liao, et al. (2017). “A lava attack on the recovery of sums of dense and sparse signals”. *Annals of Statistics* 45, pp. 39–76 (cit. on p. 2811).
- D.L. Donoho (1993). “Nonlinear wavelet methods for recovery of signals, densities, and spectra from indirect and noisy data”. In: *In Proceedings of Symposia in Applied Mathematics* (cit. on p. 2804).
- (2006). “Compressed sensing”. *IEEE Transactions on Information Theory* 52, pp. 1289–1306 (cit. on p. 2804).

- D.L. Donoho and X. Huo (2001). “Uncertainty principles and ideal atomic decomposition”. *IEEE Transactions on Information Theory* 47, pp. 2845–2862 (cit. on p. 2804).
- D.L. Donoho and J.M. Johnstone (1994). “Ideal spatial adaptation by wavelet shrinkage”. *Biometrika* 81, pp. 425–455 (cit. on p. 2804).
- R.C. Geary (1949). “Determination of linear relations between systematic parts of variables with errors of observation the variances of which are unknown”. *Econometrica: Journal of the Econometric Society*, pp. 30–58 (cit. on p. 2813).
- C. Giraud (2014). *Introduction to High-Dimensional Statistics*. CRC Press (cit. on p. 2804).
- E. Greenshtein and Y. Ritov (2004). “Persistence in high-dimensional predictor selection and the virtue of over-parametrization”. *Bernoulli* 10, pp. 971–988 (cit. on p. 2804).
- T. Haavelmo (1943). “The statistical implications of a system of simultaneous equations”. *Econometrica*, pp. 1–12 (cit. on pp. 2803, 2807).
- T. Hastie, R. Tibshirani, and M. Wainwright (2015). *Statistical Learning with Sparsity: the Lasso and Generalizations*. CRC Press (cit. on p. 2804).
- V. Koltchinskii (2009a). “Sparsity in penalized empirical risk minimization”. *Annales de l’Institut Henri Poincaré, Probabilités et Statistiques* 45, pp. 7–57 (cit. on p. 2804).
- (2009b). “The Dantzig selector and sparsity oracle inequalities”. *Bernoulli* 15, pp. 799–828 (cit. on p. 2804).
- N. Meinshausen and P. Bühlmann (2006). “High-dimensional graphs and variable selection with the Lasso”. *Annals of Statistics* 34, pp. 1436–1462 (cit. on p. 2804).
- N. Meinshausen, A. Hauser, J.M. Mooij, J. Peters, P. Versteeg, and P. Bühlmann (2016). “Methods for causal inference from gene perturbation experiments and validation”. *Proc. National Academy of Sciences USA* 113, pp. 7361–7368 (cit. on pp. 2805, 2816, 2817).
- J. Pearl (2000). *Causality: Models, Reasoning and Inference*. Cambridge University Press (cit. on pp. 2803, 2806, 2807).
- J. Peters, P. Bühlmann, and N. Meinshausen (2016). “Causal inference using invariant prediction: identification and confidence interval (with discussion)”. *J. Royal Statistical Society, Series B* 78, pp. 947–1012 (cit. on pp. 2803, 2807).
- J. Peters, D. Janzing, and B. Schölkopf (2017). *Elements of Causal Inference*. MIT Press (cit. on p. 2806).
- N. Pfister, P. Bühlmann, and J. Peters (2017). *Invariant causal prediction for sequential data*. arXiv: 1706.08058 (cit. on p. 2805).
- D. Rothenhäusler, P. Bühlmann, N. Meinshausen, and J. Peters (2018). *Anchor regression: heterogeneous data meets causality*. arXiv: 1801.06229 (cit. on pp. 2814, 2815).
- R. Tibshirani (1996). “Regression shrinkage and selection via the Lasso”. *Journal of the Royal Statistical Society, Series B* 58, pp. 267–288 (cit. on p. 2804).



- S. van de Geer and P. Bühlmann (2009). “On the conditions used to prove oracle results for the Lasso”. *Electronic Journal of Statistics* 3, pp. 1360–1392 (cit. on p. [2812](#)).
- P. Zhao and B. Yu (2006). “On model selection consistency of Lasso”. *Journal of Machine Learning Research* 7, pp. 2541–2563 (cit. on p. [2804](#)).
- H. Zou (2006). “The adaptive Lasso and its oracle properties”. *Journal of the American Statistical Association* 101, pp. 1418–1429 (cit. on p. [2804](#)).

Received 2017-11-24.

PETER BÜHLMANN  
SEMINAR FOR STATISTICS  
DEPARTMENT OF MATHEMATICS  
ETH ZÜRICH  
SWITZERLAND  
[peter.buehlmann@stat.math.ethz.ch](mailto:peter.buehlmann@stat.math.ethz.ch)

# PLANAR ISING MODEL AT CRITICALITY: STATE-OF-THE-ART AND PERSPECTIVES

DMITRY CHELKAK

## Abstract

In this essay, we briefly discuss recent developments, started a decade ago in the seminal work of Smirnov and continued by a number of authors, centered around the conformal invariance of the critical planar Ising model on  $\mathbb{Z}^2$  and, more generally, of the critical  $\mathbb{Z}$ -invariant Ising model on isoradial graphs (rhombic lattices). We also introduce a new class of embeddings of general weighted planar graphs (s-embeddings), which might, in particular, pave the way to true universality results for the planar Ising model.

## 1 Introduction

The two-dimensional Ising model, introduced by Lenz almost a hundred years ago, does not need an introduction, being probably the most famous example of a statistical mechanics system exhibiting the phase transition and the conformally invariant behavior at criticality, as well as an inspiring structure of the correlation functions both at the critical point and in a vicinity of it; e.g., see [McCoy and Wu \[1973\]](#) and [Mussardo \[2010\]](#). More recently, it became a playground for mathematicians interested in a rigorous understanding of the conjectural conformal invariance of critical lattice systems, see [Smirnov \[2006\]](#).

What makes the *planar* (a priori, not necessarily critical) Ising model particularly feasible for such a mathematical treatment (in absence of the magnetic field) is the underlying structure of *s-holomorphic spinors* (aka *fermionic observables*), essentially dating back to the work of Onsager and Kaufman and reinterpreted several times since then, notably by [Kadanoff and Ceva \[1971\]](#). From the classical analysis (or probability theory) perspective, these s-holomorphic spinors can be thought of as counterparts of discrete harmonic functions associated to random-walk-based systems. The main theme of this note is recent

---

Holder of the ENS-MHI chair funded by MHI.

MSC2010: primary 82B20; secondary 30G25, 60J67, 81T40.

Keywords: 2D Ising model, conformal invariance, s-holomorphicity, s-embeddings.

convergence results for the critical model based on an analysis of such observables in the small mesh size limit.

The text should *not* be considered as a survey: certainly, many results and references deserving to be mentioned in such a survey are missing. Its purposes are rather to give a general informal description of the current state-of-the-art of the subject (from the personal viewpoint of the author and to the best of his knowledge) for readers not interested in technical details; to provide some references for those interested; and to indicate several ongoing research directions and open questions.

I wish to thank my co-authors as well as many other colleagues for numerous helpful discussions (centered at Geneva a decade ago, worldwide nowadays) of the planar Ising model and for sharing their knowledge and ideas with others. Clearly, the progress described below was achieved by collective efforts and it is a privilege to discuss the results of hard work of a rather big community in this *essay*.

## 2 Discrete spinors and s-holomorphicity in the planar Ising model

**2.1 Notation and the Kramers–Wannier duality.** Below we consider the ferromagnetic Ising model on *faces* of a graph  $G$  embedded into the complex plane  $\mathbb{C}$  so that all its edges are straight segments. One can also work with graphs embedded into Riemann surfaces but for simplicity we prefer not to discuss such a generalization here (see [Chelkak, Cimasoni, and Kassel \[2017, Section 4\]](#) and references therein for more details). A *spin configuration*  $\sigma$  is an assignment of a  $\pm 1$  spin  $\sigma_u$  to each face of  $G$ , including the outer face  $u_{\text{out}}$ , with the spin  $\sigma_{u_{\text{out}}}$  playing the role of boundary conditions. The probability measure  $\mathbb{P}^\circ$  (the superscript  $\circ$  emphasizes the fact that the model is considered on faces of  $G$ ) on the set of spin configurations is given by

$$(2.1) \quad \mathbb{P}^\circ(\sigma) = (\mathcal{Z}^\circ(G))^{-1} \exp \left[ \beta \sum_e J_e \sigma_{u_-(e)} \sigma_{u_+(e)} \right],$$

where  $\beta > 0$  is the inverse temperature,  $J_e > 0$  are fixed interaction constants, the summation is over unoriented edges of  $G$  (an edge  $e$  separates two faces  $u_\pm(e)$ ), and  $\mathcal{Z}^\circ(G)$  is the normalization constant called the *partition function*. Note that the spin of the outer face  $u_{\text{out}}$  of  $G$  plays the role of boundary conditions and one can always break the  $\mathbb{Z}_2$  (spin-flip) symmetry of the model by assuming  $\sigma_{u_{\text{out}}} = +1$ .

Abusing the notation slightly, we also admit the situation when  $J_e = 0$  along some boundary arcs of  $G$ , which means that the corresponding near-to-boundary spins do not interact with  $\sigma_{u_{\text{out}}}$ . We call these parts of the boundary of  $G$  *free arcs* and use the name *wired arcs* for the remaining parts, across which the inner spins interact with the *same*  $\sigma_{u_{\text{out}}} = +1$ .

We will use the name *standard boundary conditions* for this setup, see also [Chelkak, Hongler, and Izzyurov \[n.d.\]](#). For simplicity, below we always assume that there exists at least one wired arc.

We denote by  $G^\bullet$  the graph obtained from  $G$  by removing the edges along free parts of the boundary and by  $G^\circ$  a graph dual to  $G$  with the following convention: instead of a single vertex corresponding to  $u_{\text{out}}$  we draw one vertex per boundary edge on wired arcs. Combinatorially, *all* these vertices of  $G^\circ$  should be thought of as wired together (hence the name) and, similarly, the vertices of  $G^\bullet$  along free parts of the boundary should be thought of as a collection of ‘macro-vertices’, one per free arc. We assume that  $G^\circ$  is also embedded into  $\mathbb{C}$  so that all its edges are straight segments and denote by  $\diamond(G)$  the set of quads  $(vuv'u')$  formed by pairs of (unoriented) dual edges  $(vv')$  and  $(uu')$  of  $G^\bullet$  and  $G^\circ$ , respectively. We also denote by  $\partial\diamond(G)$  the set of triangles  $(vuv')$  and  $(uvu')$  arising instead of quads along free and wired boundary arcs, respectively, and set  $\overline{\diamond(G)} := \diamond(G) \cup \partial\diamond(G)$ . Finally, let  $\Omega(G) \subset \mathbb{C}$  be the polygon formed by all these quads and triangles.

For an unoriented edge  $e$  of  $G^\bullet$  (or, equivalently, an element of  $\diamond(G)$ ), we define  $x_e := \exp[-2\beta J_e]$  and extend this notation to the elements of  $\partial\diamond(G)$  by setting  $x_e := 1$  on free arcs and  $x_e := 0$  on wired ones. For a subset  $C \subset \diamond(G)$ , denote  $x(C) := \prod_{e \in C} x_e$  and let  $\mathcal{E}(G)$  denote the set of all even subgraphs of  $G$ . There exists a trivial bijection of this set and the set of spin configurations on faces of  $G$ : draw edges separating misaligned spins. In particular, one sees that

$$(2.2) \quad \mathcal{Z}^\circ(G) = \prod_{e \in \diamond(G)} x_e^{-1/2} \cdot \mathcal{Z}(G), \quad \text{where} \quad \mathcal{Z}(G) := \sum_{C \in \mathcal{E}(G)} x(C),$$

this is called the *domain walls* (or low-temperature) *expansion* of  $\mathcal{Z}^\circ(G)$ . A remarkable fact (first observed by van der Waerden) is that the *same*  $\mathcal{Z}(G)$  also gives an expression for the partition function  $\mathcal{Z}^\bullet(G)$  of the Ising model on *vertices* of  $G^\bullet$ , provided that the dual parameters  $\beta^\bullet$  and  $J_e^\bullet$  satisfy the identity  $x_e = \tanh[\beta^\bullet J_e^\bullet]$ . Namely, the following *high-temperature expansion* of  $\mathcal{Z}^\bullet(G)$  holds true:

$$(2.3) \quad \mathcal{Z}^\bullet(G) = 2^{|V(G^\bullet)|} \prod_{e \in \diamond(G)} (1 - x_e^2)^{-1/2} \cdot \mathcal{Z}(G), \quad \mathcal{Z}(G) = \sum_{C \in \mathcal{E}(G)} x(C).$$

This link between the Ising models on  $G^\circ$  and  $G^\bullet$  is called the *Kramers–Wannier duality* and it is not limited to the equality between the partition functions  $\mathcal{Z}^\circ(G)$  and  $\mathcal{Z}^\bullet(G)$ . Nevertheless, it is worth mentioning that similar objects typically lead to different types of sums in the two representations. E.g., in order to compute the expectation  $\mathbb{E}^\circ[\sigma_u \sigma_{u'}]$  similarly to (2.2) one should keep track of the parity of the number of loops in  $C$  separating  $u$  and  $u'$ , while computing the expectation  $\mathbb{E}^\bullet[\sigma_v \sigma_{v'}]$  amounts to the replacement of  $\mathcal{E}(G)$  in (2.3) by the set  $\mathcal{E}(G; v, v')$  of subgraphs of  $G$  in which  $v$  and  $v'$  (but no other vertex) have odd degrees. Also, note that the boundary conditions on  $G^\circ$  and  $G^\bullet$  are *not*

fully symmetric: dual spins on different free arcs are not required to coincide, contrary to the wired ones.

It is convenient to introduce the following parametrization  $\theta_e$  of the weights  $x_e$ :

$$(2.4) \quad x_e = \exp[-2\beta J_e] = \tanh[\beta^\bullet J_e^\bullet] = \tan \frac{1}{2}\theta_e, \quad \theta_e \in [0, \frac{\pi}{2}].$$

Note that, if we similarly define  $\theta_e^\bullet$  so that  $\tan \frac{1}{2}\theta_e^\bullet = \exp[-2\beta^\bullet J_e^\bullet]$ , then  $\theta_e^\bullet = \frac{\pi}{2} - \theta_e$ .

**2.2 Spins-disorders formalism of Kadanoff and Ceva.** Following [Kadanoff and Ceva \[1971\]](#), we now describe the so-called *disorder insertions* – the objects dual to spins under the Kramers–Wannier duality. We also refer the interested reader to recent papers of Dubédat, see [Dubédat \[2011\]](#) and [Dubédat \[2011\]](#). Let  $m$  be even and  $v_1, \dots, v_m$  be a collection of vertices of  $G^\bullet$  (with the standard convention that each free arc should be considered as a single vertex). Let us fix a collection of paths  $\chi^{[v_1, \dots, v_m]}$  on  $G$  linking these vertices pairwise and change the interaction constants  $J_e$  to  $-J_e$  along these paths to get another probability measure on the spin configurations instead of (2.1). Note that one can think of this operation as putting an additional random weight  $\exp[-2\beta J_{uu'}\sigma_u\sigma_{u'}]$  along  $\gamma^{[v_1, \dots, v_m]}$  and treat this weight as a random variable, which we denote by  $\mu_{v_1} \dots \mu_{v_m}$  (note that its definition implicitly depends on the choice of disorder lines). The domain walls representation of the Ising model on  $G^\circ$  then gives

$$(2.5) \quad \mathbb{E}^\circ[\mu_{v_1} \dots \mu_{v_m}] = x(\gamma^{[v_1, \dots, v_m]}) \cdot \sum_{C \in \mathcal{E}(G)} x^{[v_1, \dots, v_m]}(C) / \mathcal{Z}(G),$$

where the weights  $x^{[v_1, \dots, v_m]}$  are obtained from  $x$  by changing  $x_e$  to  $x_e^{-1}$  on disorder lines and the first factor comes from the prefactor in (2.2). More invariantly, one can consider the sign-flip-symmetric Ising model on *faces* of the double-cover  $G^{[v_1, \dots, v_m]}$  of  $G$  ramified over the vertices  $v_1, \dots, v_m$  (the spins at two faces of  $G^{[v_1, \dots, v_m]}$  lying over the same face of  $G$  are required to have opposite values) and rewrite (2.5) as

$$(2.6) \quad \begin{aligned} \mathbb{E}^\circ[\mu_{v_1} \dots \mu_{v_m}] &= \mathcal{Z}^\circ(G^{[v_1, \dots, v_m]}) / \mathcal{Z}^\circ(G) \\ &= \sum_{C \in \mathcal{E}(G; v_1, \dots, v_m)} x(C) / \mathcal{Z}(G) = \mathbb{E}^\bullet[\sigma_{v_1} \dots \sigma_{v_m}], \end{aligned}$$

where  $\mathcal{E}(G; v_1, \dots, v_m)$  stands for the set of subgraphs of  $G$  in which all  $v_1, \dots, v_m$  (but no other vertex) have odd degrees; the last equality is the classical high-temperature expansion of spin correlations on  $G^\bullet$  mentioned above.

Vice versa, given an even  $n$  and a collection of faces  $u_1, \dots, u_n$  of  $G$ , one can write

$$(2.7) \quad \begin{aligned} \mathbb{E}^\circ[\sigma_{u_1} \dots \sigma_{u_n}] &= \sum_{C \in \mathcal{E}(G)} x_{[u_1, \dots, u_n]}(C) / \mathcal{Z}(G) \\ &= \mathcal{Z}^\bullet(G_{[u_1, \dots, u_n]}) / \mathcal{Z}^\bullet(G) = \mathbb{E}^\bullet[\mu_{u_1} \dots \mu_{u_n}], \end{aligned}$$

where the weights  $x_{[u_1, \dots, u_n]}$  are obtained from  $x$  by putting additional minus signs on the edges of  $\gamma_{[u_1, \dots, u_n]}$  and  $\mathbb{Z}^\bullet(G_{[u_1, \dots, u_n]})$  denotes the partition function of the spin-flip symmetric Ising model on *vertices* of the double-cover  $G_{[u_1, \dots, u_n]}$  of  $G$  ramified over faces  $u_1, \dots, u_m$ , treated via the high-temperature expansion. Generalizing (2.6) and (2.7), one has the following duality between spins and disorders:

$$(2.8) \quad \mathbb{E}^\circ[\mu_{v_1} \dots \mu_{v_m} \sigma_{u_1} \dots \sigma_{u_n}] = \mathbb{E}^\bullet[\sigma_{v_1} \dots \sigma_{v_m} \mu_{u_1} \dots \mu_{u_n}]$$

since both sides are equal to  $\sum_{C \in \mathcal{E}(G; v_1, \dots, v_n)} x_{[u_1, \dots, u_n]}(C) / \mathbb{Z}(G)$ . Let us emphasize that one needs to fix disorder lines in order to interpret these quantities as expectations with respect to  $\mathbb{P}^\circ$  and  $\mathbb{P}^\bullet$ , respectively. Below we prefer a more invariant approach and view  $u_q$ 's as faces of the double-cover  $G^{[v_1, \dots, v_n]}$  and the expectation taken with respect to the sign-flip symmetric Ising model defined on faces of this double-cover. To avoid possible confusion, we introduce the notation

$$(2.9) \quad \langle \mu_{v_1} \dots \mu_{v_m} \sigma_{u_1} \dots \sigma_{u_n} \rangle := \mathbb{E}_{G^{[v_1, \dots, v_m]}}^\circ[\sigma_{u_1} \dots \sigma_{u_n}]$$

instead of (2.8). Considered as a function of both  $v_p$ 's and  $u_q$ 's, (2.9) is defined on a double-cover  $G_{[\bullet, \circ]}^{m, n}$  of  $(G^\bullet)^m \times (G^\circ)^n$  and changes the sign each time when one of  $v_p$  turns around one of  $u_q$  or vice versa; we call such functions *spinors* on  $G_{[\bullet, \circ]}^{m, n}$ .

We also need some additional notation. Let  $\Lambda(G) := G^\bullet \cup G^\circ$  be the planar graph formed by the sides of quads from  $\diamond(G)$  and let  $\Upsilon(G)$  denote the medial graph of  $\Lambda(G)$ . In other words, the vertices of  $\Upsilon(G)$  correspond to edges  $(uv)$  of  $\Lambda(G)$  or to *corners* of  $G$ , while the faces of  $\Upsilon(G)$  correspond either to vertices of  $G^\bullet$  or to vertices of  $G^\circ$  or to quads from  $\diamond(G)$ . Denote by  $\Upsilon^\times(G)$  a double-cover of  $\Upsilon(G)$  which branches over each of its faces (e.g., see [Mercat \[2001, Fig. 27\]](#) or [Chelkak and Smirnov \[2012, Fig. 6\]](#)). Note that  $\Upsilon^\times(G)$  is fully defined by this condition for graphs embedded into  $\mathbb{C}$  but on Riemann surfaces there is a choice that can be rephrased as the choice of a *spin structure* on the surface. Below we discuss spinors on  $\Upsilon^\times(G)$ , i.e. the functions whose values at two vertices of  $\Upsilon^\times(G)$  lying over the same vertex of  $\Upsilon(G)$  differ by the sign. An important example of such a function is the *Dirac spinor*

$$(2.10) \quad \eta_c := \varsigma \cdot \exp[-\frac{i}{2} \arg(v(c) - u(c))], \quad \text{where } c = (u(c)v(c)) \in \Upsilon^\times(G),$$

$u(c) \in G^\circ$ ,  $v(c) \in G^\bullet$  and a global prefactor  $\varsigma : |\varsigma| = 1$  is added to the definition for later convenience. If  $G$  was embedded into a Riemann surface  $\Sigma$ , one should fix a vector field on  $\Sigma$  with zeroes of even index to define the  $\arg$  function (cf. [Chelkak, Cimasoni, and Kassel \[2017, Section 4\]](#)), in the case  $\Sigma = \mathbb{C}$  we simply consider a constant vector field  $\varsigma^2$ .

Given a corner  $c$  of  $G$ , we formally define  $\chi_c := \mu_{v(c)} \sigma_{u(c)}$ , i.e.

$$(2.11) \quad \langle \chi_c \mu_{v_1} \dots \mu_{v_{m-1}} \sigma_{u_1} \dots \sigma_{u_{n-1}} \rangle := \langle \mu_{v(c)} \mu_{v_1} \dots \mu_{v_{m-1}} \sigma_{u(c)} \sigma_{u_1} \dots \sigma_{u_{n-1}} \rangle.$$

According to the preceding discussion of mixed spins-disorders expectations, for a given collection  $\varpi := \{v_1, \dots, v_{m-1}, u_1, \dots, u_{n-1}\} \in (G^\bullet)^{m-1} \times (G^\circ)^{n-1}$ , the function (2.11) locally behaves as a spinor on  $\Upsilon^\times(G)$  but its global branching structure is slightly different as the additional sign change arises when  $c$  turns around one of  $v_p$  or  $u_q$ . Let us denote the corresponding double-cover of  $\Upsilon(G)$  by  $\Upsilon_\varpi^\times(G)$ . Finally, define  $\psi_c := \eta_c \chi_c$ , where  $\eta_c$  is defined by (2.10), and note that functions

$$(2.12) \quad \langle \psi_c \mu_{v_1} \dots \mu_{v_{m-1}} \sigma_{u_1} \dots \sigma_{u_{n-1}} \rangle := \eta_c \langle \chi_c \mu_{v_1} \dots \mu_{v_{m-1}} \sigma_{u_1} \dots \sigma_{u_{n-1}} \rangle$$

do not branch locally (because of the cancellation of sign changes of  $\chi_c$  and  $\eta_c$ ) and change the sign only when  $c$  turns around one of the vertices  $v_p$  or the faces  $u_q$ . We denote the corresponding (i.e., ramified over  $\varpi$ ) double-cover of  $\Upsilon(G)$  by  $\Upsilon_\varpi(G)$ .

**2.3 S-holomorphicity.** This section is devoted to the crucial three-term equation for the functions (2.11), the so-called *propagation equation* for spinors on  $\Upsilon^\times(G)$  (and also known as the Dotsenko–Dotsenko equation, see Mercat [2001] and Chelkak, Cimasoni, and Kassel [2017, Section 3.5]). To simplify the presentation, we introduce the following notation: for a quad  $z_e$  corresponding to an edge  $e$  of  $G$ , we denote its vertices, listed counterclockwise, by  $v_0^\bullet(z_e)$ ,  $v_0^\circ(z_e)$ ,  $v_1^\bullet(z_e)$ , and  $v_1^\circ(z_e)$ , where  $v_0^\bullet(z)$ ,  $v_1^\bullet(z) \in G^\bullet$  and  $v_0^\circ(z)$ ,  $v_1^\circ(z) \in G^\circ$  (the choice of  $v_0^\bullet(z)$  among the two vertices of  $G^\bullet$  is arbitrary). The corner of  $G$  corresponding to the edge  $(v_p^\bullet(z_e)v_q^\circ(z_e))$  of  $z_e$  is denoted by  $c_{pq}(z_e) \in \Upsilon(G)$ . For shortness, we also often omit  $z_e$  in this notation if no confusion arises.

**Definition 2.1.** A spinor  $F$  defined on  $\Upsilon^\times(G)$  or, more generally, on some  $\Upsilon_\varpi^\times(G)$  is called *s-holomorphic* if its values at any three consecutive (on  $\Upsilon_\varpi^\times(G)$ ) corners  $c_{p,1-q}(z_e)$ ,  $c_{pq}(z_e)$  and  $c_{1-p,q}(z_e)$  surrounding a quad  $z_e \in \diamond(G)$  satisfy the identity

$$(2.13) \quad F(c_{pq}) = F(c_{p,1-q}) \cos \theta_e + F(c_{1-p,q}) \sin \theta_e,$$

where  $\theta_e$  stands for the parametrization (2.4) of the Ising model weight  $x_e$  of  $e$ .

**Remark 2.1.** In fact, a straightforward computation shows that (2.13) implies the spinor property:  $F(c_{pq}^b) = -F(c_{pq}^\sharp)$  if  $c_{pq}^b, c_{pq}^\sharp \in \Upsilon_\varpi^\times(G)$  lie over the same corner  $c_{pq}$ .

The key observation is that all the Ising model observables of the form (2.11), considered as functions of  $c \in \Upsilon_\varpi^\times(G)$ , satisfy the propagation equation (2.13) (e.g., see Chelkak, Cimasoni, and Kassel [ibid., Section 3.5]). In the recent research, this equation was mostly used in the context of isoradial graphs, in which case the parameter  $\theta_e$  has also a direct geometric meaning, but in fact (2.13) is fairly abstract. In particular, Definition 2.1 does not rely upon a particular choice (up to a homotopy) of an embedding of  $\diamond(G)$  into  $\mathbb{C}$ . Contrary to (2.13), which was known for decades, the next definition first appeared in the work of Smirnov on the critical Ising model on  $\mathbb{Z}^2$ , see Smirnov [2006] and Smirnov [2010a].

**Definition 2.2.** Let  $F$  be an  $s$ -holomorphic spinor on  $\Upsilon_{\varpi}^{\times}(G)$ . Then one can define a function  $H_F$  on  $\Lambda(G)$  by specifying its increment around each quad from  $\diamond(G)$  as

$$(2.14) \quad H_F(v_p^{\bullet}) - H_F(v_q^{\circ}) := (F(c_{pq}^{\#}))^2 = (F(c_{pq}^b))^2$$

Note that, independently of  $\varpi$ ,  $H_F$  is defined on  $\Lambda(G)$  and not on its double-cover.

*Remark 2.2.* Due to (2.13), one has  $(F(c_{00}))^2 + (F(c_{11}))^2 = (F(c_{01}))^2 + (F(c_{10}))^2$ . Thus,  $H_F$  is locally (and hence globally in the simply connected setup) well-defined.

A priori, the functions  $H_F$  do not seem to be natural objects for the study of correlations in the Ising model but they turned out to be absolutely indispensable for the analysis of scaling limits of such correlations in discrete approximations to general planar domains initiated in Smirnov [2006] and Smirnov [2010a], see Section 3.3 below.

**2.4 Pfaffian structure of fermionic correlators.** Similarly to (2.11), one can consider expectations containing two or more formal variables  $\chi_c$ . We start with the following observation: despite the fact that the quantities (2.9), viewed as functions on  $G_{[\bullet, \circ]}^{m, n}$ , are symmetric with respect to permutations of  $v_p$ , as well as to those of  $u_q$ , an additional sign change appears if one exchanges  $(u(c)v(c))$  and  $(u(d)v(d))$ ; this can be viewed as a cumulative result of a ‘half-turn’ of  $u(c)$  around  $v(d)$  and a ‘half-turn’ of  $u(d)$  around  $v(c)$ . In other words, the variables  $\chi_c$  and  $\chi_d$  *anti-commute*:

$$(2.15) \quad \langle \chi_d \chi_c \mu_{v_1} \dots \mu_{v_{m-2}} \sigma_{u_1} \dots \sigma_{u_{n-2}} \rangle = - \langle \chi_c \chi_d \mu_{v_1} \dots \mu_{v_{m-2}} \sigma_{u_1} \dots \sigma_{u_{n-2}} \rangle$$

if one considers both sides as a function of  $(c, d) \in (\Upsilon_{\varpi}^{\times}(G))^{[2]}$ , where  $(\Upsilon_{\varpi}^{\times}(G))^{[2]}$  denotes the set  $(\Upsilon_{\varpi}^{\times}(G))^2 \setminus \{(c, d) : c^{\#, b} = d^{\#, b} \text{ or } c^{\#, b} = d^{b, \#}\}$ . More generally, given a collection of vertices and faces  $\varpi = \{v_1, \dots, v_{m-k}, u_1, \dots, u_{n-k}\}$ , the quantities

$$(2.16) \quad \langle \chi_{c_1} \dots \chi_{c_k} \mathcal{O}_{\varpi}[\mu, \sigma] \rangle := \langle \chi_{c_1} \dots \chi_{c_k} \mu_{v_1} \dots \mu_{v_{m-k}} \sigma_{u_1} \dots \sigma_{u_{n-k}} \rangle$$

are anti-symmetric functions on  $(\Upsilon_{\varpi}^{\times}(G))^{[k]}$ ; see Chelkak, Hongler, and Izyurov [n.d.] for more precise definitions.

Another striking observation, which is also well known in the folklore for decades, is that the correlations (2.16) satisfy the Pfaffian identities: for an even number of pairwise distinct  $c_1, \dots, c_k$  corners of  $G$ , one has

$$(2.17) \quad \langle \chi_{c_1} \dots \chi_{c_k} \mathcal{O}_{\varpi}[\mu, \sigma] \rangle \cdot \langle \mathcal{O}_{\varpi}[\mu, \sigma] \rangle^{k/2-1} = \text{Pf}[\langle \chi_{c_r} \chi_{c_s} \mathcal{O}_{\varpi}[\mu, \sigma] \rangle]_{r,s=1}^k,$$

where the diagonal entries of the matrix on the right-hand side are set to 0.

One of the most transparent explanations of this Pfaffian structure comes from a remarkable fact that the partition function  $\mathbb{Z}(G)$  of the Ising model can be also written as



the Pfaffian of some (real, anti-symmetric) matrix  $\widehat{\mathcal{K}}$ , which is a simple transform of the famous *Kac–Ward matrix*  $\mathcal{KW}(G, x) := \text{Id} - \mathbf{T}$ , where

$$(2.18) \quad T_{e,e'} = \begin{cases} \exp[\frac{i}{2}(\arg(e') - \arg(e))] \cdot (x_e x_{e'})^{1/2} & \text{if } e' \text{ prolongates } e, \\ 0 & \text{otherwise,} \end{cases}$$

and  $e, e'$  are *oriented* edges of  $G$ . Namely (see [Chelkak, Cimasoni, and Kassel \[2017\]](#) for more details, including the interpretation via a relevant dimer model on the so-called Fisher graph  $G^F$ , and [Lis \[2016\]](#) for a streamlined version of classical arguments),

$$\widehat{\mathcal{K}} = i\mathbf{U}^* \mathbf{J} \cdot \mathcal{KW}(G, x) \cdot \mathbf{U}, \quad \text{where} \quad \mathbf{U} = \text{diag}(\eta_e), \quad \eta_e := \varsigma \cdot \exp[-\frac{i}{2} \arg(e)],$$

and  $\mathbf{J}$  is composed of  $2 \times 2$  blocks  $J_{e,e} = J_{\bar{e},\bar{e}} = 0$ ,  $J_{e,\bar{e}} = J_{\bar{e},e} = 1$ , indexed by pairs  $(e, \bar{e})$  of oppositely oriented edges of  $G$ . Note the similarity between the definition of  $\eta_e$  and (2.10): essentially, one can view the former as an arbitrarily chosen section of the latter, considered on oriented edges of  $G$  instead of  $\Upsilon(G)$ .

In other words, the Hamiltonian of the Ising model can be rewritten as a quadratic form in *Grassmann variables*  $\phi_e$  (aka free *fermions*), whose (formal) correlators satisfy Pfaffian identities by definition. Then one can check (e.g., see [Chelkak, Cimasoni, and Kassel \[2017, Theorem 1.2\]](#)) that these correlators of  $\phi_e$  admit essentially the same combinatorial expansions as the expectations (2.16). In fact, if all the vertices  $v(c_r)$  and  $v_p$  are pairwise distinct, then all the expectations involved in (2.17) can be viewed as (formal) correlators of some Grassmann variables  $\chi_c$  obtained from  $\phi_e$  by a local (namely, block diagonal with blocks indexed by vertices of  $G$ ) linear change, see [Chelkak, Cimasoni, and Kassel \[ibid., Section 3.4\]](#). Therefore, the Pfaffian identities (2.17) hold true if all  $v(c_r)$  and  $v_p$  are pairwise distinct and this assumption can be removed using the propagation equation (2.13), which is satisfied (with respect to each of  $c_r$ ) by both sides of (2.17).

### 3 Holomorphic observables in the critical model on isoradial graphs

**3.1 Critical Ising model on isoradial graphs (rhombic lattices).** We now focus on the case when a weighted graph  $G$  is a part of a nice infinite grid, for instance a part of  $\mathbb{Z}^2$ . For the homogeneous (i.e., all  $J_e = 1$ ) model on  $\mathbb{Z}^2$ , the Kramers–Wannier duality (2.4) suggests that the value  $x = \tan \frac{\pi}{8} = \sqrt{2} - 1$  corresponds to the critical point of the model and indeed a second order phase transition at  $\beta = -\frac{1}{2} \log(\sqrt{2} - 1)$  can be justified in several ways. More generally, one can consider an arbitrary infinite tiling  $\diamond$  of the complex plane by *rhombi* with angles uniformly bounded from below, split the vertices of the bipartite graph  $\Lambda$  formed by the vertices of these rhombi into two *isoradial graphs*  $\Gamma^\bullet$  and  $\Gamma^\circ$ , and define the Ising model weights on  $\Gamma^\bullet$  by setting  $x_e := \tan \frac{1}{2}\theta_e$ , where  $\theta_e$  is the

half-angle of the corresponding rhombus at vertices from  $\Gamma^\bullet$ . This model is called a *self-dual Z-invariant Ising model* on isoradial graphs and it can be viewed as a particular point in a family of the so-called Z-invariant Ising models studied by Baxter and parameterized by an elliptic parameter  $k$ . Recently, it was shown by Boutillier, de Tilière and Raschel that, similarly to the case of regular lattices, the Z-invariant Ising model on a given isoradial graph exhibits a second order phase transition at its self-dual point  $k = 0$ , see [Boutillier, De Tilière, and Raschel \[2016\]](#) and references therein for more details.

Below we are mostly interested in the following setup: let  $\Omega \subset \mathbb{C}$  be a bounded simply connected domain and let  $\Omega^\delta = \Omega(G^\delta)$  be a sequence of its polygonal discretizations on isoradial graphs of mesh size  $\delta$  (in other words, the corresponding rhombic lattice is formed by rhombi with side lengths  $\delta$ ) with  $\delta \rightarrow 0$ . Below we use the notation  $V^\bullet(\Omega^\delta)$ ,  $V^\circ(\Omega^\delta)$ ,  $V^\Delta(\Omega^\delta)$ ,  $V^\diamond(\Omega^\delta)$ ,  $V^\Upsilon(\Omega^\delta)$  and  $V_\varpi^\Upsilon(\Omega^\delta)$  for the sets of vertices of  $(G^\delta)^\bullet$ ,  $(G^\delta)^\circ$ ,  $\Lambda(G^\delta)$ ,  $\diamond(G^\delta)$ ,  $\Upsilon(G^\delta)$  and  $\Upsilon_\varpi(G^\delta)$ , respectively, and we always assume that the Ising model weighs  $x_e = \tan \frac{1}{2}\theta_e$  are chosen according to the geometry of these isoradial graphs, as explained above. In particular, the reader can think of a sequence of discrete domains  $\Omega^\delta$  drawn on square grids of mesh sizes  $\delta \rightarrow 0$  and the homogeneous Ising model with the critical weight  $x = \sqrt{2}-1$ .

**3.2 From Kadanoff–Ceva fermions to discrete holomorphic functions.** As already mentioned in [Section 2.3](#), all the Ising model observables of the form

$$\langle \psi_c \mathcal{O}_\varpi[\mu, \sigma] \rangle := \eta_c \langle \chi_c \mathcal{O}_\varpi[\mu, \sigma] \rangle, \quad \varpi = \{v_1, \dots, v_{m-1}, u_1, \dots, u_{m-1}\},$$

considered as functions of  $c \in V_\varpi^\Upsilon(\Omega^\delta)$ , always satisfy a three-term propagation equation, which is obtained from (2.13) by multiplying each of the terms by the Dirac spinor (2.10). For the critical Ising model on isoradial graphs, it admits a particularly nice interpretation as the *discrete holomorphicity* of  $F_\varpi(c)$ . One has

$$\eta_c = \varsigma \cdot \exp[-\frac{i}{2} \arg(v^\bullet(c) - v^\circ(c))] = \varsigma \delta^{1/2} \cdot (v^\bullet(c) - v^\circ(c))^{-1/2},$$

and a straightforward computation shows that (2.13) can be rewritten as

$$(3.1) \quad \langle \psi_{c_{00}} \mathcal{O}_\varpi[\mu, \sigma] \rangle + \langle \psi_{c_{11}} \mathcal{O}_\varpi[\mu, \sigma] \rangle = \langle \psi_{c_{01}} \mathcal{O}_\varpi[\mu, \sigma] \rangle + \langle \psi_{c_{10}} \mathcal{O}_\varpi[\mu, \sigma] \rangle.$$

Taking the complex conjugate and using  $(v^\bullet(c) - v^\circ(c)) \cdot \eta_c = \varsigma^2 \delta \cdot \bar{\eta}_c$ , one gets

$$(v_0^\circ - v_0^\bullet) \Phi(c_{00}) + (v_1^\bullet - v_0^\circ) \Phi(c_{10}) + (v_1^\circ - v_1^\bullet) \Phi(c_{11}) + (v_0^\bullet - v_1^\circ) \Phi(c_{01}) = 0,$$

where  $\Phi(c) := \langle \psi_c \mathcal{O}_\varpi[\mu, \sigma] \rangle$ , which can be thought of as a vanishing discrete contour integral around a given rhombus. However, there exists an even better way to interpret (3.1), adopted in [Smirnov \[2006\]](#), [Smirnov \[2010a\]](#) and further works. For  $z \in V_\varpi^\diamond(\Omega^\delta)$ , denote

$$(3.2) \quad F_\varpi(z) := \langle (\psi_{c_{00}(z)} + \psi_{c_{11}(z)}) \mathcal{O}_\varpi[\mu, \sigma] \rangle = \langle (\psi_{c_{01}(z)} + \psi_{c_{10}(z)}) \mathcal{O}_\varpi[\mu, \sigma] \rangle.$$

It turns out that there exists a natural discrete Cauchy–Riemann operator  $\partial_\Lambda^*$  (acting on functions defined on  $V^\diamond(\Omega^\delta)$  and returning functions on  $V^\Lambda(\Omega^\delta)$ ) such that

$$(3.3) \quad \partial_\Lambda^* F_\varpi = 0 \quad \text{in } V^\Lambda(\Omega^\delta) \setminus \{v_1, \dots, v_{m-1}, u_1, \dots, u_{n-1}\}.$$

More precisely,  $\partial_\Lambda^*$  is the (formally) adjoint operator to

$$(3.4) \quad [\partial_\Lambda H](z) := \frac{1}{2} \left[ \frac{H(v_1^\bullet(z)) - H(v_0^\bullet(z))}{v_1^\bullet(z) - v_0^\bullet(z)} + \frac{H(v_1^\circ(z)) - H(v_0^\circ(z))}{v_1^\circ(z) - v_0^\circ(z)} \right],$$

see [Mercat \[2001\]](#), [Kenyon \[2002\]](#), [Chelkak and Smirnov \[2011\]](#) and [Chelkak and Smirnov \[2012\]](#) for more details. One of the advantages of (3.3) is that we now have roughly the same number of equations as the number of unknowns  $F(z)$  and do not need to keep track of the fact that the values  $\Phi(c_{pq})$  discussed above have prescribed complex phases (note that the number of unknowns  $\Phi(c)$  is roughly twice the number of vanishing elementary contour integrals around rhombi). However, this information on complex phases is not fully encoded by (3.3). In fact, a slightly stronger condition holds: for two rhombi  $z, z'$  adjacent to the same  $c \in V_\varpi^Y(\Omega^\delta)$ ,

$$(3.5) \quad \Pr[F_\varpi(z); \eta_c \mathbb{R}] = \Pr[F_\varpi(z'); \eta_c \mathbb{R}],$$

where  $\Pr[F; \eta \mathbb{R}] := \bar{\eta}^{-1} \text{Re}[\bar{\eta} F]$ . Actually, one can easily see from (3.1) that both sides of (3.5) are equal to  $\langle \psi_c \mathcal{O}_\varpi[\mu, \sigma] \rangle$  and it is not hard to check that the condition (3.5) indeed implies (3.3), see [Chelkak and Smirnov \[ibid., Section 3.2\]](#).

*Remark 3.1.* In [Chelkak and Smirnov \[ibid.\]](#), the term *s-holomorphicity* was introduced for complex-valued functions  $F(z)$  defined on  $V^\diamond(\Omega^\delta)$  and satisfying (3.5), in particular to indicate that such functions also satisfy (3.3). In view of (3.2), this is essentially the same notion as the one introduced in [Definition 2.1](#) for real-valued spinors defined on  $V^\times(\Omega^\delta)$ . Still, it is worth mentioning that (2.13) does not rely upon the very specific (isoradial) choice of the embedding of  $G^\delta$  and the Ising weights, while (3.3) does; see also [Chelkak, Cimasoni, and Kassel \[2017, Sections 3.5, 3.6\]](#) for a discussion of (3.5) in the general case.

*Remark 3.2.* In breakthrough works [Smirnov \[2006\]](#), [Smirnov \[2010a\]](#), [Duminil-Copin and Smirnov \[2012\]](#), [Smirnov \[2010b\]](#) of Smirnov on the critical Ising model (see also [Chelkak and Smirnov \[2012\]](#) and [Hongler and Smirnov \[2013\]](#)), discrete holomorphic fermions were introduced in a purely combinatorial way, as a particular case of parafermionic observables, and the s-holomorphicity condition (3.5) was verified combinatorially as well (e.g., see [Smirnov \[2010b, Fig. 5\]](#) or [Chelkak and Smirnov \[2012, Fig. 5\]](#)). Following this route, more general spinor observables (3.2) were also treated combinatorially, and not via the Kadanoff–Ceva formalism, in [Chelkak and Izyurov \[2013\]](#) and [Chelkak, Hongler, and Izyurov \[2015\]](#).

**3.3 Boundary conditions and the key role of the function  $H_F$ .** We now come back to [Definition 2.2](#) suggested in [Smirnov \[2006\]](#) and [Smirnov \[2010a\]](#) as a crucial tool for the analysis of fermionic observables via boundary value problems for s-holomorphic functions in  $\Omega^\delta$ . Let  $F = F^\delta$  be such a function and  $H_F$  be defined via [\(2.14\)](#) from the corresponding (see [Remark 3.1](#)) real-valued s-holomorphic spinor on  $V^\times(\Omega^\delta)$ . A straightforward computation shows that, for each  $z \in V^\circ(\Omega^\delta)$ , one has

$$(3.6) \quad \begin{aligned} H_F(v_1^\bullet) - H_F(v_0^\bullet) &= \operatorname{Re}[\bar{\varsigma}^2 \delta^{-1}(F(z))^2(v_1^\bullet - v_0^\bullet)] = \operatorname{Im}[\delta^{-1}(F(z))^2(v_1^\bullet - v_0^\bullet)], \\ H_F(v_1^\circ) - H_F(v_0^\circ) &= \operatorname{Re}[\bar{\varsigma}^2 \delta^{-1}(F(z))^2(v_1^\circ - v_0^\circ)] = \operatorname{Im}[\delta^{-1}(F(z))^2(v_1^\circ - v_0^\circ)], \end{aligned}$$

provided that the global prefactor in the definition [\(2.10\)](#) is chosen<sup>1</sup> as  $\varsigma = e^{i\frac{\pi}{4}}$ . In other words, the real-valued function  $H_F$  turns out to be a proper discrete analogue of the expression  $\int \operatorname{Im}[(F(z))^2 dz]$ . Even more importantly,  $H_F$  encodes the boundary conditions of the Ising model in the form of Dirichlet boundary conditions. Namely, one can choose an additive constant in its definition so that

$$(3.7) \quad \begin{aligned} &\bullet H_F = 0 \text{ and } \partial_{v_{\text{in}}} H \geq 0 \text{ on all the wired boundary arcs,} \\ &\bullet H_F \text{ is constant and } \partial_{v_{\text{in}}} H \leq 0 \text{ on each of the free boundary arcs,} \end{aligned}$$

where  $\partial_{v_{\text{in}}}$  stands for a discrete analogue of the inner normal derivative, see [Chelkak and Smirnov \[2012\]](#) and [Izyurov \[2015\]](#). In particular, the values of the s-holomorphic spinor  $F$  at two endpoints of a free arc have the same absolute value (which, according to [\(2.14\)](#), is equal to the square root of the value of  $H$  on this free arc) and in fact one can also match their signs by tracking the Dirac spinor [\(2.10\)](#) along the boundary; see [Izyurov \[ibid.\]](#) and [Chelkak, Hongler, and Izyurov \[n.d.\]](#) for more details.

One can also check that, for  $\varpi = \{v_1, \dots, v_{m-1}, u_1, \dots, u_{n-1}\}$  and  $H_{F_\varpi}$  obtained from the s-holomorphic function  $F_\varpi$  given by [\(3.2\)](#) (or, equivalently, from the s-holomorphic spinor  $\langle \chi_c \mathcal{O}_\varpi[\mu, \sigma] \rangle$ ), the following is fulfilled:

- the minimum of  $H_F$  in a vicinity of each of  $v_p$  is attained at the boundary,
- the maximum of  $H_F$  in a vicinity of each of  $u_q$  is attained at the boundary.

Imagine now that instead of a sequence of discrete s-holomorphic functions  $F^\delta$  on  $\Omega^\delta$  we had a continuous holomorphic function  $f : \Omega \rightarrow \mathbb{C}$  such that the harmonic function  $h_f := \int \operatorname{Im}[(f(z))^2 dz]$  satisfied the conditions listed above. We could then hope to identify  $f$  as a solution to such a boundary value problem in  $\Omega$ , provided that this solution is unique (up to a normalization to be fixed). This not necessarily true in full generality (e.g., some degeneracy might appear in presence of several spins and several

<sup>1</sup>This is a matter of convenience. E.g., the notation in [Hongler and Smirnov \[2013\]](#), [Hongler and Kytölä \[2013\]](#), [Chelkak, Hongler, and Izyurov \[2015\]](#) and [Gheissari, Hongler, and Park \[2013\]](#) corresponds to  $\varsigma = i$ .

disorders in  $\mathcal{O}_{\mathcal{W}}[\mu, \sigma]$ ) but there are enough situations in which a relevant uniqueness theorem ‘in continuum’ is fulfilled and thus one can hope to prove the convergence of  $F^\delta$  to  $f$  as  $\delta \rightarrow 0$ ; see [Section 4.1](#) and [Chelkak, Hongler, and Izyurov \[n.d.\]](#) for more details.

**3.4 Smirnov’s sub-/super-harmonicity.** The interpretation of  $s$ -holomorphic functions  $F^\delta$  as solutions to discrete boundary value problems described above is implicitly based on the idea that one can think of  $H_{F^\delta}$  as of a discrete harmonic function on  $\Lambda(\Omega^\delta)$ . However, this cannot be true literally: the functions  $(F^\delta)^2$  are not discrete holomorphic, thus there is no hope that  $H_{F^\delta}$  are exactly discrete harmonic. Nevertheless, there is a miraculous positivity phenomenon, first observed in [Smirnov \[2006\]](#) and [Smirnov \[2010a\]](#) on the square grid and later generalized to the isoradial setup in [Chelkak and Smirnov \[2012\]](#).

**Lemma 3.1** (see [Smirnov \[2010a\]](#), Lemma 3.8) and [Chelkak and Smirnov \[2012\]](#), Proposition 3.6(iii)). *Let  $F$  be defined on quads surrounding a given vertex  $v \in V^\Lambda(\Omega^\delta)$  and satisfy the  $s$ -holomorphicity identities (3.5). If  $H_F$  is constructed via (3.6) (or, equivalently, via (2.14)), then*

$$(3.8) \quad [\Delta^\bullet H_F](v) \geq 0 \text{ if } v \in V^\bullet(\Omega^\delta), \quad \text{and} \quad [\Delta^\circ H_F](v) \leq 0 \text{ if } v \in V^\circ(\Omega^\delta),$$

where the Laplacian operator  $\Delta^\bullet$  (and, similarly,  $\Delta^\circ$ ) is defined as

$$(3.9) \quad [\Delta^\bullet H](v) := \sum_{v_1: v_1 \sim v} \tan \theta_{(vv_1)} \cdot (H(v_1) - H(v)).$$

It is easy to check that both discrete operators  $\Delta^\bullet$  and  $\Delta^\circ$  approximate the standard Laplacian as  $\delta \rightarrow 0$  in a quite strong (local) sense; see [Chelkak and Smirnov \[2011\]](#) for details. Recall that, according to [Definition 2.2](#) the discrete *subharmonic* function  $H_{F^\delta}|_{V^\bullet(\Omega^\delta)}$  is pointwise (i.e., at any pair of neighboring vertices  $v^\bullet(c)$  and  $v^\circ(c)$ ) greater or equal to the discrete *superharmonic* function  $H_{F^\delta}|_{V^\circ(\Omega^\delta)}$ . Therefore, provided that the difference  $\delta H_{F^\delta}(v^\bullet(c)) - \delta H_{F^\delta}(v^\circ(c)) = (F^\delta(c))^2$  is small inside of  $\Omega^\delta$ , both  $\delta H_{F^\delta}|_{V^\bullet(\Omega^\delta)}$  and  $\delta H_{F^\delta}|_{V^\circ(\Omega^\delta)}$  should have the same *harmonic* limit as  $\delta \rightarrow 0$ .

*Remark 3.3.* Still, there is a question of how to show that  $F^\delta$  is small. In the pioneering work [Smirnov \[2006\]](#) devoted to basic fermionic observables on  $\mathbb{Z}^2$ , this fact was derived from monotonicity arguments and the magnetization estimate; see [Smirnov \[2010a\]](#), Lemma A.1]. Shortly afterwards, an *a priori* regularity theory for functions  $H_F$  were developed in [Chelkak and Smirnov \[2012\]](#), Section 3]. These *a priori* estimates were later applied to various Ising-model observables, in particular when no simple monotonicity arguments are available.

*Remark 3.4.* It is commonly believed that one cannot *directly* generalize [Lemma 3.1](#) for more general graphs or Ising-model weights  $x_e = \frac{1}{2} \tan \theta_e$ : the existence of *some*

Laplacians  $\Delta^\bullet$  and  $\Delta^\circ$  such that (3.8) holds true seems to be equivalent to the conditions  $\sum_{v_1: v_1 \sim v} \arctan x_{(vv_1)} \in \frac{\pi}{2} + \pi\mathbb{Z}$ , which lead to an isoradial embedding of  $(G^\bullet, G^\circ)$ , possibly with  $(2\pi + 4\pi\mathbb{Z})$  conical singularities at vertices.

## 4 Convergence of correlations

In this section we very briefly discuss the convergence results for correlation functions (fermions, energy densities, spins etc) obtained during the last decade for the critical model on  $\mathbb{Z}^2$  (see also Remark 4.2 on the critical Z-invariant model). We refer the interested reader to upcoming Chelkak, Hongler, and Izyurov [n.d.], see also Chelkak [2016a, Section 4].

**4.1 Convergence of s-holomorphic observables.** As already mentioned above, techniques developed in Chelkak and Smirnov [2012] essentially allow one to think of sub/super-harmonic functions  $H^\delta := \delta^{-1} H_{F^\delta}$  as of harmonic ones. In particular, a uniform boundedness of the family  $H^\delta$  on an open set implies that both  $H^\delta$  and the original observables  $F^\delta$  are equicontinuous on compact subsets of this open set.

Imagine now that we want to prove the convergence of (properly normalized) observables  $F^\delta_\varpi$ ,  $\varpi = \{v_1, \dots, v_{m-1}, u_1, \dots, u_{n-1}\}$  as  $\delta \rightarrow 0$ . Assuming that the corresponding functions  $H^\delta_\varpi$  are uniformly bounded away from  $v_p$ 's and  $u_q$ 's, the Arzelà–Ascoli theorem ensures that, *inside of*  $\Omega \setminus \varpi$ , at least subsequential limits  $H^\delta_\varpi \rightarrow h_\varpi$  and  $F^\delta_\varpi \rightarrow f_\varpi$  exist. Since the functions  $F^\delta_\varpi$  are discrete holomorphic, their limit  $f_\varpi$  is a holomorphic function and one also has  $h_\varpi = \int \text{Im}[(f_\varpi(z))^2 dz]$ . Moreover, one can show that the *boundary conditions* (3.7) survive as  $\delta \rightarrow 0$ , see Chelkak and Smirnov [ibid., Remark 6.3] and Izyurov [2015]. Clearly,  $h_\varpi$  also inherits from  $H^\delta_\varpi$  the semi-boundedness from below near  $v_p$  and from above near  $u_q$ . Thus, only two questions remain:

- (i) to show that  $f$  and  $h$  are uniquely characterized by the above properties;
- (ii) to justify that the functions  $\delta^{-1} H_\delta$  are uniformly bounded away from  $\varpi$ .

In general (i.e., in presence of several disorders and spins in  $\mathcal{O}_\varpi[\mu, \sigma]$ ), the uniqueness (i) may fail. Fortunately, there exists a principal setup in which it holds true:

$$(4.1) \quad F^\delta_\varpi(z) := \frac{\delta^{-1} \langle \psi(z) \chi_d \sigma_{u_1} \dots \sigma_{u_{n-2}} \rangle}{\langle \sigma_{u_1} \dots \sigma_{u_{n-2}} \rangle}, \quad \psi(z) := \psi_{c_{00}(z)} + \psi_{c_{11}(z)},$$

where  $d$  and  $u_1, \dots, u_{n-2}$  are assumed to approximate distinct points of  $\Omega$  as  $\delta \rightarrow 0$ . (Recall also that the roles of spins and disorders are not fully symmetric since our standard boundary conditions are not fully symmetric under the Kramers–Wannier duality). Note that such functions  $F^\delta_\varpi$  have ‘standard’ discrete singularities at  $d$ , leading to a ‘standard’ singularity (simple pole with a fixed residue) of  $f_\varpi$  at  $d$ .

Finally, a useful trick allowing to deduce (ii) from (i) was suggested in [Chelkak, Hongler, and Izyurov \[2015, Section 3.4\]](#): if the functions  $H_{\overline{w}}^{\delta}$  were unbounded, then one could renormalize them in order to obtain non-trivial limits  $\widetilde{f}_{\overline{w}}$  and  $\widetilde{h}_{\overline{w}} = \int \text{Im}[(\widetilde{f}_{\overline{w}}(z))^2 dz]$ , which would solve the same boundary value problem as  $f_{\overline{w}}$  and  $h_{\overline{w}}$  but without the pole at  $d$ , killed by such an additional renormalization. Due to (i), this boundary value problem has no nontrivial solution, which gives a contradiction; see also [Chelkak, Hongler, and Izyurov \[n.d.\]](#).

**4.2 Fusion: from s-holomorphic observables to  $\varepsilon, \sigma$  and  $\mu$ .** Once the convergence of observables (4.1) is established, one can first use the Pfaffian structure (2.17) of fermionic correlators to obtain the convergence of the quantities

$$(4.2) \quad \frac{\langle \chi_{c_1} \dots \chi_{c_k} \sigma_{u_1} \dots \sigma_{u_{n-k}} \rangle}{\langle \sigma_{u_1} \dots \sigma_{u_{n-k}} \rangle} = \text{Pf} \left[ \frac{\langle \chi_{c_r} \chi_{c_s} \sigma_{u_1} \dots \sigma_{u_{n-k}} \rangle}{\langle \sigma_{u_1} \dots \sigma_{u_{n-k}} \rangle} \right]_{s,r=1}^k.$$

In particular, the simplest case  $n - k = 0$  allows one to study (see [Hongler and Smirnov \[2013\]](#) and [Hongler \[2010\]](#)) the scaling limit of correlations of the *energy density* field, defined for  $z \in V^{\diamond}(\Omega^{\delta})$  as

$$\varepsilon_z := \sigma_{v_0^{\diamond}(z)} \sigma_{v_1^{\diamond}(z)} - 2^{-\frac{1}{2}} = \pm 2^{-\frac{1}{2}} \chi_{c_{00}(z)} \chi_{c_{11}(z)} = 2^{-\frac{1}{2}} - \mu_{v_0^{\bullet}(z)} \mu_{v_1^{\bullet}(z)}.$$

Note that a careful analysis of the singularity of (4.1) as  $z \rightarrow d$  is required in order to handle the next-to-leading term  $\varepsilon_z$  appearing in the fusion of two fermions  $\chi_c \chi_d$ .

In the same spirit, recursively analyzing the behavior of (4.2) near discrete singularities as  $c_{k-s} \rightarrow u_{n-k-s}$ ,  $s = 0, \dots, r-1$ , one obtains scaling limits of the ratios

$$(4.3) \quad \langle \chi_{c_1} \dots \chi_{c_{k-r}} \mu_{u_{n-k-r+1}^{\bullet}} \dots \mu_{u_{n-k}^{\bullet}} \sigma_{u_1} \dots \sigma_{u_{n-k-r}} \rangle \cdot \langle \sigma_{u_1} \dots \sigma_{u_{n-k}} \rangle^{-1},$$

where  $u_p^{\bullet} \in V^{\bullet}(\Omega^{\delta})$  denotes one of the neighboring to  $u_p$  vertices of  $\Gamma^{\bullet}$ . The remaining ingredient is the convergence of the denominators  $\langle \sigma_{u_1} \dots \sigma_{u_{n-k}} \rangle$ , which do not contain any discrete holomorphic variable. Such correlations were treated in [Chelkak, Hongler, and Izyurov \[2015\]](#) using the following idea: fusing  $\chi_d$  and  $\sigma_{u_1}$  in (4.1) one gets the observable

$$\langle \psi(z) \mu_{u_1^{\bullet}} \sigma_{u_2} \dots \sigma_{u_{n-2}} \rangle \cdot \langle \sigma_{u_1} \dots \sigma_{u_{n-2}} \rangle^{-1}$$

and then analyzes its behavior (as that of a discrete holomorphic function in  $z$ ) near the vertex  $u_1^{\bullet} \sim u_1$ . This analysis provides an access to the ratio

$$(4.4) \quad \langle \sigma_{u_1^{\diamond}} \sigma_{u_2} \dots \sigma_{u_{n-2}} \rangle \cdot \langle \sigma_{u_1} \dots \sigma_{u_{n-2}} \rangle^{-1},$$

where  $u_1^{\diamond} \sim u_1^{\bullet} \sim u_1$  can be any neighboring to  $u_1$  face of  $\Omega^{\delta}$ . The next-to-leading term in the asymptotics of this ratio as  $\delta \rightarrow 0$  encodes the discrete spatial derivative

of  $\log\langle\sigma_{u_1}\dots\sigma_{u_{n-2}}\rangle$ , which eventually allows one to identify its scaling limit by fusing spins to each other and thus reducing  $n$ ; see [Chelkak, Hongler, and Izyurov \[ibid.\]](#) and [Chelkak, Hongler, and Izyurov \[n.d.\]](#) for more details.

*Remark 4.1.* The above scheme allows one to prove the convergence of arbitrary correlations of disorders, spins and fermions under *standard* boundary conditions in  $\Omega^\delta$ . In fact, it can be further generalized to include (a) the (common) spin  $\sigma_{u_{\text{out}}}$  of the wired boundary arcs; (b) one or several disorders assigned to free boundary arcs; (c) fermions on the boundary of  $\Omega^\delta$ . In particular, this allows one to handle an arbitrary mixture of ‘+’, ‘−’ and ‘free’ boundary conditions; see [Chelkak, Hongler, and Izyurov \[ibid.\]](#) for details.

*Remark 4.2.* The convergence results for s-holomorphic observables (4.1) and the energy density correlations can be also proved in the isoradial setup ad verbum. Nevertheless, the passage to (4.3) and, especially, the analysis of the spatial derivatives (4.4) of spin correlations require some additional work. We believe that all the key ingredients can be extracted from [Dubédat \[2015\]](#) but it is worth mentioning that such a generalization has not appeared yet even for the honeycomb/triangular grids.

**4.3 More CFT on the lattice.** From the Conformal Field Theory perspective (e.g., see [Mussardo \[2010\]](#)), the scaling limits of correlations of fermions, spins, disorders and energy-densities are those of *primary fields* (non-local ones in the case of  $\psi$  and  $\mu$ ). It is a subject of the ongoing research to construct the corresponding CFT, as fully as possible, directly on the lattice level (and not in the limit as  $\delta \rightarrow 0$ ). Below we mention several results and research projects in this direction:

- The analysis of general *lattice fields* (i.e., functions of several neighboring spins) was started in [Gheissari, Hongler, and Park \[2013\]](#), though at the moment only their leading terms, which converge to either  $1, \varepsilon, \sigma$  or the spatial derivative of  $\sigma$  in the scaling limit, are treated.

- The action of the *Virasoro algebra* on such lattice fields was recently defined in [Hongler, Kytölä, and Viklund \[2017\]](#), via the so-called Sugawara construction applied to discrete fermions.

- An ‘infinitesimal non-planar deformation’ approach to the *stress-energy tensor* of the Ising model on faces of the honeycomb grid was recently suggested in [Chelkak, Glazman, and Smirnov \[2016\]](#).

We also believe that one can use s-embeddings (see [Section 6](#)) of weighted planar graphs  $(G, x)$  to properly interpret an infinitesimal change of the Ising weights  $x_e$  as a *vector-field* in  $\mathbb{C}$ , thus providing yet another approach to the stress-energy tensor.



## 5 Interfaces and loop ensembles

In the 20th century, the (conjectural) conformal invariance of critical lattice models was typically understood via the scaling limits of correlation functions, which are fundamental objects studied by the classical CFT (e.g., see [Mussardo \[2010\]](#)), though see [Aizenman and Burchard \[1999\]](#) and [Langlands, Lewis, and Saint-Aubin \[2000\]](#). The introduction of SLEs by Schramm (see [Schramm \[2000\]](#)), and later developments on CLEs (Conformal Loop Ensembles, see [Sheffield \[2009\]](#), [Sheffield and Werner \[2012\]](#), [Miller, Sheffield, and Werner \[2017\]](#) and references therein) provided quite a different perspective of studying the *geometry* of either particular interfaces (e.g., domain walls in the critical Ising model) – conjecturally converging to SLEs – or the full collection of interfaces – conjecturally converging to CLEs – as  $\Omega^\delta \rightarrow \Omega$ .

For the critical Ising model, both in its classical and the so-called random-cluster (or Fortuin–Kastelen, see below) representations, the convergence of interfaces generated by Dobrushin boundary conditions was obtained already in the pioneering work of Smirnov (see [Smirnov \[2006, Theorem 1\]](#) and references therein) via the application of the so-called *martingale principle* (e.g., see [Smirnov \[ibid., Section 5.2\]](#)) and the convergence results for basic fermionic observables (see [Smirnov \[2010a\]](#) and [Chelkak and Smirnov \[2012\]](#)). However, the improvement of the *topology of convergence* from that of functional parameters (aka driving forces) in the Loewner equation to the convergence of curves themselves required some additional efforts: the appropriate framework was provided in [Kemppainen and Smirnov \[2017\]](#); see also [Chelkak, Duminil-Copin, Hongler, Kemppainen, and Smirnov \[2014\]](#) and references therein. Basing on this framework, the convergence of the full *branching tree* of interfaces in the FK-representation (announced in [Smirnov \[2006\]](#) and [Smirnov \[2010a\]](#)) to CLE(16/3) was eventually justified by Kemppainen and Smirnov in [Kemppainen and Smirnov \[2015\]](#) and [Kemppainen and Smirnov \[n.d.\]](#). In parallel, an exploration algorithm aiming at the proof of the convergence of the classical domain walls ensemble to CLE(3) was suggested in [Hongler and Kytölä \[2013\]](#) and later, via the convergence of the so-called *free arc ensemble* established in [Benoist, Duminil-Copin, and Hongler \[2016\]](#), this convergence to CLE(3) has also been justified by Benoist and Hongler in [Benoist and Hongler \[2016\]](#). Recently, another approach to derive the convergence of the domain walls loop ensemble to CLE(3) from that of the random cluster one to CLE(16/3) was suggested in [Miller, Sheffield, and Werner \[2017\]](#).

Certainly, it is absolutely impossible to provide details of these advanced developments in a short note, thus we refer the interested reader to the original articles mentioned above and hope that such a survey will appear one day. Below we only emphasize several ingredients coming from the complex analysis side and indicate the role played by the Ising model observables discussed above.

**5.1 FK-Ising (random cluster) representation and crossing probabilities.** Recall that the random-cluster representation of the critical Ising model on  $\mathbb{Z}^2$  (e.g., see [Smirnov \[2006, Section 2.3\]](#)) is a probability measure on the configurations of edges of  $\Gamma^\circ$  (each edge is declared open or closed), proportional to

$$(5.1) \quad x_{\text{crit}}^{\#\text{closed edges}} (1 - x_{\text{crit}})^{\#\text{open edges}} \cdot 2^{\#\text{clusters}} \sim \sqrt{2}^{\#\text{loops}},$$

where  $\#\text{clusters}$  stands for the number of connected components in a given configuration and  $\#\text{loops}$  denotes the number of loops separating these clusters of vertices of  $\Gamma^\circ$  from dual ones, living on  $\Gamma^\bullet$  (and formed by the edges of  $\Gamma^\bullet$  dual to the closed ones of  $\Gamma^\circ$ ); e.g., see [Smirnov \[ibid., Fig. 3\]](#). Through the *Edward–Sokal coupling*, it is intimately related to the original spin model: to obtain a random cluster configuration from  $\sigma$ , one tosses a (biased according to the Ising weight  $x_e$ ) coin for each edge  $e$  of  $\Gamma^\circ$  connecting aligned spins; and, inversely, tosses a fair coin to assign a  $\pm 1$  spin to each of the clusters of a given random cluster configuration. In particular,

$$(5.2) \quad \mathbb{E}^\circ[\sigma_{u_1} \dots \sigma_{u_n}] = \mathbb{P}^{\text{FK}}[\text{each cluster contains an even number of } u_1, \dots, u_n].$$

*Remark 5.1.* The fact that the probability measure (5.1) is proportional to  $\sqrt{2}^{\#\text{loops}}$  relies upon the duality of the graphs  $\Gamma^\bullet$  and  $\Gamma^\circ$ , as embedded into a *sphere*. Recall that we consider all the vertices of  $\Gamma^\circ$  on wired arcs as a *single* ‘macro-vertex’ while in  $\Gamma^\bullet$  there is one ‘macro-vertex’ for *each* of the free arcs, so this duality holds.

A particularly important application of the identity (5.2) (applied to disorders on  $\Gamma^\bullet$  rather than to spins on  $\Gamma^\circ$ ) appears in the *quadrilateral* setup, when there are two wired  $(ab)$ ,  $(cd)$  and two free  $(bc)$ ,  $(da)$  arcs on  $\partial\Omega^\delta$ . Namely, one has

$$(5.3) \quad \langle \mu_{(da)} \mu_{(bc)} \rangle = \mathbb{P}^{\text{FK}}[(da) \leftrightarrow (bc)] = \varrho \left( \mathbb{P}^{\text{loops}}[(da) \leftrightarrow (bc)] \right),$$

where  $\varrho(p) := p \cdot (p + \sqrt{2}(1-p))^{-1}$  and  $\mathbb{P}^{\text{loops}} \sim \sqrt{2}^{\#\text{loops}}$  denotes the probability measure on FK-Ising configurations, where only the loops lying inside of  $\Omega^\delta$  are counted. In other words, in this *self-dual* setup the two wired arcs are not connected, similarly to the free ones; cf. [Remark 5.1](#). Note that, though  $\mathbb{P}^{\text{loops}}$  does not literally correspond to any FK-Ising measure, it is absolutely continuous with respect to  $\mathbb{P}^{\text{FK}}$ , with the density proportional to  $1 + (\sqrt{2} - 1) \cdot \mathbf{1}[(da) \leftrightarrow (bc)]$ .

**5.2 Crossing estimates and tightness.** When studying the convergence of random curves  $\gamma^\delta$  as  $\delta \rightarrow 0$  (e.g.,  $\gamma^\delta$  can be an interface generated by Dobrushin boundary conditions, see [Chelkak, Duminil-Copin, Hongler, Kemppainen, and Smirnov \[2014\]](#), or a branch of the FK-Ising tree, see [Kemppainen and Smirnov \[2015\]](#) and [Kemppainen and](#)

Smirnov [n.d.], or a branch of the free arc ensemble, see Benoist, Duminil-Copin, and Hongler [2016] and Benoist and Hongler [2016]), an important step is to establish the tightness of this family in the topology induced by the natural metric on the space of curves considered up to reparametrizations. Departing from the classical results of Aizenman and Burchard [1999], which appeared even before the introduction of SLE, Kemppainen and Smirnov showed in Kemppainen and Smirnov [2017] that a very weak estimate on the probability of an *annulus crossing* implies not only the tightness of  $\gamma^\delta$  themselves but also the tightness of the corresponding random driving forces in the Loewner equations and a uniform bound on exponential moments of these driving forces. Below we only emphasize the following two points of the study of crossing estimates; see the original article for more details.

- The annulus crossing estimate mentioned above (*geometric* Condition G in Kemppainen and Smirnov [ibid.]) was shown to be equivalent to a similar estimate on crossings of general topological quadrilaterals (*conformal* Condition C in Kemppainen and Smirnov [ibid.]), uniform in the *extremal length* of a quadrilateral. The latter is conformally invariant by definition, which makes the framework developed in Kemppainen and Smirnov [ibid.] extremely well-suited to the SLE theory.
- Provided suitable monotonicity (with respect to the position of the boundary and the boundary conditions) arguments are available, it is enough to obtain the required crossing estimates for *two standard quadrilaterals* only, with alternating (i.e., wired/free/wired/free or ‘+ / − / + / −’, respectively) boundary conditions.

For the FK-Ising model, it is then enough to prove a uniform (in  $\delta$ ) lower bound for the quantities (5.3). In fact, using techniques described in Section 4, one can even find the limit of these correlations as  $\delta \rightarrow 0$  (see also Chelkak and Smirnov [2012, Theorem 6.1] for a shortcut suggested earlier by Smirnov, which reduces (5.3) to the analysis of basic s-holomorphic observables). A similar crossing estimate for the spin-Ising model can be then easily deduced via the Edwards–Sokal coupling and another application of the FKG inequality, see Chelkak, Duminil-Copin, Hongler, Kemppainen, and Smirnov [2014, Remark 4] and Chelkak, Duminil-Copin, and Hongler [2016, Section 5.3] (a more involved but self-contained argument can be found in Kemppainen and Smirnov [2017, Section 4.2]).

*Remark 5.2.* In absence of the required monotonicity arguments, one can always use the ‘strong’ RSW-type theory developed for the critical FK-Ising model in Chelkak, Duminil-Copin, and Hongler [2016] (basing, in particular, on techniques from Chelkak [2016b]) to verify Condition C of Kemppainen and Smirnov [2017].

**5.3 Martingale observables and convergence of interfaces.** According to the classical martingale principle, aiming to describe the scaling limit of a single interface  $\gamma^\delta$

running in  $\Omega^\delta$  from a marked boundary point  $\gamma^\delta(0) = a^\delta$ , one can (try to) find an observable  $M^\delta(z) = M_{\Omega^\delta, a^\delta}(z)$  such that, for each  $z \in \Omega^\delta$ , the value  $M_{\Omega^\delta \setminus \gamma^\delta[0; t], \gamma^\delta(t)}(z)$  is a martingale with respect to the filtration generated by  $\gamma^\delta[0; t]$ . Provided that the scaling limit  $M(z)$  of  $M^\delta(z)$  as  $\delta \rightarrow 0$  can be identified and the tightness conditions discussed in the previous section are fulfilled, one thus gets a *family* (indexed by  $z \in \Omega$ ) of martingales  $M_{\Omega \setminus \gamma[0; t], \gamma(t)}(z)$  for a (subsequential) limit  $\gamma$  of  $\gamma^\delta$ , which is enough to identify its law as that of  $\text{SLE}(\kappa)$ . Note, however, that some additional analysis is usually required when working with more general  $\text{SLE}(\kappa, \rho)$  curves in order to control their behavior on the set of times when the corresponding Bessel process hits 0, see [Kemppainen and Smirnov \[2015\]](#), [Kemppainen and Smirnov \[n.d.\]](#) and [Benoist, Duminil-Copin, and Hongler \[2016\]](#), [Benoist and Hongler \[2016\]](#) for details.

The convergence results on correlation functions described above provide such martingale observables  $M^\delta(z)$ , amenable to the analysis in the limit  $\delta \rightarrow 0$ , for a huge variety of setups, in both spin- and FK-representations of the model. E.g.,

- $\langle \psi(z) \chi_a \rangle / \langle \mu_b \mu_a \rangle$  is a martingale for the *domain wall* (converging to  $\text{SLE}(3)$  as  $\delta \rightarrow 0$ , see [Chelkak, Duminil-Copin, Hongler, Kemppainen, and Smirnov \[2014\]](#)) generated by Dobrushin (i.e., ‘−’ on  $(ab)$ , ‘+’ on  $(ba)$ ) boundary conditions;
- more generally,  $\langle \psi(z) \chi_a \rangle / \langle \mu_{(bc)} \mu_a \rangle$  is a martingale for the domain wall emanating from  $a$  under ‘+ / − / free’ boundary conditions (this interface converges, see [Hongler and Kytölä \[2013\]](#) or [Izyurov \[2015\]](#), to the dipolar  $\text{SLE}(3, -3/2, -3/2)$ , an important building block of [Benoist, Duminil-Copin, and Hongler \[2016\]](#), [Benoist and Hongler \[2016\]](#));
- $\langle \psi(z) \mu_{(ba)} \sigma_{(ab)} \rangle$  is a martingale for the *FK-interface* generated by Dobrushin (wired on  $(ab)$ , free on  $(ba)$ ) boundary conditions (this is the observable introduced by Smirnov in [Smirnov \[2006\]](#), [Smirnov \[2010a\]](#) to prove the convergence of this interface to  $\text{SLE}(16/3)$ ).

*Remark 5.3.* Following [Smirnov \[2006\]](#), [Smirnov \[2010a\]](#), [Chelkak and Smirnov \[2012\]](#), [Duminil-Copin and Smirnov \[2012\]](#), these martingale observables are usually defined in a purely combinatorial way in the existing literature (and not via the Kadanoff–Ceva formalism, cf. [Remark 3.2](#)). One of the advantages of this approach is that the martingale property becomes a triviality. On the other hand, the origin of the crucial s-holomorphicity property becomes less transparent.

**5.4 Convergence of loop ensembles.** As already mentioned above, we do not intend to overview the details of [Kemppainen and Smirnov \[2015\]](#), [Kemppainen and Smirnov \[n.d.\]](#) (convergence of the FK-Ising loop ensemble to  $\text{CLE}(16/3)$ ) and of [Benoist, Duminil-Copin, and Hongler \[2016\]](#), [Benoist and Hongler \[2016\]](#) (convergence of the domain walls ensemble to  $\text{CLE}(3)$ ) in this essay and refer the interested reader to the original articles. In both projects, some iterative procedure is used: a branching exploration of loops in the former (which is a prototype of the branching  $\text{SLE}$ -tree from [Sheffield \[2009\]](#)) and

an alternate exploration of FK-Ising clusters and free arc ensembles (proposed in [Hongler and Kytölä \[2013\]](#)) in the latter. Because of the hierarchical nature of these algorithms, the following subtlety arises: even if the convergence of each of FK-interfaces to SLE(16/3) curves is known, one still needs to control the behavior of its double points, which split the current domain into smaller pieces to be explored (and also of the endpoints of arcs constituting the free arc ensemble in [Benoist, Duminil-Copin, and Hongler \[2016\]](#)), in order to guarantee that the exploration algorithm in discrete does not deviate too much from the continuum one. This is done<sup>2</sup> using the strong RSW theory developed in [Chelkak, Duminil-Copin, and Hongler \[2016\]](#), which guarantees that all such ‘pivotal’ points in continuum are the limits of those in discrete, uniformly in the shape of subdomains obtained along the way and boundary conditions, cf. [Remark 5.2](#).

## 6 Towards universality beyond isoradial graphs

As discussed above, the fundamental questions of convergence and conformal invariance of the critical Ising model on the square grid are now relatively well understood, both for correlation functions and loop ensembles. Moreover, a great part, if not all, of these results can be generalized from  $\mathbb{Z}^2$  to the critical *Z-invariant* model using the already existing techniques. Nevertheless, this does not give a fully satisfactory understanding of the *universality* phenomenon since the cornerstone ‘sub-/super-harmonicity’ [Lemma 3.1](#) does not admit a direct generalization beyond the isoradial case, see [Remark 3.4](#). E.g., even the convergence of Ising interfaces on doubly-periodic weighted graphs has never been treated though the criticality condition on the weights in this case is well known; see [Cimasoni and Duminil-Copin \[2013\]](#) and references therein.

The main purpose of this section is to discuss a new class of embeddings of *generic* planar weighted graphs carrying the Ising model into the complex plane, with the emphasis on an analogue of the ‘s-Lemma’ [3.1](#). Below we only sketch some important features of the construction, details will appear elsewhere. Independently of our paper, a special class of s-embeddings – circle patterns – was studied in [Lis \[2017\]](#) and the *criticality* was proven in the case of uniformly bounded faces.

**6.1 S-embeddings of weighted planar graphs.** Below we adopt the notation of [Section 2](#). To construct a particular embedding into  $\mathbb{C}$  of a given planar weighted graph  $(G, x)$  we choose a pair  $\mathfrak{F}_1, \mathfrak{F}_2$  of *s-holomorphic spinors* on  $\Upsilon^\times(G)$  and denote  $\mathfrak{F} := \mathfrak{F}_1 + i\mathfrak{F}_2$ . For instance, one can imagine one of the following setups:

<sup>2</sup>The results of [Kempainen and Smirnov \[2015\]](#) and [Kempainen and Smirnov \[n.d.\]](#) can be made self-contained though the current version relies upon [Chelkak, Duminil-Copin, and Hongler \[2016\]](#) (while checking the required crossing estimates for the full tree of interfaces in the proof of [Kempainen and Smirnov \[2015, Theorem 3.4\]](#)) in order to lighten the presentation (S. Smirnov, private communication).

- $G$  is an infinite graph,  $\mathfrak{F}_1, \mathfrak{F}_2$  are s-holomorphic everywhere on  $\Upsilon^\times(G)$ ;
- $G$  has the topology of a sphere, the condition (2.13) is relaxed on a ‘root’ quad (note that  $\mathfrak{F}_1, \mathfrak{F}_2$  cannot be s-holomorphic everywhere on  $\Upsilon^\times(G)$ );
- a finite graph  $G$ , the s-holomorphicity of  $\mathfrak{F}_1, \mathfrak{F}_2$  is relaxed at the boundary.

Clearly, the propagation equation (2.13) still holds for the *complex-valued* spinor  $\mathfrak{F}$  and hence one can define a complex-valued function  $\mathcal{S} := H_{\mathfrak{F}}$  on  $G^\bullet \cup G^\circ$  by (2.14). We interpret  $\mathcal{S}$  as an embedding of  $G$  into  $\mathbb{C}$  and call it *s-embedding*. Note that, similarly to Russkikh [2016, Section 3.2], the function  $\mathcal{S}$  can be also defined on  $z \in \diamond(G)$  by

$$(6.1) \quad \begin{aligned} \mathcal{S}(v_p^\bullet) - \mathcal{S}(z) &:= \cos \theta \cdot \mathfrak{F}(c_{p,0})\mathfrak{F}(c_{p,1}), \\ \mathcal{S}(z) - \mathcal{S}(v_q^\circ) &:= \sin \theta \cdot \mathfrak{F}(c_{0,q})\mathfrak{F}(c_{1,q}), \end{aligned}$$

where the two corners  $c_{p,0}$  and  $c_{p,1}$  (resp.,  $c_{0,q}$  and  $c_{1,q}$ ) are chosen on the same sheet of  $\Upsilon^\times(G)$ ; these equations are consistent with (2.14) due to (2.13).

It is easy to see that (2.13) also implies the identity

$$(6.2) \quad \begin{aligned} |\mathcal{S}(v_0^\bullet) - \mathcal{S}(v_0^\circ)| + |\mathcal{S}(v_1^\bullet) - \mathcal{S}(v_1^\circ)| &= |\mathfrak{F}(c_{00})|^2 + |\mathfrak{F}(c_{11})|^2 \\ &= |\mathfrak{F}(c_{01})|^2 + |\mathfrak{F}(c_{10})|^2 = |\mathcal{S}(v_0^\bullet) - \mathcal{S}(v_1^\circ)| + |\mathcal{S}(v_1^\bullet) - \mathcal{S}(v_0^\circ)|, \end{aligned}$$

which means that  $\mathcal{S}(v_0^\bullet), \mathcal{S}(v_0^\circ), \mathcal{S}(v_1^\bullet)$  and  $\mathcal{S}(v_1^\circ)$  form a *tangential* (though possibly non-convex) quadrilateral in the plane. In fact, one can easily see that  $\mathcal{S}(z)$ , if defined according to (6.1), is the center of the circle inscribed into this quadrilateral.

*Remark 6.1.* (i) Let us emphasize that the parameters  $\theta_e$  in the s-holomorphicity condition (2.13) are *defined* as  $\theta_e := 2 \arctan x_e$  and have no straightforward geometrical meaning similar to isoradial embeddings discussed above (though see (6.3)).

(ii) One can easily check that the isoradial embedding of the critical Z-invariant model is a particular case of the above construction in which  $\mathfrak{F}(c) = \varsigma \delta^{1/2} \cdot \bar{\eta}_c$ ; note that, in this case, the Dirac spinor (2.10) satisfies the propagation equation (2.13).

A priori, there is no guarantee that the combinatorics of the s-embedding constructed above matches the one of  $\diamond(G)$ , considered as an abstract *topological* (i.e., embedded into  $\mathbb{C}$  up to homotopies) graph: the images  $(\mathcal{S}(v_0^\bullet)\mathcal{S}(v_0^\circ)\mathcal{S}(v_1^\bullet)\mathcal{S}(v_1^\circ))$  of quads  $(v_0^\bullet v_0^\circ v_1^\bullet v_1^\circ)$  might overlap. Below we assume that this does *not* happen and, moreover, all  $(\mathcal{S}(v_0^\bullet)\mathcal{S}(v_0^\circ)\mathcal{S}(v_1^\bullet)\mathcal{S}(v_1^\circ))$  are nondegenerate and oriented counterclockwise (except maybe the ‘root’ one). We call such  $\mathcal{S}$  *proper s-embeddings*.

We now introduce a set of geometric parameters characterizing an s-embedding  $\mathcal{S}$  up to translations and rotations. For a quad  $(v_0^\bullet v_0^\circ v_1^\bullet v_1^\circ) = z \in \diamond(G)$ , let  $r_z$  denote the radius

of the circle inscribed into its image  $(\mathcal{S}(v_0^\bullet)\mathcal{S}(v_0^\circ)\mathcal{S}(v_1^\bullet)\mathcal{S}(v_1^\circ))$  and let

$$\phi_{zv_p^\bullet} := \frac{1}{2} \arg \frac{\mathcal{S}(v_{1-p}^\circ) - \mathcal{S}(v_p^\bullet)}{\mathcal{S}(v_p^\circ) - \mathcal{S}(v_p^\bullet)}, \quad \phi_{zv_q^\circ} := \frac{1}{2} \arg \frac{\mathcal{S}(v_q^\bullet) - \mathcal{S}(v_q^\circ)}{\mathcal{S}(v_{1-q}^\bullet) - \mathcal{S}(v_q^\circ)}$$

be the half-angles of  $(\mathcal{S}(v_0^\bullet)\mathcal{S}(v_0^\circ)\mathcal{S}(v_1^\bullet)\mathcal{S}(v_1^\circ))$ , note that  $\phi_{zv_0^\bullet} + \phi_{zv_0^\circ} + \phi_{zv_1^\bullet} + \phi_{zv_1^\circ} = \pi$  and  $|\mathcal{S}(v) - \mathcal{S}(z)| = (\sin \phi_{zv})^{-1} r_z$  for each of the four vertices  $v = v_0^\bullet, v_0^\circ, v_1^\bullet, v_1^\circ$ . A straightforward computation shows that

$$(6.3) \quad \tan \theta_{v_0^\bullet v_1^\bullet} = \left( \frac{\cot \phi_{zv_0^\circ} + \cot \phi_{zv_1^\circ}}{\cot \phi_{zv_0^\bullet} + \cot \phi_{zv_1^\bullet}} \right)^{1/2} = \left( \frac{\sin \phi_{zv_0^\bullet} \sin \phi_{zv_1^\bullet}}{\sin \phi_{zv_0^\circ} \sin \phi_{zv_1^\circ}} \right)^{1/2},$$

where  $\theta_e = 2 \arctan x_e$  is the standard parametrization of the Ising model weights.

*Remark 6.2.* It is worth mentioning that the construction of an s-embedding described above is *reversible*. Namely, given a proper embedding  $\mathcal{S}$  of  $(G^\bullet, G^\circ)$  into the complex plane formed by non-degenerate tangential quads, one can define a complex-valued spinor  $\mathcal{F}(c) := (\mathcal{S}(v^\bullet(c)) - \mathcal{S}(v^\circ(c)))^{1/2}$  on  $\Upsilon^\times(G)$  and deduce from (6.2) that (2.13) holds true for *some*  $\theta_e$  (which must then coincide with (6.3)). In other words, given  $\mathcal{S}$ , one can find Ising weights  $x_e = \tan \frac{1}{2} \theta_e$  on  $G$  and a pair  $\mathcal{F}_1, \mathcal{F}_2$  of real-valued spinors satisfying (2.13) with these  $\theta_e$  such that  $\mathcal{S} = \mathcal{H}_{\mathcal{F}}$ .

**6.2 S-subharmonicity.** Miraculously enough, it turns out that the cornerstone Lemma 3.1 actually *admits* a generalization to the setup described above, though not a straightforward one, cf. Remark 3.4. Let  $H$  be a function defined in a vicinity of a given vertex  $v^\bullet \in G^\bullet$  or  $v^\circ \in G^\circ$ . We define its *s-Laplacian*  $\Delta_{\mathcal{S}} H$  as

$$\begin{aligned} [\Delta_{\mathcal{S}} H](v^\bullet) &:= \sum_{v_1^\bullet \sim v^\bullet} a_{v^\bullet v_1^\bullet} (H(v_1^\bullet) - H(v^\bullet)) + \sum_{v^\circ \sim v^\bullet} b_{v^\bullet v^\circ} (H(v^\circ) - H(v^\bullet)), \\ [\Delta_{\mathcal{S}} H](v^\circ) &:= \sum_{v^\bullet \sim v^\circ} b_{v^\circ v^\bullet} (H(v^\bullet) - H(v^\circ)) - \sum_{v_1^\circ \sim v^\circ} a_{v^\circ v_1^\circ} (H(v_1^\circ) - H(v^\circ)), \end{aligned}$$

where, for each quad  $(v_0^\bullet v_0^\circ v_1^\bullet v_1^\circ) = z \in \diamond(G)$ , one has

$$(6.4) \quad a_{v_0^\bullet v_1^\bullet} = a_{v_1^\bullet v_0^\bullet} := r_z^{-1} \sin^2 \theta_{v_0^\bullet v_1^\bullet}, \quad a_{v_0^\circ v_1^\circ} = a_{v_1^\circ v_0^\circ} := r_z^{-1} \cos^2 \theta_{v_0^\bullet v_1^\bullet},$$

and, for each edge  $(v^\bullet v^\circ) = (v_0^\bullet(z) v_1^\circ(z)) = (v_0^\bullet(z') v_1^\circ(z'))$  separating  $z, z' \in \diamond(G)$ ,

$$(6.5) \quad b_{v^\bullet v^\circ} := a_{v^\circ v_0^\circ(z)} - \frac{r_z^{-1} \cot \phi_{zv^\bullet}}{\cot \phi_{zv^\bullet} + \cot \phi_{zv^\circ}} + a_{v^\circ v_1^\circ(z')} - \frac{r_{z'}^{-1} \cot \phi_{z'v^\bullet}}{\cot \phi_{z'v^\bullet} + \cot \phi_{z'v^\circ}}$$

and  $b_{v^\circ v^\bullet} := b_{v^\bullet v^\circ}$ , so that  $\Delta_{\mathcal{S}} = \Delta_{\mathcal{S}}^\top$  is symmetric (though not sign-definite).

*Remark 6.3.* For isoradial embeddings of graphs carrying the critical Z-invariant Ising model one has  $a_{v_0 \bullet v_1} = \delta^{-1} \tan \theta_{v_0 \bullet v_1}$ ,  $a_{v_0^\circ v_1^\circ} = \delta^{-1} \cot \theta_{v_0 \bullet v_1}$  and  $b_{v_0 v_1} = 0$ , thus  $\Delta_{\mathfrak{s}}$  is simply the direct sum of the two *signed* Laplacians  $\delta^{-1} \Delta^\bullet$  and  $-\delta^{-1} \Delta^\circ$ .

**Lemma 6.1.** *Let  $\mathfrak{s} = H_{\mathfrak{T}_1 + i\mathfrak{T}_2}$  be a proper  $s$ -embedding and a function  $H_F$  be obtained via (2.14) from a real-valued  $s$ -holomorphic spinor  $F$  defined on  $\Upsilon^\times(G)$  in a vicinity of a vertex  $v \in \Lambda(G)$ . Then,  $[\Delta_{\mathfrak{s}} H_F] \geq 0$ . Moreover,  $[\Delta_{\mathfrak{s}} H_F] = 0$  if and only if  $F$  is a linear combination of  $\mathfrak{T}_1$  and  $\mathfrak{T}_2$  on corners of  $G$  incident to  $v$ .*

*Proof.* As in the isoradial case (see Chelkak and Smirnov [2012, Proposition 3.6(iii)]), we are only able to check this by brute force though clearly a more conceptual explanation of this phenomenon must exist. E.g., if one numbers the quads  $z_1, \dots, z_n$  around  $v_0^\bullet \in G^\bullet$  counterclockwise and adopts the notation  $z_s = (v_0^\bullet v_{s-1}^\circ v_s^\circ v_s^\bullet)$ ,  $c_s = (v_0^\bullet v_s^\circ)$ , and  $\mathfrak{F}(c_s) = \alpha e^{i(\phi_1 + \dots + \phi_s)} \rho_s$  with  $\alpha = \exp[i \arg \mathfrak{F}(c_0)]$  and  $\rho_s = |\mathfrak{F}(c_s)| > 0$ , then

$$(6.6) \quad a_{v_0^\bullet v_s^\bullet} = \frac{\sin \theta_s \tan \theta_s}{\rho_{s-1} \rho_s \sin \phi_s}, \quad a_{v_{s-1}^\circ v_s^\circ} = \frac{\cos \theta_s}{\rho_{s-1} \rho_s \sin \phi_s},$$

$$(6.7) \quad b_{v_0^\bullet v_s^\circ} = \frac{\cos \theta_s}{\rho_{s-1} \rho_s \sin \phi_s} + \frac{\cos \theta_{s+1}}{\rho_s \rho_{s+1} \sin \phi_{s+1}} - \frac{\sin(\phi_s + \phi_{s+1})}{\rho_s^2 \sin \phi_s \sin \phi_{s+1}},$$

and  $[\Delta_{\mathfrak{s}} H](v_0^\bullet)$  turns out to be a non-negative quadratic form in the variables  $F(c_s)$ :

$$[\Delta_{\mathfrak{s}} H_F](v_0^\bullet) = Q_{\phi_1, \dots, \phi_n}^{(n)}(\rho_0^{-1} F(c_0), \dots, \rho_{n-1}^{-1} F(c_{n-1})) \geq 0,$$

see Chelkak and Smirnov [ibid., p. 543] for the definition of  $Q_{\phi_1, \dots, \phi_n}^{(n)}$ . The case  $v \in G^\circ$  is similar.  $\square$

**Definition 6.2.** *We call a function defined on (a subset of)  $\Lambda(G)$   $s$ -subharmonic if the inequality  $\Delta_{\mathfrak{s}} H \geq 0$  holds true pointwise and  $s$ -harmonic if  $\Delta_{\mathfrak{s}} H = 0$  pointwise.*

*Remark 6.4.* Though this is not fully clear at the moment, we hope that, at least in some situations of interest (e.g., see Section 6.4 or Lis [2017]),  $s$ -subharmonic functions  $H_F$  obtained via (2.14) are *a priori* close to  $s$ -harmonic ones; recall that this is exactly the viewpoint developed in Chelkak and Smirnov [2012, Section 3] for the critical Z-invariant model. Also, note that extending the domain of definition of  $H_F$  to  $\diamond(G)$  similarly to (6.1), one can easily see that thus obtained functions satisfy the maximum principle.

**6.3 Factorization of the  $s$ -Laplacian.** An important feature of the isoradial setup is the following factorization of the direct *sum* of  $\Delta^\bullet$  and  $\Delta^\circ$ , e.g. see Kenyon [2002] or Chelkak and Smirnov [2011]:

$$(6.8) \quad -\delta^{-1}(\Delta^\bullet + \Delta^\circ) = 16\partial_\Lambda^* R \partial_\Lambda = 16\bar{\partial}_\Lambda^* R \bar{\partial}_\Lambda,$$



where the Cauchy-Riemann operator  $\partial_\Lambda$  is given by (3.4) and  $R := \text{diag}\{r_z\}_{z \in \diamond(G)}$ ; note that one has  $r_z = \frac{1}{4}\delta^{-1}|v_1^\bullet - v_0^\bullet||v_1^\circ - v_0^\circ|$  in this special case. Since the s-Laplacian  $\Delta_\mathcal{S}$  is not sign-definite (recall that, on rhombic lattices,  $\Delta_\mathcal{S}$  is the *difference*  $\delta^{-1}(\Delta^\bullet - \Delta^\circ)$ ), the factorization (6.8) cannot be generalized directly. However, there exists a way to rewrite<sup>3</sup> it in the full generality of s-embeddings:

$$(6.9) \quad \Delta_\mathcal{S} = 16\partial_\mathcal{S}^* U^{-1} R \bar{\partial}_\mathcal{S} = 16\bar{\partial}_\mathcal{S}^* \bar{U}^{-1} R \partial_\mathcal{S},$$

where

$$[\bar{\partial}_\mathcal{S} H](z) := \frac{\mu_z}{4} \left[ \frac{H(v_0^\bullet)}{\mathcal{S}(v_0^\bullet) - \mathcal{S}(z)} + \frac{H(v_1^\bullet)}{\mathcal{S}(v_1^\bullet) - \mathcal{S}(z)} - \frac{H(v_0^\circ)}{\mathcal{S}(v_0^\circ) - \mathcal{S}(z)} - \frac{H(v_1^\circ)}{\mathcal{S}(v_1^\circ) - \mathcal{S}(z)} \right],$$

the prefactor  $\mu_z$  is chosen so that  $[\bar{\partial}_\mathcal{S} \bar{\mathcal{S}}](z) = 1$ , the operator  $\partial_\mathcal{S}$  is defined so that  $\bar{\partial}_\mathcal{S} \bar{H} = \bar{\partial}_\mathcal{S} \bar{H}$ , and  $U := \text{diag}(\mu_z)_{z \in \diamond(G)}$ . Moreover, the following is fulfilled:

**Lemma 6.3.** *A real-valued function  $H_1$  is (locally) s-harmonic on  $\Lambda(G)$  if and only if there exists (locally and hence globally in the simply connected setup) another real-valued s-harmonic function  $H_2$  on  $\Lambda(G)$  such that  $\bar{\partial}_\mathcal{S}(H_1 + iH_2) = 0$  on  $\diamond(G)$ . In other words, s-harmonic functions are real parts of those lying in the kernel of  $\bar{\partial}_\mathcal{S}$ .*

Another straightforward computation shows that  $\bar{\partial}_\mathcal{S} H = 0$  if  $H$  is a constant and that  $\bar{\partial}_\mathcal{S} \mathcal{S} = 0$ , which (together with the normalization  $\bar{\partial}_\mathcal{S} \bar{\mathcal{S}} = 1$ ) establishes some link of the difference operator  $\mathcal{S}$  with the standard complex structure on  $\mathbb{C}$ . In addition, one can check that  $\bar{\partial}_\mathcal{S} L_\mathcal{S} = 0$ , where the real-valued function  $L_\mathcal{S}$  is defined on  $\Lambda(G)$ , up to an additive constant, by  $L_\mathcal{S}(v^\bullet(c)) - L_\mathcal{S}(v^\circ(c)) := |\mathcal{S}(c)|^2$ .

**6.4 Doubly-periodic graphs.** We now briefly discuss s-embeddings of doubly-periodic graphs  $G$  carrying a *critical* Ising model. It was shown in Cimasoni and Duminil-Copin [2013] that the criticality condition is equivalent to the existence of two periodic functions in the kernel of the Kac–Ward matrix (2.18), which means (e.g., see Chelkak, Cimasoni, and Kassel [2017]) the existence of two linearly independent *periodic* (real-valued) *spinors*  $\mathfrak{F}_1, \mathfrak{F}_2$  on  $\Upsilon^\times(G)$ . Thus, up to a global scaling and rotation (corresponding to the multiplication of  $\mathfrak{F}$  by a constant), it remains to tune one *complex-valued parameter*  $\kappa$  in order to construct a periodic s-embedding  $\mathcal{S} = H_{\mathfrak{F}_1 + \kappa \mathfrak{F}_2}$  of  $G$ . The choice of  $\kappa$ , in particular, corresponds to the choice of the *conformal modulus*  $\tau$  of (the image under  $\mathcal{S}$  of) a fundamental domain of  $G$ . However, note that this dependence is not trivial: according to (2.14),  $\tau = \tau(\kappa)$  is the ratio of two quadratic polynomials constructed from  $\mathfrak{F}_1$  and  $\mathfrak{F}_2$ .

<sup>3</sup>For rhombic lattices, one has a very special intertwining identity  $\bar{\partial}_\Lambda = U \partial_\Lambda (-\text{Id}^\bullet + \text{Id}^\circ)$ .

Using (6.6) and (6.7), one can check that the  $s$ -Laplacian  $\Delta_{\mathfrak{s}}$  is essentially independent of the choice of  $\mathfrak{T}$ : changing  $\kappa$  results in the multiplication of all the coefficients of  $\Delta_{\mathfrak{s}}$  by a constant. Nevertheless, the operator  $\bar{\partial}_{\mathfrak{s}}$  is much more sensitive to this choice. We believe that the following picture is true<sup>4</sup>:

- Provided  $\mathfrak{T}_1$  and  $\mathfrak{T}_2$  are linearly independent, any choice of  $\kappa \notin \mathbb{R}$  leads either to a proper  $s$ -embedding or to the conjugate of a proper  $s$ -embedding.
- The kernel of  $\Delta_{\mathfrak{s}}$  in the space of periodic functions is two-dimensional.
- There exists a unique (up to conjugation) value  $\kappa = \kappa_L$  such that the function  $L_{\mathfrak{s}}$  is periodic. For  $\kappa \neq \kappa_L$ , the kernel of  $\bar{\partial}_{\mathfrak{s}}$  in the space of periodic functions consists of constants only. For  $\kappa = \kappa_L$ , it coincides with the kernel of  $\Delta_{\mathfrak{s}}$  and is spanned by constants and  $L_{\mathfrak{s}}$ .
- If  $\kappa = \kappa_L$ , there exists a *periodic* (and hence bounded) function  $\rho$  on  $\Lambda(G)$  such that the complex-valued function  $\mathfrak{S}^2 + \rho$  is  $s$ -harmonic (i.e., both its real and imaginary parts are  $s$ -harmonic, note that  $\rho = 0$  for rhombic lattices).

*Remark 6.5.* The last claim reveals the rotational (and, eventually, conformal) symmetry of the critical Ising model on  $(G, x)$ , which must show up if the conformal modulus  $\tau = \tau(\kappa_L)$  of the fundamental domain is tuned properly. Note that one should not expect that  $s$ -harmonic functions on a general  $s$ -embedding  $\mathfrak{S}$  of  $G$  behave like continuous harmonic functions even on large scales, not to mention periodic or quasi-periodic fluctuations. Indeed, as mentioned above, the operator  $\Delta_{\mathfrak{s}}$  is essentially independent of the choice of  $\kappa$  (and thus  $\tau(\kappa)$ ), hence in general one cannot hope for more than a skewed rotational symmetry despite of Lemma 6.3 and the basic properties  $\bar{\partial}_{\mathfrak{s}}1 = \bar{\partial}_{\mathfrak{s}}\mathfrak{S} = 0$  of the Cauchy–Riemann type operator  $\bar{\partial}_{\mathfrak{s}}$ . Nevertheless, by analogy with usual discrete harmonic functions, one can hope that some form of an invariance principle for  $s$ -harmonic functions can be found.

## 7 Open questions

Above, we already indicated several promising research directions basing on the analysis of  $s$ -holomorphic spinors, notably ‘CFT on the lattice level’ projects (see Section 4.3) and the study of  $s$ -embeddings of planar graphs. Besides universality questions, one can apply them to *random maps* (finite or infinite) carrying the Ising model, in an attempt to understand their conformal structure in the large size limit (one of the most straightforward

---

<sup>4</sup>At the moment we do not have a full proof. However, let us mention that one can justify the missing ingredients for  $s$ -embeddings close enough to isoradial ones using continuity arguments.

setups is to interpret a random quadrangulation as  $\diamond(G)$  and to work with the self-dual Ising weights  $x_{\text{sd}} = \sqrt{2} - 1$  on  $G^\bullet, G^\circ$ ).

Another challenging research direction, which cannot be reduced to the analysis of s-holomorphic observables and requires some other techniques to be developed, is a better understanding of *topological correlators*, cf. [Ikhlef, Jacobsen, and Saleur \[2016\]](#). E.g., one can consider the FK-Ising model with  $2k$  marked points on the boundary of  $\Omega^\delta$  and alternating wired/free/.../wired/free boundary conditions, cf. [\(5.3\)](#). There is the Catalan number  $C_k$  of possible patterns of interfaces matching these marked points and only  $2^{k-1} \ll C_k$  correlations of disorders on free arcs to handle. The situation with domain walls is even worse: no geometric information on those can be extracted directly *in discrete*: already in the simplest possible setup with four marked points and ‘+ / - / + / -’ boundary conditions, the only available way to prove the convergence of the crossing probabilities is first to prove the convergence of interfaces to hypergeometric SLEs and then to do computations *in continuum*, see [Izyurov \[2015\]](#).

*Remark 7.1.* In the context of double-dimer and CLE(4) loop ensembles, it was recently demonstrated by Dubédat in [Dubédat \[2014\]](#) that topological correlators can be treated via tau-functions associated with  $\text{SL}(2)$ -representations of the fundamental group of a punctured domain  $\Omega^\delta$  (see also [Basok and Chelkak \[n.d.\]](#)). This remarkable development raises the following question: could one attack topological correlators corresponding to  $\text{CLE}(\kappa)$ ,  $\kappa \neq 4$ , replacing  $\text{SL}(2)$  by relevant quantum groups? If so, could one use the critical Ising model, once again, as a laboratory to reveal and to analyze these structures in discrete? Note also that a detailed understanding of such topological correlators would also pave the way to a better understanding of the famous Coulomb gas approach to the critical lattice models; see [Nienhuis \[1984\]](#) and [Smirnov \[2006, Section 5.3\]](#).

As discussed above, both the convergence of critical Ising correlation functions (to CFT ones) and that of loop ensembles (to CLEs) as  $\Omega^\delta \rightarrow \Omega$  are now understood in detail. Moreover, a scaling limit  $\sigma_\Omega(z)$ ,  $z \in \Omega$ , of the spin field  $\{\sigma_u\}_{u \in V^\circ(\Omega^\delta)}$  viewed as a random distribution (generalized function) was constructed in the work of Camia, Garban and Newman, see [Camia, Garban, and Newman \[2015\]](#). This leads to natural measurability questions: e.g., is it true that  $\sigma_\Omega$  is (not) measurable with respect to the nested  $\text{CLE}(3)$  – the limit of domain walls – and vice versa? In fact, [Camia, Garban, and Newman \[ibid.\]](#) ensures the measurability of  $\sigma_\Omega$  with respect to the limit of FK-Ising clusters –  $\text{CLE}(16/3)$  – but the latter contains more information, cf. [Miller, Sheffield, and Werner \[2017\]](#). Also, one can wonder whether it is possible to construct the *energy density* correlation functions out of these CLEs (e.g., via some regularized ‘occupation density’ of loops)? If so, one could then try to generalize such a construction to  $\kappa \neq 3$  (note that the Ising spin field  $\sigma$  is more model-specific from the CLE perspective than the energy operator  $\varepsilon \sim \phi_{3,1}$ , cf. [Dotsenko and Fateev \[1984, p. 319\]](#)).

We conclude this essay on the *critical* Ising model by a famous question on the *super-critical* one: to prove that each fixed value  $x > x_{\text{crit}}$  gives rise to the CLE(6) as the scaling limit of domain walls configurations. Note that a possible approach to this via the study of *massive* theories was suggested in [Makarov and Smirnov \[2010, Question 4.8\]](#).

## References

- M. Aizenman and A. Burchard (1999). “Hölder regularity and dimension bounds for random curves”. *Duke Math. J.* 99.3, pp. 419–453. MR: [1712629](#) (cit. on pp. [2834](#), [2836](#)).
- Mikhail Basok and Dmitry Chelkak (n.d.). *Tau-functions à la Dubédat and probabilités of cylindric events for double-dimers and CLE(4)*. In preparation (cit. on p. [2844](#)).
- Stéphane Benoist, Hugo Duminil-Copin, and Clément Hongler (2016). “Conformal invariance of crossing probabilities for the Ising model with free boundary conditions”. *Ann. Inst. Henri Poincaré Probab. Stat.* 52.4, pp. 1784–1798. MR: [3573295](#) (cit. on pp. [2834](#), [2836–2838](#)).
- Stéphane Benoist and Clément Hongler (Apr. 2016). “The scaling limit of critical Ising interfaces is CLE (3)”. arXiv: [1604.06975](#) (cit. on pp. [2834](#), [2836](#), [2837](#)).
- Cédric Boutillier, Béatrice De Tilière, and Kilian Raschel (Dec. 2016). “The Z-invariant Ising model via dimers”. arXiv: [1612.09082](#) (cit. on p. [2827](#)).
- Federico Camia, Christophe Garban, and Charles M. Newman (2015). “Planar Ising magnetization field I. Uniqueness of the critical scaling limit”. *Ann. Probab.* 43.2, pp. 528–571. MR: [3305999](#) (cit. on p. [2844](#)).
- Dmitry Chelkak (2016a). “2D Ising model: correlation functions at criticality via Riemann-type boundary value problems”. To appear in Proceedings of the 7ECM. arXiv: [1605.09035](#) (cit. on p. [2831](#)).
- (2016b). “Robust discrete complex analysis: a toolbox”. *Ann. Probab.* 44.1, pp. 628–683. MR: [3456348](#) (cit. on p. [2836](#)).
- Dmitry Chelkak, David Cimasoni, and Adrien Kassel (2017). “Revisiting the combinatorics of the 2D Ising model”. *Ann. Inst. Henri Poincaré D* 4.3, pp. 309–385. MR: [3713019](#) (cit. on pp. [2820](#), [2823](#), [2824](#), [2826](#), [2828](#), [2842](#)).
- Dmitry Chelkak, Hugo Duminil-Copin, and Clément Hongler (2016). “Crossing probabilities in topological rectangles for the critical planar FK-Ising model”. *Electron. J. Probab.* 21, Paper No. 5, 28. MR: [3485347](#) (cit. on pp. [2836](#), [2838](#)).
- Dmitry Chelkak, Hugo Duminil-Copin, Clément Hongler, Antti Kemppainen, and Stanislav Smirnov (2014). “Convergence of Ising interfaces to Schramm’s SLE curves”. *C. R. Math. Acad. Sci. Paris* 352.2, pp. 157–161. MR: [3151886](#) (cit. on pp. [2834–2837](#)).
- Dmitry Chelkak, Alexander Glazman, and Stanislav Smirnov (Apr. 2016). “Discrete stress-energy tensor in the loop  $O(n)$  model”. arXiv: [1604.06339](#) (cit. on p. [2833](#)).

- Dmitry Chelkak, Clément Hongler, and Konstantin Izyurov (n.d.). “Spins, disorders and fermions in the critical 2D Ising model: convergence and fusion rules” (cit. on pp. [2821](#), [2825](#), [2829–2833](#)).
- Dmitry Chelkak, Clément Hongler, and Konstantin Izyurov (2015). “[Conformal invariance of spin correlations in the planar Ising model](#)”. *Ann. of Math.* (2) 181.3, pp. 1087–1138. MR: [3296821](#) (cit. on pp. [2828](#), [2829](#), [2832](#), [2833](#)).
- Dmitry Chelkak and Konstantin Izyurov (2013). “[Holomorphic spinor observables in the critical Ising model](#)”. *Comm. Math. Phys.* 322.2, pp. 303–332. MR: [3077917](#) (cit. on p. [2828](#)).
- Dmitry Chelkak and Stanislav Smirnov (2011). “[Discrete complex analysis on isoradial graphs](#)”. *Adv. Math.* 228.3, pp. 1590–1630. MR: [2824564](#) (cit. on pp. [2828](#), [2830](#), [2841](#)).
- (2012). “[Universality in the 2D Ising model and conformal invariance of fermionic observables](#)”. *Invent. Math.* 189.3, pp. 515–580. MR: [2957303](#) (cit. on pp. [2823](#), [2828–2831](#), [2834](#), [2836](#), [2837](#), [2841](#)).
- David Cimasoni and Hugo Duminil-Copin (2013). “[The critical temperature for the Ising model on planar doubly periodic graphs](#)”. *Electron. J. Probab.* 18, no. 44, 18. MR: [3040554](#) (cit. on pp. [2838](#), [2842](#)).
- Vl. S. Dotsenko and V. A. Fateev (1984). “[Conformal algebra and multipoint correlation functions in 2D statistical models](#)”. *Nuclear Phys. B* 240.3, pp. 312–348. MR: [762194](#) (cit. on p. [2844](#)).
- Julien Dubédat (Dec. 2011). “[Exact bosonization of the Ising model](#)”. arXiv: [1112.4399](#) (cit. on p. [2822](#)).
- Julien Dubédat (2011). “[Topics on abelian spin models and related problems](#)”. *Probab. Surv.* 8, pp. 374–402. MR: [2861134](#) (cit. on p. [2822](#)).
- Julien Dubédat (Mar. 2014). “[Double dimers, conformal loop ensembles and isomonodromic deformations](#)”. arXiv: [1403.6076](#) (cit. on p. [2844](#)).
- Julien Dubédat (2015). “[Dimers and families of Cauchy-Riemann operators I](#)”. *J. Amer. Math. Soc.* 28.4, pp. 1063–1167. MR: [3369909](#) (cit. on p. [2833](#)).
- Hugo Duminil-Copin and Stanislav Smirnov (2012). “Conformal invariance of lattice models”. In: *Probability and statistical physics in two and more dimensions*. Vol. 15. Clay Math. Proc. Amer. Math. Soc., Providence, RI, pp. 213–276. MR: [3025392](#) (cit. on pp. [2828](#), [2837](#)).
- Reza Gheissari, Clément Hongler, and Sung Chul Park (Dec. 2013). “[Ising Model: Local Spin Correlations and Conformal Invariance](#)”. arXiv: [1312.4446](#) (cit. on pp. [2829](#), [2833](#)).
- Clement Hongler (2010). “Conformal invariance of Ising model correlations”. No. 4228. PhD thesis. University of Geneva (cit. on p. [2832](#)).

- Clément Hongler and Kalle Kytölä (2013). “[Ising interfaces and free boundary conditions](#)”. *J. Amer. Math. Soc.* 26.4, pp. 1107–1189. MR: [3073886](#) (cit. on pp. [2829](#), [2834](#), [2837](#), [2838](#)).
- Clément Hongler, Kalle Kytölä, and Fredrik Viklund (2017). “[Conformal Field Theory at the lattice level: discrete complex analysis and Virasoro structure](#)”. arXiv: [1307.4104](#) (cit. on p. [2833](#)).
- Clément Hongler and Stanislav Smirnov (2013). “[The energy density in the planar Ising model](#)”. *Acta Math.* 211.2, pp. 191–225. MR: [3143889](#) (cit. on pp. [2828](#), [2829](#), [2832](#)).
- Yacine Ikhlef, Jesper Lykke Jacobsen, and Hubert Saleur (2016). “[Three-point functions in  \$c \leq 1\$  Liouville theory and conformal loop ensembles](#)”. *Phys. Rev. Lett.* 116.13, pp. 130601, 5. MR: [3600546](#) (cit. on p. [2844](#)).
- Konstantin Izyurov (2015). “[Smirnov’s observable for free boundary conditions, interfaces and crossing probabilities](#)”. *Comm. Math. Phys.* 337.1, pp. 225–252. MR: [3324162](#) (cit. on pp. [2829](#), [2831](#), [2837](#), [2844](#)).
- Leo P. Kadanoff and Horacio Ceva (1971). “Determination of an operator algebra for the two-dimensional Ising model”. *Phys. Rev. B* (3) 3, pp. 3918–3939. MR: [0389111](#) (cit. on pp. [2819](#), [2822](#)).
- Antti Kemppainen and Stanislav Smirnov (n.d.). “Conformal invariance in random cluster models. II. Full scaling limit as a branching SLE” (cit. on pp. [2834](#), [2835](#), [2837](#), [2838](#)).
- (Sept. 2015). “[Conformal invariance of boundary touching loops of FK Ising model](#)”. arXiv: [1509.08858](#) (cit. on pp. [2834](#), [2835](#), [2837](#), [2838](#)).
  - (2017). “[Random curves, scaling limits and Loewner evolutions](#)”. *Ann. Probab.* 45.2, pp. 698–779. MR: [3630286](#) (cit. on pp. [2834](#), [2836](#)).
- R. Kenyon (2002). “[The Laplacian and Dirac operators on critical planar graphs](#)”. *Invent. Math.* 150.2, pp. 409–439. MR: [1933589](#) (cit. on pp. [2828](#), [2841](#)).
- Robert P. Langlands, Marc-André Lewis, and Yvan Saint-Aubin (2000). “[Universality and conformal invariance for the Ising model in domains with boundary](#)”. *J. Statist. Phys.* 98.1-2, pp. 131–244. MR: [1745839](#) (cit. on p. [2834](#)).
- Marcin Lis (2016). “[A short proof of the Kac-Ward formula](#)”. *Ann. Inst. Henri Poincaré D* 3.1, pp. 45–53. MR: [3462629](#) (cit. on p. [2826](#)).
- (2017). “[Circle patterns and critical Ising models](#)”. arXiv: [1712.08736](#) (cit. on pp. [2838](#), [2841](#)).
- Nikolai Makarov and Stanislav Smirnov (2010). “[Off-critical lattice models and massive SLEs](#)”. In: *XVIth International Congress on Mathematical Physics*. World Sci. Publ., Hackensack, NJ, pp. 362–371. MR: [2730811](#) (cit. on p. [2845](#)).
- Barry M. McCoy and Tai Tsun Wu (1973). *The two-dimensional Ising model*. Harvard University Press, Cambridge, MA, pp. xvi+418. MR: [3618829](#) (cit. on p. [2819](#)).
- Christian Mercat (2001). “[Discrete Riemann surfaces and the Ising model](#)”. *Comm. Math. Phys.* 218.1, pp. 177–216. MR: [1824204](#) (cit. on pp. [2823](#), [2824](#), [2828](#)).

- Jason Miller, Scott Sheffield, and Wendelin Werner (2017). “CLE percolations”. *Forum Math. Pi* 5, e4, 102. MR: [3708206](#) (cit. on pp. [2834](#), [2844](#)).
- Giuseppe Mussardo (2010). *Statistical field theory*. Oxford Graduate Texts. An introduction to exactly solved models in statistical physics. Oxford University Press, Oxford, pp. xxii+755. MR: [2559725](#) (cit. on pp. [2819](#), [2833](#), [2834](#)).
- Bernard Nienhuis (1984). “Critical behavior of two-dimensional spin models and charge asymmetry in the Coulomb gas”. *J. Statist. Phys.* 34.5-6, pp. 731–761. MR: [751711](#) (cit. on p. [2844](#)).
- Marianna Russkikh (Nov. 2016). “Dimers in piecewise Temperley domains”. arXiv: [1611.07884](#) (cit. on p. [2839](#)).
- Oded Schramm (2000). “Scaling limits of loop-erased random walks and uniform spanning trees”. *Israel J. Math.* 118, pp. 221–288. MR: [1776084](#) (cit. on p. [2834](#)).
- Scott Sheffield (2009). “Exploration trees and conformal loop ensembles”. *Duke Math. J.* 147.1, pp. 79–129. MR: [2494457](#) (cit. on pp. [2834](#), [2837](#)).
- Scott Sheffield and Wendelin Werner (2012). “Conformal loop ensembles: the Markovian characterization and the loop-soup construction”. *Annals of Math. (2)* 176.3, pp. 1827–1917. MR: [2979861](#) (cit. on p. [2834](#)).
- Stanislav Smirnov (2006). “Towards conformal invariance of 2D lattice models”. In: *International Congress of Mathematicians. Vol. II*. Eur. Math. Soc., Zürich, pp. 1421–1451. MR: [2275653](#) (cit. on pp. [2819](#), [2824](#), [2825](#), [2827–2830](#), [2834](#), [2835](#), [2837](#), [2844](#)).
- (2010a). “Conformal invariance in random cluster models. I. Holomorphic fermions in the Ising model”. *Ann. of Math. (2)* 172.2, pp. 1435–1467. MR: [2680496](#) (cit. on pp. [2824](#), [2825](#), [2827–2830](#), [2834](#), [2837](#)).
  - (2010b). “Discrete complex analysis and probability”. In: *Proceedings of the International Congress of Mathematicians. Volume I*. Hindustan Book Agency, New Delhi, pp. 595–621. MR: [2827906](#) (cit. on p. [2828](#)).

Received 2017-12-12.

DMITRY CHELKAK  
 DÉPARTEMENT DE MATHÉMATIQUES ET APPLICATIONS DE L’ENS  
 ÉCOLE NORMALE SUPÉRIEURE PSL RESEARCH UNIVERSITY  
 CNRS UMR 8553, PARIS 5ÈME  
 FRANCE

and

ST. PETERSBURG DEPARTMENT OF STEKLOV MATHEMATICAL INSTITUTE RAS  
 ST. PETERSBURG  
 RUSSIA  
[Dmitry.Chelkak@ens.fr](mailto:Dmitry.Chelkak@ens.fr)

# SIXTY YEARS OF PERCOLATION

HUGO DUMINIL-COPIN

## Abstract

Percolation models describe the inside of a porous material. The theory emerged timidly in the middle of the twentieth century before becoming one of the major objects of interest in probability and mathematical physics. The golden age of percolation is probably the eighties, during which most of the major results were obtained for the most classical of these models, named Bernoulli percolation, but it is really the two following decades which put percolation theory at the crossroad of several domains of mathematics. In this broad review, we propose to describe briefly some recent progress as well as some famous challenges remaining in the field. This review is not intended to probabilists (and a fortiori not to specialists in percolation theory): the target audience is mathematicians of all kinds.

## 1 A brief history of Bernoulli percolation

**1.1 What is percolation?** Intuitively, it is a simplistic probabilistic model for a porous stone. The inside of the stone is described as a random maze in which water can flow. The question then is to understand which part of the stone will be wet when immersed in a bucket of water. Mathematically, the material is modeled as a random subgraph of a reference graph  $\mathbb{G}$  with (countable) vertex-set  $\mathbb{V}$  and edge-set  $\mathbb{E}$  (this is a subset of unordered pairs of elements in  $\mathbb{V}$ ).

Percolation on  $\mathbb{G}$  comes in two kinds, *bond* or *site*. In the former, each edge  $e \in \mathbb{E}$  is either *open* or *closed*, a fact which is encoded by a function  $\omega$  from the set of edges to  $\{0, 1\}$ , where  $\omega(e)$  is equal to 1 if the edge  $e$  is open, and 0 if it is closed. We think of an open edge as being open to the passage of water, while closed edges are not. A bond percolation model then consists in choosing edges of  $\mathbb{G}$  to be open or closed at random.

---

This research was funded by an IDEX Chair from Paris Saclay, by the NCCR SwissMap from the Swiss NSF and the ERC grant 757296 CRIBLAM. We thank David Cimasoni, Sébastien Martineau, Aran Raoufi and Vincent Tassion for their comments on the manuscript.



Site percolation is the same as bond percolation except that, this time, vertices  $v \in \mathbb{V}$  are either open or closed, and therefore  $\omega$  is a (random) function from  $\mathbb{V}$  to  $\{0, 1\}$ .

The simplest and oldest model of bond percolation, called *Bernoulli percolation*, was introduced by [Broadbent and Hammersley \[1957\]](#). In this model, each edge is open with probability  $p$  in  $[0, 1]$  and therefore closed with probability  $1 - p$ , independently of the state of other edges. Equivalently, the  $\omega(e)$  for  $e \in \mathbb{E}$  are independent Bernoulli random variables of parameter  $p$ .

Probabilists are interested in connectivity properties of the random object obtained by taking the graph induced by  $\omega$ . In the case of bond percolation, the vertices of this graph are the vertices of  $\mathbb{G}$ , and the edges are given by the open edges only. In the case of site percolation, the graph is the subgraph of  $\mathbb{G}$  induced by the open vertices, i.e. the graph composed of open vertices and edges between them.

Let us focus for a moment on Bernoulli percolation on the hypercubic lattice  $\mathbb{Z}^d$  with vertex-set given by the points of  $\mathbb{R}^d$  with integer coordinates, and edges between vertices at Euclidean distance 1 of each other. The simplest connectivity property to study is the fact that the connected component of the origin is finite or not. Set  $\theta(p)$  for the probability that the origin is in an infinite connected component of  $\omega$ . The union bound easily implies that the probability that the origin is connected to distance  $n$  is smaller than  $(2dp)^n$  (simply use the fact that one of the less than  $(2d)^n$  self-avoiding paths of length  $n$  starting from the origin must be made of open edges only, as well as the union bound). As a consequence, one deduces that  $\theta(p) = 0$  as soon as  $p < 1/(2d)$ . This elementary argument was described in the first paper [Broadbent and Hammersley \[ibid.\]](#) on percolation theory. On  $\mathbb{Z}^2$ , Harris drastically improved this result in [Harris \[1960\]](#) by showing that  $\theta(\frac{1}{2}) = 0$ .

A slightly harder argument [Hammersley \[1959\]](#) involving Peierls's argument (left to the reader) shows that when  $d \geq 2$  and  $p$  is close to 1, then  $\theta(p)$  is strictly positive. This suggests the existence of a *phase transition* in the model: for some values of  $p$ , connected components are all finite, while for others, there exists an infinite connected component in  $\omega$ . One can in fact state a more precise result [Broadbent and Hammersley \[1957\]](#). For Bernoulli percolation on transitive<sup>1</sup> (infinite) graphs, there exists  $p_c(\mathbb{G}) \in [0, 1]$  such that the probability that there is an infinite connected component in  $\omega$  is zero if  $p < p_c(\mathbb{G})$ , and one if  $p > p_c(\mathbb{G})$  (note that nothing is said about what happens at criticality). This is an archetypical example of a phase transition in statistical physics: as the parameter  $p$  (which can be interpreted physically as the porosity of the stone) is varied continuously through the value  $p_c(\mathbb{G})$ , the probability of having an infinite connected component jumps from 0 to 1.

---

<sup>1</sup>A graph is *transitive* if its group of automorphisms acts transitively on its vertices.

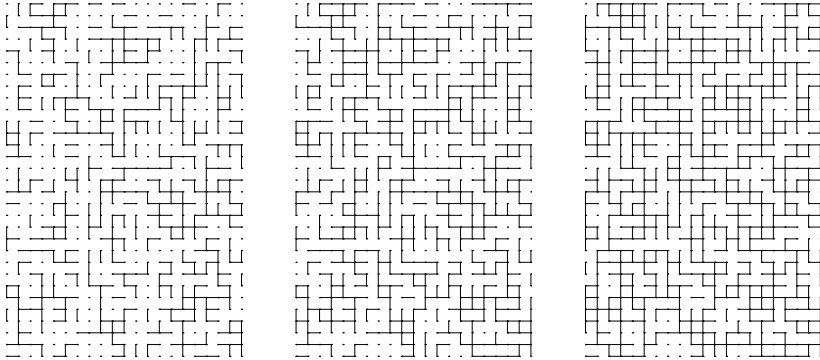


Figure 1: A sampled configuration  $\omega$  of Bernoulli bond percolation on the square lattice  $\mathbb{Z}^2$  for the values of the parameter  $p < 1/2$ ,  $p = 1/2$  and  $p > 1/2$ .

**1.2 The eighties: the Golden Age of Bernoulli percolation.** The eighties are famous for pop stars like Michael Jackson and Madonna, and a little bit less for probabilists such as Harry Kesten and Michael Aizenman. Nonetheless, these mathematicians participated intensively in the amazing progress that the theory underwent during this period.

The decade started in a firework with Kesten's Theorem [Kesten \[1980\]](#) showing that the critical point of Bernoulli bond percolation on the square lattice  $\mathbb{Z}^2$  is equal to  $1/2$ . This problem drove most of the efforts in the field until its final solution by Kesten, and some of the ideas developed in the proof became instrumental in the thirty years that followed. The strategy is based on an important result obtained simultaneously by [Russo \[1978\]](#) and [Seymour and Welsh \[1978\]](#), which will be discussed in more details in the next sections.

While the two-dimensional case concentrated most of the early focus, the model is, of course, not restricted to the graph  $\mathbb{Z}^2$ . On  $\mathbb{Z}^d$ , [Menshikov \[1986\]](#) at the same time as [Aizenman and Barsky \[1987\]](#) showed that not only the probability of being connected to distance  $n$  is going to 0 when  $p < p_c(\mathbb{Z}^d)$ , but that in fact this quantity is decaying exponentially fast, in the sense that there exists  $c > 0$  depending on  $p$  (and  $d$ ) such that

$$\theta_n(p) \stackrel{\text{def}}{=} \mathbb{P}_p[0 \text{ is connected to distance } n] \leq \exp(-cn)$$

for every  $n \geq 1$ . This result, known under the name of *sharpness of the phase transition*, provides a very strong control of the size of connected components. In particular it says that, when  $p < p_c$ , the largest connected component in a box of size  $n$  is typically of size  $\log n$ . It is the cornerstone of the understanding of percolation in the *subcritical regime*  $p < p_c$ , and as such, represents an important breakthrough in the field.

Properties of the *supercritical regime*  $p > p_c(\mathbb{Z}^d)$  were also studied in detail during this period. For instance, it is natural to ask oneself whether the infinite connected component is unique or not in this regime. In 1987, [Aizenman, Kesten, and Newman \[1987\]](#) showed that this is indeed the case<sup>2</sup>. The proof relied on delicate properties of Bernoulli percolation, and did not extend easily to more general models. Two years later, Burton and Keane proposed a beautiful argument [Burton and Keane \[1989\]](#), which probably deserves its place in The Book, showing by ergodic means that a large class of percolation models has a unique infinite connected component in the supercritical regime. A consequence of this theorem is the continuity of  $p \mapsto \theta(p)$  when  $p > p_c(\mathbb{Z}^d)$ . Of course, many other impressive results concerning the supercritical regime were obtained around the same period, but the lack of space refrains us from describing them in detail.

The understanding of percolation at  $p = p_c(\mathbb{Z}^d)$  also progressed in the late eighties and in the beginning of the nineties. Combined with the early work of [Harris \[1960\]](#) who proved  $\theta(1/2) = 0$  on  $\mathbb{Z}^2$ , Kesten's result directly implies that  $\theta(p_c) = 0$ . In dimension  $d \geq 19$ , [Hara and Slade \[1990\]](#) used a technique known under the name of *lace expansion* to show that critical percolation exhibits a *mean-field behavior*, meaning that the critical exponents describing the phase transition are matching those predicted by the so-called mean-field approximation. In particular, the mean-field behavior implies that  $\theta(p_c)$  is equal to 0. Each few years, more delicate uses of the lace-expansion enable to reduce the dimension starting at which the mean-field behavior can be proved: the best known result today is  $d \geq 11$  [Fitzner and van der Hofstad \[2015\]](#).

One may wonder whether it would be possible to use the lace expansion to prove that  $\theta(p_c)$  is equal to 0 for every dimension  $d \geq 3$ . Interestingly, the mean-field behavior is expected to hold only when  $d \geq 6$ , and to fail for dimensions  $d \leq 5$  (making the lace expansion obsolete). This leaves the intermediate dimensions 3, 4 and 5 as a beautiful challenge to mathematicians. In particular, the following question is widely considered as the major open question in the field.

**Conjecture 1.** *Show that  $\theta(p_c) = 0$  on  $\mathbb{Z}^d$  for every  $d \geq 3$ .*

This conjecture, often referred to as the “ $\theta(p_c) = 0$  conjecture”, is one of the problems that Harry Kesten was describing in the following terms in his famous 1982 book [Kesten \[1982\]](#):

*“Quite apart from the fact that percolation theory had its origin in an honest applied problem, it is a source of fascinating problems of the best kind a mathematician can wish for: problems which are easy to state with a minimum of preparation, but whose solutions are (apparently) difficult and require new methods.”*

---

<sup>2</sup>Pictorially, in two dimensions, the infinite connected component has properties similar to those of  $\mathbb{Z}^d$  and can be seen as an ocean. The finite connected components can then be seen as small lakes separated from the sea by the closed edges (which somehow can be seen as the land forming finite islands).

It would be unfair to say that the understanding of critical percolation is non-existent for  $d \in \{3, 4, 5\}$ . Barsky, G. R. Grimmett, and Newman [1991] proved that the probability that there exists an infinite connected component in  $\mathbb{N} \times \mathbb{Z}^{d-1}$  is zero for  $p = p_c(\mathbb{Z}^d)$ . It seems like a small step to bootstrap this result to the non-existence of an infinite connected component in the full space  $\mathbb{Z}^d$  ... But it is not. More than twenty five years after Barsky, G. R. Grimmett, and Newman [ibid.], the conjecture still resists and basically no improvement has been obtained.

**1.3 The nineties: the emergence of new techniques.** Percolation theory underwent a major mutation in the 90's and early 00's. While some of the historical questions were solved in the previous decade, new challenges appeared at the periphery of the theory. In particular, it became clear that a deeper understanding of percolation would require the use of techniques coming from a much larger range of mathematics. As a consequence, Bernoulli percolation took a new place at the crossroad of several domains of mathematics, a place in which it is not just a probabilistic model anymore.

**1.3.1 Percolation on groups.** In a beautiful paper entitled *Percolation beyond  $\mathbb{Z}^d$ , many questions and a few answers* Benjamini and Schramm [1996] underlined the relevance of Bernoulli percolation on Cayley graphs<sup>3</sup> of finitely generated infinite groups by proposing a list of problems relating properties of Bernoulli percolation to properties of groups. The paper triggered a number of new problems in the field and drew the attention of the community working in geometric group theory on the potential applications of percolation theory.

A striking example of a connection between the behavior of percolation and properties of groups is provided by the following conjecture, known as the “ $p_c < p_u$  conjecture”. Let  $p_u(\mathbb{G})$  be the smallest value of  $p$  for which the probability that there exists a *unique* infinite connected component is one. On the one hand, the uniqueness result Burton and Keane [1989] mentioned in the previous section implies that  $p_c(\mathbb{Z}^d) = p_u(\mathbb{Z}^d)$ . On the other hand, one can easily convince oneself that, on an infinite tree  $\mathbb{T}_d$  in which each vertex is of degree  $d + 1$ , one has  $p_c(\mathbb{T}_d) = 1/d$  and  $p_u(\mathbb{T}_d) = 1$ . More generally, the  $p_c < p_u$  conjecture relates the possibility of infinitely many connected components to the property of non-amenability<sup>4</sup> of the underlying group.

<sup>3</sup>The Cayley graph  $\mathbb{G} = \mathbb{G}(G, S)$  of a finitely generated group  $G$  with a symmetric system of generators  $S$  is the graph with vertex-set  $\mathbb{V} = G$  and edge-set given by the unordered pairs  $\{x, y\} \subset G$  such that  $yx^{-1} \in S$ . For instance,  $\mathbb{Z}^d$  is a Cayley graph for the free abelian group with  $d$  generators.

<sup>4</sup> $G$  is *non-amenable* if for any Cayley graph  $\mathbb{G}$  of  $G$ , the infimum of  $|\partial A|/|A|$  on non-empty finite subsets  $A$  of  $G$  is strictly positive, where  $\partial A$  denotes the *boundary* of  $A$  (i.e. the set of  $x \in A$  having one neighbor outside  $A$ ) and  $|B|$  is the cardinality of the set  $B$ .

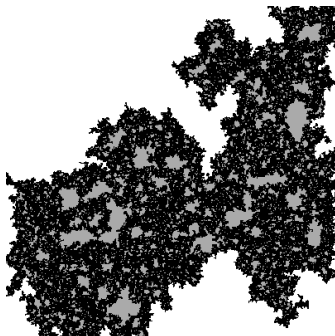


Figure 2: A large connected component of  $\omega$  for critical bond percolation on the square lattice  $\mathbb{Z}^2$  (simulation by Vincent Beffara).

**Conjecture 2** (Benjamini-Schramm). *Consider a Cayley graph  $\mathbb{G}$  of a finitely generated (infinite) group  $G$ . Then*

$$p_c(\mathbb{G}) < p_u(\mathbb{G}) \iff G \text{ is non-amenable.}$$

The most impressive progress towards this conjecture was achieved by [Pak and Smirnova-Nagnibeda \[2000\]](#) who provided a group theoretical argument showing that any non-amenable group possesses a multi-system of generators for which the corresponding Cayley graph satisfies  $p_c < p_u$ . This is a perfect example of an application of Geometric Group Theory to probability. Nicely enough, the story sometimes goes the other way and percolation shed a new light on classical questions on group theory. The following example perfectly illustrates this cross-fertilization.

In 2009, [Gaboriau and Lyons \[2009\]](#) provided a measurable solution to von Neumann's (and Day's) famous problem on non-amenable groups. While it is simple to show that a group containing the free group  $F_2$  as a subgroup is non-amenable, it is non-trivial to determine whether the converse is true. [Ol'sanskii \[1980\]](#) showed in 1980 that this is not the case, but [Whyte \[1999\]](#) gave a very satisfactory geometric solution: a finitely generated group is non-amenable if and only if it admits a partition into pieces that are all uniformly bi-lipschitz equivalent to the regular four-valent tree  $\mathbb{T}_3$ . Bernoulli percolation was used by Gaboriau and Lyons to show the measurable counterpart of this theorem, a result which has many important applications in the ergodic theory of group actions.

**1.3.2 Complex analysis and percolation.** The nineties also saw a renewed interest in questions about planar percolation. The impressive developments in the eighties of

Conformal Field Theory, initiated by [Belavin, Polyakov, and Zamolodchikov \[1984\]](#), suggested that the scaling limit of planar Bernoulli percolation is conformally invariant at criticality. From a mathematical perspective, the notion of conformal invariance of the entire model is ill-posed, since the meaning of scaling limit depends on the object under study (interfaces, size of connected components, crossings, etc). In 1992, the observation that properties of interfaces should also be conformally invariant led [Langlands, Pouliot, and Saint-Aubin \[1994\]](#) to publish numerical values in agreement with the conformal invariance in the scaling limit of crossing probabilities; see [Section 3.1](#) for a precise definition (the authors attribute this conjecture to Aizenman). The same year, the physicist [J. L. Cardy \[1992\]](#) proposed an explicit formula for the limit of crossing probabilities in rectangles of fixed aspect ratio.

These two papers, while numerical (for the first one) and physical (for the second one), attracted many mathematicians to the domain. In 2001, [Smirnov \[2001\]](#) proved Cardy's formula for critical site percolation on the triangular lattice, hence rigorously providing a concrete example of a conformally invariant property of the model. In parallel, a major breakthrough shook the field of probability. In 2000, [Schramm \[2000\]](#) introduced a random process, now called the *Schramm-Loewner Evolution*, describing the behavior of interfaces between open and closed sites. Very shortly after Smirnov's proof of Cardy's formula, the complete description of the scaling limit of site percolation was obtained, including the description of the full "loop ensemble" corresponding to the interfaces bordering each connected component by [Camia and Newman \[2006\]](#), see [Beffara and Duminil-Copin \[2013\]](#) for more references on this beautiful subject.

Smirnov's theorem and the Schramm-Loewner Evolution share a common feature: they both rely on complex analysis in a deep way. The first result uses discrete complex analysis, i.e. the study of functions on graphs that approximate holomorphic functions, to prove the convergence of certain observables of the model to conformal maps<sup>5</sup>. The second revisits Loewner's deterministic evolutions (which were used to solve the Bieberbach conjecture) to construct random processes whose applications now spread over all probability theory.

**1.3.3 Discrete Fourier analysis, concentration inequalities... and percolation.** The end of the nineties witnessed the appearance of two important new problems regarding Bernoulli percolation. [Häggström, Peres, and Steif \[1997\]](#) introduced a simple time dynamics in Bernoulli percolation obtained by resampling each edge at an exponential rate 1. More precisely, an exponential clock is attached to each edge of the lattice, and each time

---

<sup>5</sup>Nicely enough, the story can again go the other way: Smirnov's argument can also be used to provide an alternative proof of Riemann's mapping theorem [Smirnov \[n.d.\]](#).

the clock of an edge rings, the state of the edge is resampled. It turns out that dynamical percolation is a very interesting model which exhibits quite rich phenomena.

As mentioned above, it is known since Harris that  $\theta(1/2) = 0$  on  $\mathbb{Z}^2$ . Since Bernoulli percolation is the invariant measure for the dynamics, this easily implies that for any fixed time  $t \geq 0$ , the probability that dynamical percolation at time  $t$  does not contain an infinite connected component is zero. Fubini's theorem shows a stronger result: with probability 1, the set of times at which the configuration contains an infinite connected component is of Lebesgue measure 0. Nonetheless, this statement does not exclude the possibility that this set of times is non-empty.

In 1999, [Benjamini, Kalai, and Schramm \[1999\]](#) initiated the study of the Fourier spectrum of percolation and its applications to the noise sensitivity of the model measuring how much connectivity properties of the model are robust under the dynamics. This work advertised the usefulness of concentration inequalities and discrete Fourier analysis for the understanding of percolation: they provide information on the model which is invisible from the historical probabilistic and geometric approaches. We will see below that these tools will be crucial to the developments of percolation theory of dependent models.

We cannot conclude this section without mentioning the impressive body of works by [Garban, Pete, and Schramm \[2011, 2013\]](#) and [Garban, Pete, and Schramm \[2017\]](#) describing in detail the noise sensitivity and dynamical properties of planar percolation. These works combine the finest results on planar Bernoulli percolation and provide a precise description of the behavior of the model, in particular proving that the Hausdorff dimension of the set of times at which there exists an infinite connected component is equal to  $31/36$ .

*At this stage of the review, we hope that the reader gathered some understanding of the problematic about Bernoulli percolation, and got some idea of the variety of fields of mathematics involved in the study of the model. We will now try to motivate the introduction of more complicated percolation models before explaining, in [Section 3](#), how some of the techniques mentioned above can be used to study these more general models. We refer to [G. Grimmett \[1999\]](#) for more references.*

## 2 Beyond Bernoulli percolation

While the theory of Bernoulli percolation still contains a few gems that deserve a solution worthy of their beauty, recent years have revived the interest for more general percolation models appearing in various areas of statistical physics as natural models associated with other random systems. While Bernoulli percolation is a product measure, the states of edges in these percolation models are typically not independent. Let us discuss a few ways of introducing these “dependent” percolation models.

**2.1 From spin models to bond percolation.** Dependent percolation models are often associated with lattice spin models. These models have been introduced as discrete models for real life experiments and have been later on found useful to model a large variety of phenomena and systems ranging from ferromagnetic materials to lattice gas. They also provide discretizations of Euclidean and Quantum Field Theories and are as such important from the point of view of theoretical physics. While the original motivation came from physics, they appeared as extremely complex and rich mathematical objects, whose study required the developments of new tools that found applications in many other domains of mathematics.

The archetypical example of the relation spin model/percolation is provided by the *Potts model* and FK percolation (defined below). In the former, spins are chosen among a set of  $q$  colors (when  $q = 2$ , the model is called the *Ising model*, and spins are seen as  $\pm 1$  variables), and the distribution depends on a parameter  $\beta$  called the *inverse temperature*. We prefer not to define the models too formally here and refer to [Duminil-Copin \[2017\]](#) for more details. We rather focus on the relation between these models and a dependent *bond* percolation model called *Fortuin-Kasteleyn (FK) percolation*, or *random-cluster model*.

FK percolation was introduced by [Fortuin and Kasteleyn \[1972\]](#) as a unification of different models of statistical physics satisfying series/parallel laws when modifying the underlying graph. The probability measure on a finite graph  $\mathbb{G}$ , denoted by  $\mathbb{P}_{\mathbb{G},p,q}$ , depends on two parameters – the *edge-weight*  $p \in [0, 1]$  and the *cluster-weight*  $q > 0$  – and is defined by the formula

$$(1) \quad \mathbb{P}_{\mathbb{G},p,q}[\{\omega\}] := \frac{p^{|\omega|}(1-p)^{|\mathbb{E}|-|\omega|}q^{k(\omega)}}{Z(\mathbb{G},p,q)} \quad \text{for every } \omega \in \{0, 1\}^{\mathbb{E}},$$

where  $|\omega|$  is the number of open edges and  $k(\omega)$  the number of connected components in  $\omega$ . The constant  $Z(\mathbb{G}, p, q)$  is a normalizing constant, referred to as the *partition function*, defined in such a way that the sum over all configurations equals 1. For  $q = 1$ , the model is Bernoulli percolation, but for  $q \neq 1$ , the probability measure is different, due to the term  $q^{k(\omega)}$  taking into account the number of connected components. Note that, at first sight, the definition of the model makes no sense on infinite graphs (contrarily to Bernoulli percolation) since the number of open edges (or infinite connected components) would be infinite. Nonetheless, the model can be extended to infinite graphs by taking the (weak) limit of measures on finite graphs. For more on the model, we refer to the comprehensive reference book [G. Grimmett \[2006\]](#).

As mentioned above, FK percolation is connected to the Ising and the Potts models via what is known as the *Edwards-Sokal coupling*. It is straightforward to describe this coupling in words. Let  $\omega$  be a percolation configuration sampled according to the FK percolation with edge-weight  $p \in [0, 1]$  and cluster-weight  $q \in \{2, 3, 4, \dots\}$ . The random coloring of  $\mathbb{V}$  obtained by assigning to connected components of  $\omega$  a color chosen



uniformly (among  $q$  fixed colors) and independently for each connected component, and then giving to a vertex the color of its connected component, is a realization of the  $q$ -state Potts model at inverse temperature  $\beta = -\frac{1}{2} \log(1 - p)$ . For  $q = 2$ , the colors can be understood as  $-1$  and  $+1$  and one ends up with the Ising model.

The relation between FK percolation and the Potts model is not an exception. Many other lattice spin models also possess their own Edwards-Sokal coupling with a dependent percolation model. This provides us with a whole family of natural dependent percolation models that are particularly interesting to study. We refer the reader to [Chayes and Machta \[1997\]](#) and [Pfister and Velenik \[1997\]](#) for more examples.

**2.2 From loop models to site percolation models.** In two dimensions, there is another recipe to obtain dependent percolation models, but on sites this time. Each site percolation configuration on a *planar graph* is naturally associated, by the so-called *low-temperature expansion*, with a bond percolation of a very special kind, called a *loop model*, on the dual graph  $\mathbb{G}^* = (\mathbb{V}^*, \mathbb{E}^*)$ , where  $\mathbb{V}^*$  is the set of faces of  $\mathbb{G}$ , and  $\mathbb{E}^*$  the set of unordered pairs of adjacent faces. More precisely, if  $\omega$  is an element of  $\{0, 1\}^{\mathbb{V}}$  (which can be seen as an attribution of 0 or 1 to faces of  $\mathbb{G}^*$ ), define the percolation configuration  $\eta \in \{0, 1\}^{\mathbb{E}^*}$  by first extending  $\omega$  to the exterior face  $x$  of  $\mathbb{G}^*$  by arbitrarily setting  $\omega_x = 1$ , and then saying that an edge of  $\mathbb{G}^*$  is open if the two faces  $x$  and  $y$  of  $\mathbb{G}^*$  that it borders satisfy  $\omega_x \neq \omega_y$ . In physics terminology, the configuration  $\eta$  corresponds to the *domain walls* of  $\omega$ . Notice that the degree of  $\eta$  at every vertex is necessarily even, and that therefore  $\eta$  is necessarily an even subgraph.

But one may go the other way: to any even subgraph of  $\mathbb{G}^*$ , one may associate a percolation configuration on  $\mathbb{V}$  by setting  $+1$  for the exterior face of  $\mathbb{G}^*$ , and then attributing 0/1 values to the other faces of  $\mathbb{G}^*$  by switching between 0 and 1 when crossing an edge. This reverse procedure provides us with a recipe to construct new dependent site percolation models: construct first a loop model, and then look at the percolation model it creates.

When starting with the Ising model on the triangular lattice (which is indeed a site percolation model: a vertex is open if the spin is  $+1$ , and closed if it is  $-1$ ), the low-temperature expansion gives rise to a model of random loops on the hexagonal lattice, for which the weight of an even subgraph  $\eta$  is proportional to  $\exp(-2\beta|\eta|)$ . This loop model was generalized by [Domany, Mukamel, Nienhuis, and Schwimmer \[1981\]](#) to give the *loop  $O(n)$  model* depending on two parameters, an *edge-weight*  $x > 0$  and a *loop-weight*  $n \geq 0$ . It is defined on the hexagonal lattice  $\mathbb{H} = (\mathbb{V}, \mathbb{E})$  as follows: the probability of  $\eta$  on  $\mathbb{H}$  is given by

$$\mu_{\mathbb{G}, x, n}[\{\eta\}] = \frac{x^{|\eta|} n^{\ell(\eta)}}{Z(\mathbb{G}, x, n)}$$

(where  $\ell(\eta)$  is the number of loops in  $\eta$  if  $\eta$  is an even subgraph, and 0 otherwise).

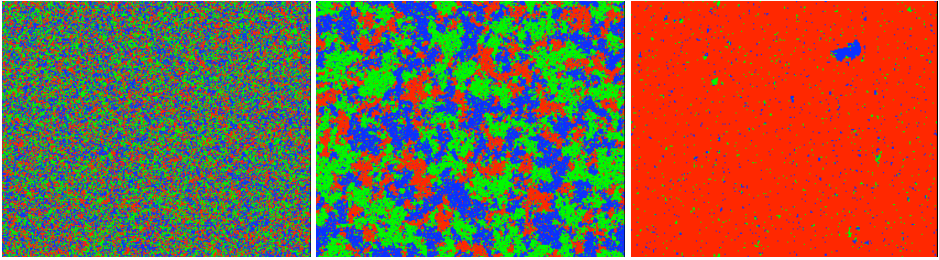


Figure 3: Simulations by Vincent Beffara of the three-state planar Potts model obtained from the FK percolation with parameter  $p < p_c$ ,  $p = p_c$  and  $p > p_c$ . On the right, every vertex of the infinite connected component receives the same color, therefore one of the colors wins over the other ones, while this is not the case at criticality or below it.

From this loop model, one obtains a site percolation model on the triangular lattice resembling FK percolation. We will call this model the *FK representation of the dilute Potts model* (for  $n = 1$ , it is simply the Ising model mentioned above).

*We hope that this section underlined the relevance of some dependent percolation models, and that the previous one motivated questions for Bernoulli percolation that possess natural counterparts for these dependent percolation models. The next sections describe the developments and solutions to these questions.*

### 3 Exponential decay of correlations in the subcritical regime

As mentioned in the first section, proving exponential decay of  $\theta_n(p)$  when  $p < p_c$  was a milestone in the theory of Bernoulli percolation since it was the key to a deep understanding of the subcritical regime. The goal of this section is to discuss the natural generalizations of these statements to FK percolation with cluster-weight  $q > 1$ . Below, we set  $\theta_n(p, q)$  and  $\theta(p, q)$  for the probabilities of being connected to distance  $n$  and to infinity respectively. Also, we define

$$p_c(q) \stackrel{\text{def}}{=} \inf\{p \in [0, 1] : \theta(p, q) > 0\},$$

$$p_{\text{exp}}(q) \stackrel{\text{def}}{=} \sup\{p \in [0, 1] : \exists c > 0, \forall n \geq 0, \theta_n(p, q) \leq \exp(-cn)\}.$$

Exponential decay in the subcritical regime gets rephrased as  $p_c(q) = p_{\text{exp}}(q)$ . Exactly like in the case of Bernoulli percolation, the result was first proved in two dimensions, and then in higher dimensions, so that we start by the former. Interestingly, both proofs (the



probabilities of crossing rectangles in the hard direction are not going to 1, then the same holds for squares. At the light of the previous paragraph, what we really want to exclude now is the existence of a whole interval of values of  $p$  for which the probabilities of crossing squares remain bounded away from 0 and 1 uniformly in  $n$ .

In order to exclude this possibility, we invoke a second ingredient, which is of a very different nature. It consists in proving that probabilities of crossing squares go quickly from a value close to 0 to a value close to 1. Kesten originally proved this result by hand by showing that the derivative of the function  $p \mapsto \mathbf{H}_{n,n}(p)$  satisfies a differential inequality of the form

$$(2) \quad \mathbf{H}'_{n,n} \geq c \log n \cdot \mathbf{H}_{n,n}(1 - \mathbf{H}_{n,n}).$$

This differential inequality immediately shows that the plot of the function  $\mathbf{H}_{n,n}$  has an  $S$  shape as on [Figure 4](#), and that  $\mathbf{H}_{n,n}(p)$  therefore goes from  $\varepsilon$  to  $1 - \varepsilon$  in an interval of  $p$  of order  $O(1/\log n)$ . In particular, it implies that only one value of  $p$  can be such that  $\mathbf{H}_{n,n}(p)$  remains bounded away from 0 and 1 uniformly in  $n$ , hence concluding the proof.

In [Bollobás and Riordan \[2006b\]](#) proposed an alternative strategy to prove (2). They suggested using a concept long known to combinatorics: a finite random system undergoes a *sharp threshold* if its qualitative behavior changes quickly as the result of a small perturbation of the parameters ruling the probabilistic structure. The notion of sharp threshold emerged in the combinatorics community studying graph properties of random graphs, in particular in the work of [Erdős and Rényi \[1959\]](#) investigating the properties of Bernoulli percolation on the complete graph.

Historically, the general theory of sharp thresholds for discrete product spaces was developed by [Kahn, Kalai, and Linial \[1988\]](#) in the case of the uniform measure on  $\{0, 1\}^n$ , i.e. in the case of  $\mathbb{P}_p$  with  $p = 1/2$  (see also an earlier non-quantitative version by [Russo \[1982\]](#)). There, the authors used the Bonami-Beckner inequality [Beckner \[1975\]](#) and [Bonami \[1970\]](#) together with discrete Fourier analysis to deduce inequalities between the variance of a Boolean function and the covariances (often called *influences*) of this function with the random variables  $\omega(e)$ . [Bourgain, Kahn, Kalai, Katznelson, and Linial \[1992\]](#) extended these inequalities to product spaces  $[0, 1]^n$  endowed with the uniform measure (see also [Talagrand \[1994\]](#)), a fact which enables to cover the case of  $\mathbb{P}_p$  for every value of  $p \in [0, 1]$ .

Roughly speaking, the statement can be read as follows: for any increasing<sup>6</sup> (boolean) function  $\mathbf{f} : \{0, 1\}^{\mathbb{E}} \rightarrow \{0, 1\}$ ,

$$(3) \quad \text{Var}_p(\mathbf{f}) \leq c(p) \sum_{e \in \mathbb{E}} \frac{\text{Cov}_p[\mathbf{f}, \omega(e)]}{\log(1/\text{Cov}_p[\mathbf{f}, \omega(e)])},$$

<sup>6</sup>Here, increasing is meant with respect to the partial order on  $\{0, 1\}^{\mathbb{E}}$  defined by  $\omega \leq \omega'$  if  $\omega(e) \leq \omega'(e)$  for every edge  $e \in \mathbb{E}$ . Then,  $\mathbf{f}$  is *increasing* if  $\omega \leq \omega'$  implies  $\mathbf{f}(\omega) \leq \mathbf{f}(\omega')$ .

where  $c$  is an explicit function of  $p$  that remains bounded away from 0 and 1 when  $p$  is away from 0 and 1.

Together with the following differential formula (which can be obtained by simply differentiating  $\mathbb{E}_p[\mathbf{f}]$ )

$$(4) \quad \frac{d}{dp} \mathbb{E}_p[\mathbf{f}] = \frac{1}{p(1-p)} \sum_{e \in \mathbb{E}} \text{Cov}_p[\mathbf{f}, \omega(e)]$$

for the indicator function  $\mathbf{f}$  of the event that the square  $[0, n]^2$  is crossed horizontally, we deduce that

$$(5) \quad \mathbf{H}'_{n,n} \geq \frac{4}{c(p) \log(1/\max_e \text{Cov}_p[\mathbf{f}_n, \omega(e)])} \cdot \mathbf{H}_{n,n}(1 - \mathbf{H}_{n,n}).$$

This inequality can be used as follows. The covariance between the existence of an open path and an edge  $\omega(e)$  can easily be bounded by the fact that one of the two endpoints of the edge  $e$  is connected to distance  $n/2$  (indeed, for  $\omega(e)$  to influence the outcome of  $\mathbf{f}_n$ , there must be an open crossing going through  $e$  when  $e$  is open). But, in the regime where crossing probabilities are bounded away from 1, the probability of being connected to distance  $n/2$  can easily be proved to decay polynomially fast, so that in fact  $\mathbf{H}'_{n,n} \geq c \log n \cdot \mathbf{H}_{n,n}(1 - \mathbf{H}_{n,n})$  as required.

**FK percolation.** What survives for dependent percolation models such as FK percolation? The good news is that the BKKKL result can be extended to this context [Graham and G. R. Grimmett \[2006\]](#) to obtain (3). Equation (4) is obtained in the same way by elementary differentiation. It is therefore the RSW result which is tricky to extend.

While mathematicians are in possession of many proofs of this theorem for Bernoulli percolation [Bollobás and Riordan \[2006b,a, 2010\]](#), [Russo \[1978\]](#), [Seymour and Welsh \[1978\]](#), and [Tassion \[2014, 2016\]](#), one had to wait for thirty years to obtain the first proof of this theorem for dependent percolation model.

The following result is the most advanced result in this direction. Let  $\mathbf{H}_{n,m}(p, q)$  be the probability that  $[0, n] \times [0, m]$  is crossed horizontally for FK percolation.

**Theorem 3.1.** *For any  $\rho > 0$ , there exists a constant  $C = C(\rho, \varepsilon, q) > 0$  such that for every  $p \in [\varepsilon, 1 - \varepsilon]$  and  $n \geq 1$ ,*

$$\mathbf{H}_{n,n}(p, q)^C \leq \mathbf{H}_{\rho n, n}(p, q) \leq 1 - (1 - \mathbf{H}_{n,n}(p, q))^C.$$

A consequence of all this is the following result [Beffara and Duminil-Copin \[2012\]](#) (see also [Duminil-Copin and Manolescu \[2016\]](#) and [Duminil-Copin, Raoufi, and Tassion \[2018\]](#)).

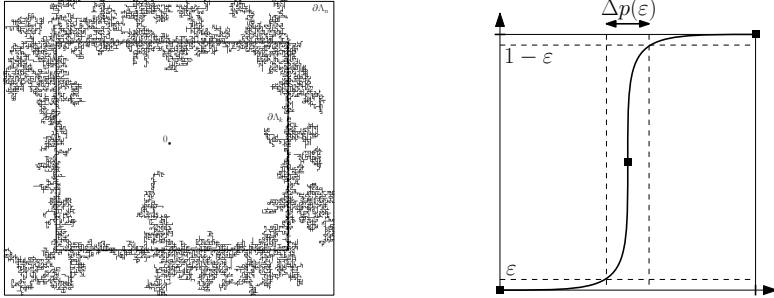


Figure 4: **Left.** The randomized algorithm is obtained as follows: pick a distance  $k$  to the origin uniformly and then explore “from the inside” all the connected components intersecting the boundary of the box of size  $k$ . If one of these connected components intersect both  $0$  and the boundary of the box of size  $n$ , then we know that  $0$  is connected to distance  $n$ , if none of them do, then the converse is true. **Right.** The  $S$  shape of the function  $p \mapsto \mathbf{H}_{n+1,n}(p)$ . The function goes very quickly from  $\varepsilon$  to  $1 - \varepsilon$  ( $\Delta p(\varepsilon)$  is small).

**Theorem 3.2.** *Consider the FK model with cluster weight  $q \geq 1$ . Then, for any  $p < p_c$ , there exists  $c = c(p, q) > 0$  such that for every  $n \geq 1$ ,*

$$\theta_n(p, q) \leq \exp(-cn).$$

### 3.2 Higher dimensions.

**Bernoulli percolation.** Again, let us start with the discussion of this simpler case. When working in higher dimensions, one can still consider crossing probabilities of boxes but one is soon facing some substantial challenges. Of course, some of the arguments of the previous section survive. For instance, one can adapt the two-dimensional proof to show that if the box  $[0, n] \times [0, 2n]^{d-1}$  is crossed from left to right with probability smaller than some constant  $\varepsilon = \varepsilon(d) > 0$ , then  $\theta_n(p)$  decays exponentially fast. One can also prove the differential inequality (5) without much trouble. But that is basically it. One cannot prove that, if the probability of the box  $[0, 2n] \times [0, n]^{d-1}$  is crossed from left to right with probability close to one, then  $\theta(p) > 0$ . Summarizing, we cannot (yet) exclude a regime of values of  $p$  for which crossing probabilities tend to 1 but the probability that there exists an infinite connected component is zero<sup>7</sup>.

<sup>7</sup>It is in fact the case that for  $p = p_c$  and  $d \geq 6$ , crossing probabilities tend to 1 but  $\theta(p_c) = 0$ . What we wish to exclude is a whole range of parameters for which this happens.

We are therefore pushed to abandon crossing probabilities and try to work directly with  $\theta_n$ . Applying the BKKKL result to  $\theta_n$  implies, when  $p < p_c$ , an inequality (basically) stating

$$\theta'_n \geq c \log n \cdot \theta_n(1 - \theta_n).$$

This differential inequality is unfortunately not useful to exclude a regime where  $\theta_n$  would decay polynomially fast. For this reason, we need to strengthen it. In order to do this, we will not rely on a concentration inequality coming from discrete Fourier analysis like in the two-dimensional case, but rather on another concentration-type inequality used in computer science.

Informally speaking, a *randomized algorithm* associated with a boolean function  $\mathbf{f}$  takes  $\omega \in \{0, 1\}^{\mathbb{E}}$  as an input, and reveals algorithmically the value of  $\omega$  at different edges one by one until the value of  $\mathbf{f}(\omega)$  is determined. At each step, which edge will be revealed next depends on the values of  $\omega$  revealed so far. The algorithm stops as soon as the value of  $\mathbf{f}$  is the same no matter the values of  $\omega$  on the remaining coordinates.

The OSSS inequality, originally introduced by O'Donnell, Saks, Schramm, and Servedio [2005] as a step toward a conjecture of Yao [1977], relates the variance of a boolean function to the influence and the computational complexity of a randomized algorithm for this function. More precisely, consider  $p \in [0, 1]$  and  $n \in \mathbb{N}$ . Fix an increasing boolean function  $\mathbf{f} : \{0, 1\}^{\mathbb{E}} \rightarrow \{0, 1\}$  and an algorithm  $\mathbf{T}$  for  $\mathbf{f}$ . We have

$$(\text{OSSS}) \quad \text{Var}_p(\mathbf{f}) \leq 2 \sum_{e \in \mathbb{E}} \delta_e(\mathbf{T}) \text{Cov}_p[\mathbf{f}, \omega(e)],$$

where  $\delta_e(\mathbf{T})$  is the probability that the edge  $e$  is revealed by the algorithm before it stops. One will note the similarity with (3), where the term  $\delta_e(\mathbf{T})$  replaces  $-1$  divided by the logarithm of the covariance.

The interest of (OSSS) comes from the fact that, if there exists a randomized algorithm for  $\mathbf{f} = \mathbb{1}[0 \text{ connected to distance } n]$  for which each edge has small probability of being revealed, then the inequality implies that the derivative of  $\mathbb{E}_p[\mathbf{f}]$  is much larger than the variance  $\theta_n(1 - \theta_n)$  of  $\mathbf{f}$ . Of course, there are several possible choices for the algorithm. Using the one described in Figure 4, one deduces that the probability of being revealed is bounded by  $cS_n/n$  uniformly for every edge, where  $S_n := \sum_{k=0}^{n-1} \theta_k$ . We therefore deduce an inequality of the form

$$(6) \quad \theta'_n \geq c' \frac{n}{S_n} \theta_n(1 - \theta_n).$$

Note that the quantity  $n/S_n(p)$  is large when the values  $\theta_k(p)$  are small, which is typically the case when  $p < p_c$ . In particular, one can use this differential inequality to prove the sharpness of the phase transition on any transitive graph. Equation (6) already appeared in Menshikov's 1986 proof while Aizenman and Barsky [1987] and later Duminil-Copin and Tassion [2016] invoked alternative differential inequalities.

**FK percolation.** As mentioned in the previous paragraph, the use of differential inequalities to prove sharpness of the phase transition is not new, and even the differential inequality (6) chosen above already appeared in the literature. Nonetheless, the existing proofs of these differential inequalities all had one feature in common: they relied on a special correlation inequality for Bernoulli percolation known as the BK inequality, which is not satisfied for most dependent percolation models, so that the historical proofs did not extend easily to FK percolation, contrarily to the approach using the OSSS inequality proposed in the previous section.

Indeed, while the OSSS inequality uses independence, it does not rely on it in a substantial way. In particular, the OSSS inequality can be extended to FK percolation, very much like [Graham and G. R. Grimmett \[2006\]](#) generalized (4). This generalization enables one to show (6) for a large class of models including dependent percolation models or so-called continuum percolation models [Duminil-Copin, Raoufi, and Tassion \[2017b\]](#) and [Duminil-Copin, Raoufi, and Tassion \[2017a\]](#). In particular,

**Theorem 3.3** ([Duminil-Copin, Raoufi, and Tassion \[2017c\]](#)). *Fix  $q \geq 1$  and  $d \geq 2$ . Consider FK percolation on  $\mathbb{Z}^d$  with cluster-weight  $q \geq 1$ . For any  $p < p_c$ , there exists  $c = c(p, q) > 0$  such that for every  $n \geq 1$ ,*

$$\theta_n(p, q) \leq \exp(-cn).$$

Exactly as for Bernoulli percolation, one can prove many things using this theorem. Of special interest are the consequences for the Potts model (and its special case the Ising model): the exponential decay of  $\theta_n(p, q)$  implies the exponential decay of correlations in the disordered phase.

*The story of the proof of exponential decay of  $\theta_n(p, q)$  is typical of percolation. Some proofs first appeared for Bernoulli percolation. These proofs were then made more robust using some external tools, here discrete analysis (the BKKKL concentration inequality or the OSSS inequality), and finally extended to more general percolation models. The next section provides another example of such a succession of events.*

## 4 Computation of critical points in two dimensions

It is often convenient to have an explicit formula for the critical point of a model. In general, one cannot really hope for such a formula but in some cases, one is saved by specific properties of the model, which can be of (at least) two kinds: *self-duality* or *exact integrability*.

### 4.1 Computation of the critical point using self-duality.



**Bernoulli percolation.** One can easily guess why the critical point of Bernoulli percolation on  $\mathbb{Z}^2$  should be equal to  $1/2$ . Indeed, every configuration  $\omega$  is naturally associated with a dual configuration  $\omega^*$  defined on the dual lattice  $(\mathbb{Z}^2)^* = (\frac{1}{2}, \frac{1}{2}) + \mathbb{Z}^2$  of  $\mathbb{Z}^2$ : for every edge  $e$ , set

$$\omega^*(e^*) \stackrel{\text{def}}{=} 1 - \omega(e),$$

where  $e^*$  is the unique edge of the dual lattice crossing the edge  $e$  in its middle. In words, a dual edge is open if the corresponding edge of the primal lattice is closed, and vice versa. If  $\omega$  is sampled according to Bernoulli percolation of parameter  $p$ , then  $\omega^*$  is sampled according to a Bernoulli percolation on  $(\mathbb{Z}^2)^*$  of parameter  $p^* := 1 - p$ . The value  $1/2$  therefore emerges as the self-dual value for which  $p = p^*$ .

It is not a priori clear why the self-dual value should be the critical one, but armed with the theorems of the previous section, we can turn this observation into a rigorous proof. Indeed, one may check (see [Figure 5](#)) that for every  $n \geq 1$ ,

$$\mathbf{H}_{n+1,n}(\tfrac{1}{2}) = \tfrac{1}{2}.$$

Yet, an outcome of [Section 3.1](#) is that crossing probabilities are tending to 0 when  $p < p_c$  and to 1 when  $p > p_c$ . As a consequence, the only possible value for  $p_c$  is  $1/2$ .

**FK percolation.** The duality relation generalizes to cluster-weights  $q \neq 1$ : if  $\omega$  is sampled according to a FK percolation measure with parameters  $p$  and  $q$ , then  $\omega^*$  is sampled according to a FK percolation measure with parameters  $p^*$  and  $q^*$  satisfying

$$\frac{pp^*}{(1-p)(1-p^*)} = q \quad \text{and} \quad q^* = q.$$

The proof of this fact involves Euler's relation for planar graphs. Let us remark that readers trying to obtain such a statement as an exercise will encounter a small difficulty due to boundary effects on  $\mathbb{G}$ ; we refer to [Duminil-Copin \[2017\]](#) for details how to handle such *boundary conditions*. The formulas above imply that for every  $q \neq 0$ , there exists a unique point  $p_{\text{sd}}(q)$  such that

$$p_{\text{sd}}(q) = p_{\text{sd}}(q)^* = \frac{\sqrt{q}}{1 + \sqrt{q}}.$$

Exactly as in the case of Bernoulli percolation, one may deduce from self-duality some estimates on crossing probabilities at  $p = p_{\text{sd}}(q)$ , which imply in the very same way the following theorem.

**Theorem 4.1** ([Beffara and Duminil-Copin \[2012\]](#)). *The critical point of FK percolation on the square lattice with cluster-weight  $q \geq 1$  is equal to the self-dual point  $\sqrt{q}/(1 + \sqrt{q})$ .*

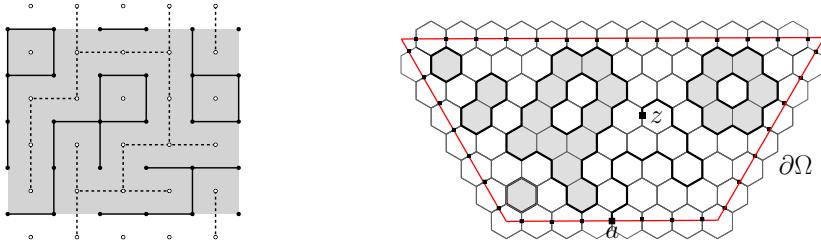


Figure 5: **Left.** If  $[0, n + 1] \times [0, n]$  is not crossed from left to right, then the boundary of the connected components touching the right side is a dual path going from top to bottom. **Right.** The domain  $\Omega$  with its boundary  $\partial\Omega$ . The configuration  $\omega$  corresponds to the interfaces between the gray and the white hexagons. The path  $\gamma$  runs from  $a$  to  $z$  without intersecting  $\omega$ .

**4.2 Computation via parafermionic observables.** Sometimes, no obvious self-duality relation helps us identify the critical point, but one can be saved by a second strategy. In order to illustrate it, consider the loop  $O(n)$  model with parameters  $x > 0$  and  $n \in [0, 2]$  and its associated FK representation described in [Section 2.2](#). Rather than referring to duality (in this case, none is available as for today), the idea consists in introducing a function that satisfies some specific integrability/local relations at a given value of the parameter.

Take a self-avoiding polygon on the dual (triangular) lattice of the hexagonal lattice; see [Figure 5](#). By definition, this polygon divides the hexagonal lattice into two connected components, a bounded one and an unbounded one. Call the bounded one  $\Omega$  and, by analogy with the continuum, denote the self-avoiding polygon by  $\partial\Omega$ .

Define the *parafermionic observable* introduced in [Smirnov \[2010b\]](#), as follows (see [Figure 5](#)): for a mid-edge  $z$  in  $\Omega$  and a mid-edge  $a$  in  $\partial\Omega$ , set

$$F(z) = F(\Omega, a, z, n, x, \sigma) \stackrel{\text{def}}{=} \sum_{\substack{\omega, \gamma \subset \Omega \\ \omega \cap \gamma = \emptyset}} e^{-i\sigma \mathbf{W}_\gamma(a, z)} x^{|\gamma| + |\omega|} n^{\ell(\omega)}$$

(recall that  $\ell(\omega)$  is the number of loops in  $\omega$ ), where the sum is over pairs  $(\omega, \gamma)$  with  $\omega$  a loop configuration, and  $\gamma$  a self-avoiding path from  $a$  to  $z$ . The quantity  $\mathbf{W}_\gamma(a, z)$ , called the *winding term*, is equal to  $\frac{\pi}{3}$  times the number of left turns minus the number of right turns made by the walk  $\gamma$  when going from  $a$  to  $z$ . It corresponds to the total rotation of the oriented path  $\gamma$ .

The interest of  $F$  lies in a special property satisfied when the parameters of the model are tuned properly. More precisely, if  $\sigma = \sigma(n)$  is well chosen (the formula is explicit but

irrelevant here) and

$$x = x_c(n) \stackrel{\text{def}}{=} \frac{1}{\sqrt{2 + \sqrt{2 - n}}},$$

the function satisfies that for any self-avoiding polygon  $\mathbf{C} = (c_0, c_1, \dots, c_k = c_0)$  on  $\mathbb{T}$  that remains within the bounded region delimited by  $\partial\Omega$ ,

$$(7) \quad \oint_{\mathbf{C}} F(z) dz \stackrel{\text{def}}{=} \sum_{i=1}^k (c_i - c_{i-1}) F\left(\frac{c_{i-1} + c_i}{2}\right) = 0.$$

Above, the quantity  $c_i$  is considered as a complex number, in such a way that the previous definition matches the intuitive notion of contour integral for functions defined on the middle of the edges of the hexagonal lattice.

In words, (7) means that for special values of  $x$  and  $\sigma$ , discrete contour integrals of the parafermionic observable vanish. In the light of Morera's theorem, this property is a glimpse of conformal invariance of the model in the sense that the observable satisfies a weak notion of discrete holomorphicity. This singles out  $x_c(n)$  as a very peculiar value of the parameter  $x$ .

The existence of such a discrete holomorphic observable at  $x_c(n)$  did not really come as a surprise. In the case of the loop  $O(n)$  model with loop-weight  $n \in [0, 2]$ , the physicist [Nienhuis \[1982\]](#) predicted that  $x_c(n)$  is a critical value for the loop model, in the following sense:

- when  $x < x_c(n)$ , loops are typically small: the probability that the loop of the origin is of length  $k$  decays exponentially fast in  $k$ .
- when  $x \geq x_c(n)$ , loops are typically large: the probability that the loop of the origin is of length  $k$  decays slower than polynomially fast in  $k$ .

Physically, this criticality of  $x_c(n)$  has an important consequence. As (briefly) mentioned in [Section 1.3.2](#), Bernoulli percolation and more generally two-dimensional models at criticality are predicted to be conformally invariant. This prediction has a concrete implication on critical models: some observables<sup>8</sup> should converge in the scaling limit to conformally invariant/covariant objects. In the continuum, typical examples of such objects are provided by harmonic and holomorphic solutions to Boundary Value Problems; it is thus natural to expect that some observables of the loop  $O(n)$  model at criticality are discrete holomorphic.

**Remark 4.2.** *Other than being interesting in themselves, discrete holomorphic functions have found several applications in geometry, analysis, combinatorics, and probability.*

<sup>8</sup>Roughly speaking, observables are averages of random quantities defined locally in terms of the system.

*The use of discrete holomorphicity in statistical physics appeared first in the case of dimers Kenyon [2000] and has since then been extended to several statistical physics models, including the Ising model Chelkak and Smirnov [2011, 2012] (see also Duminil-Copin and Smirnov [2012a] for references).*

The definition of discrete holomorphicity usually imposes stronger conditions on the function than just zero contour integrals (for instance, one often asks for suitable discretizations of the Cauchy-Riemann equations). In particular, in our case the zero contour integrals do not uniquely determine  $F$ . Indeed, there is one variable  $F(z)$  by edge, but the number of independent linear relations provided by the zero contour integrals is equal to the number of vertices<sup>9</sup>. In conclusion, there are much fewer relations than unknown and it is completely unclear whether one can extract any relevant information from (7).

Anyway, one can still try to harvest this “partial” information to identify rigorously the critical value of the loop  $O(n)$  model. For  $n = 0$  (in this case there is no loop configuration and just one self-avoiding path), the parafermionic observable was used to show that the *connective constant* of the hexagonal lattice Duminil-Copin and Smirnov [2012b], i.e.

$$\mu_c \stackrel{\text{def}}{=} \lim_{n \rightarrow \infty} \#\{\text{self-avoiding walks of length } n \text{ starting at the origin}\}^{1/n}$$

is equal to  $\sqrt{2 + \sqrt{2}}$ . For  $n \in [1, 2]$ , the same observable was used to show that at  $x = x_c(n)$ , the probability of having a loop of length  $k$  decays slower than polynomially fast Duminil-Copin, Glazman, Peled, and Spinka [2017], thus proving part of Nienhuis prediction (this work has also applications for the corresponding site percolation model described in Section 2.2).

Let us conclude this section by mentioning that the parafermionic observable defined for the loop  $O(n)$  model can also be defined for a wide variety of models of planar statistical physics; see e.g. Ikhlef and J. Cardy [2009], Ikhlef, Weston, Wheeler, and Zinn-Justin [2013], and Rajabpour and J. Cardy [2007]. This leaves hope that many more models from planar statistical physics can be studied using discrete holomorphic functions. For the FK percolation of parameter  $q \geq 4$ , a parafermionic observable was used to show that  $p_c(q) = \sqrt{q}/(+\sqrt{q})$ , thus providing an alternative proof to Beffara and Duminil-Copin [2012] (this proof was in fact obtained prior to the proof of Beffara and Duminil-Copin [ibid.]). Recently, the argument was generalized to the case  $q \in [1, 4]$  in *Computation of the critical point for random-cluster models via the parafermionic observable* [n.d.].

---

<sup>9</sup>Indeed, it is sufficient to know that discrete contour integrals vanish for a basis of the  $\mathbb{Z}$ -module of cycles (which are exactly the contours) on the triangular lattice staying in  $\Omega$  to obtain all the relations in (7). A natural choice for such a basis is provided by the triangular cycles around each face of the triangular lattice inside  $\partial\Omega$ , hence it has exactly as many elements as vertices of the hexagonal lattice in  $\Omega$ .

*In the last two sections, we explained how the study of crossing probabilities can be combined with duality or parafermionic observables to identify the critical value of some percolation models. In the next one, we go further and discuss how the same tools can be used to decide whether the phase transition is continuous or discontinuous (see definition below).*

## 5 The critical behavior of planar dependent percolation models

**5.1 Renormalization of crossing probabilities.** For Bernoulli percolation, we mentioned that crossing probabilities remain bounded away from 0 and 1 uniformly in the size of the rectangle (provided the aspect ratio stays bounded away from 0 or 1). For more complicated percolation models, the question is more delicate, in particular due to the long-range dependencies. For instance, it may be that crossing probabilities tend to zero when conditioning on edges outside the box to be closed, and to 1 if these edges are conditioned to be open. To circumvent this problem, we introduce a new property.

Consider a percolation measure  $\mathbb{P}$  (one can think of a FK percolation measure for instance). Let  $\Lambda_n$  be the box of size  $n$  around the origin. We say that  $\mathbb{P}$  satisfies the (*polynomial*) *mixing property* if

**(Mix)** *there exist  $c, C \in (0, \infty)$  such that for every  $N \geq 2n$  and every events  $A$  and  $B$  depending on edges in  $\Lambda_n$  and outside  $\Lambda_N$  respectively, we have that*

$$|\mathbb{P}[A \cap B] - \mathbb{P}[A]\mathbb{P}[B]| \leq C \left(\frac{n}{N}\right)^c \cdot \mathbb{P}[A]\mathbb{P}[B].$$

This property has many implications for the study of the percolation model, mostly because it enables one to decorrelate events happening in different parts of the space.

It is a priori unclear how one may prove the mixing property in general. Nonetheless, for critical FK percolation, it can be shown that (wMix) is equivalent to the *strong box crossing property*: uniformly on the states of edges outside of  $\Lambda_{2n}$ , crossing a rectangle of aspect ratio  $\rho$  included in  $\Lambda_n$  remains bounded away from 0 and 1 uniformly in  $n$ . Note that the difference with the previous sections comes from the fact that we consider crossing probabilities conditioned on the state of edges at distance  $n$  of the rectangle (of course, when considering Bernoulli percolation, this does not change anything, but this is not the case anymore when  $q > 1$ ).

The mixing property is not always satisfied at criticality. Nevertheless, in [Duminil-Copin, Sidoravicius, and Tassion \[2017\]](#) the following dichotomy result was obtained.

**Theorem 5.1** (The continuous/discontinuous dichotomy). *For any  $q \geq 1$ ,*

- *either (wMix) is satisfied. In such case:*

- $\theta(p, q)$  tends to 0 as  $p \searrow p_c$ ;
  - There exists  $c > 0$  such that  $cn^{-1} \leq \theta_n(p_c, q) \leq n^{-c}$ .
  - Crossing probabilities of a rectangle of size roughly  $n$  remain bounded away from 0 and 1 uniformly in the state of edges at distance  $n$  of the rectangle;
  - The rate of exponential decay of  $\theta_n(p, q)$  goes to 0 as  $p \nearrow p_c$ .
- or (wMix) is not satisfied and in such case:
    - $\theta(p, q)$  does not tend to 0 as  $p \searrow p_c$ ;
    - There exists  $c > 0$  such that for every  $n \geq 1$ ,  $\theta_n(p_c, q) \leq \exp(-cn)$ ;
    - Crossing probabilities of a rectangle of size roughly  $n$  tend to 0 (resp. 1) when conditioned on the state of edges at distance  $n$  of the rectangle to be closed (resp. open);
    - The rate of exponential decay of  $\theta_n(p, q)$  does not go to 0 as  $p \nearrow p_c$ .

In the first case, we say that the phase transition is *continuous* in reference to the fact that  $p \mapsto \theta(p, q)$  is continuous at  $p_c$ . In the second case, we say that the phase transition is *discontinuous*. Interestingly, this result also shows that a number of (potentially different) definitions of continuous/discontinuous phase transitions sometimes used in physics are in fact the same one in the case of FK percolation.

The proof of the dichotomy is based on a renormalization scheme for crossing probabilities when conditioned on edges outside a box to be closed. Explaining the strategy would lead us too far, and we refer to [Duminil-Copin and Tassion \[n.d.\]](#) and [Duminil-Copin, Sidoravicius, and Tassion \[2017\]](#) for more details. Let us simply add that the proof is not specific to FK percolation and has been extended to other percolation models (see for instance [Duminil-Copin, Glazman, Peled, and Spinka \[2017\]](#) and [Duminil-Copin and Tassion \[n.d.\]](#)), so one should not think of this result as an isolated property of FK percolation, but rather as a general feature of two-dimensional dependent percolation models.

**5.2 Deciding the dichotomy.** As mentioned above, critical planar dependent percolation models can exhibit two different types of critical behavior: continuous or discontinuous. In order to decide which one of the two it is, one needs to work a little bit harder. Let us (briefly) describe two possible strategies. We restrict to the case of FK percolation, for which [Baxter \[1973\]](#) conjectured that for  $q \leq q_c(2) = 4$ , the phase transition is continuous, and for  $q > q_c(2)$ , the phase transition is discontinuous; see [Figure 6](#).

To prove that the phase transition is discontinuous for  $q > 4$ , we used a method going back to early works on the six-vertex model [Duminil-Copin, Gagnebin, Harel, Manolescu, and Tassion \[2016\]](#). The six-vertex model was initially proposed by Pauling in 1931 in order to study the thermodynamic properties of ice. While we are mainly interested in its connection to the previously discussed models, the six-vertex model is a major object of

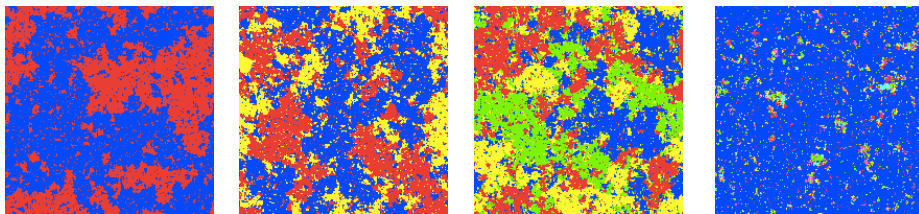


Figure 6: Simulations, due to Vincent Beffara, of the critical planar Potts model with  $q$  equal to 2, 3, 4, and 9 respectively. The behavior for  $q \leq 4$  is clearly different from the behavior for  $q = 9$ . In the first three pictures, each color seems to play the same role, while in the last three, one color wins over the other ones. This is coherent since the phase transition of the associated FK percolation is continuous for  $q \leq 4$  and discontinuous for  $q \geq 5$ .

study on its own right; we refer to [Reshetikhin \[2010\]](#) and Chapter 8 of [Baxter \[1989\]](#) (and references therein) for the definition and a bibliography on the subject.

The utility of the six-vertex model stems from its solvability using the transfer-matrix formalism. More precisely, the partition function of a six-vertex model on a torus of size  $N$  times  $M$  may be expressed as the trace of the  $M$ -th power of a matrix  $V$  (depending on  $N$ ) called the *transfer matrix*. This property can be used to rigorously compute the Perron-Frobenius eigenvalues of the diagonal blocks of the transfer matrix, whose ratios are then related to the rate of exponential decay  $\tau(q)$  of  $\theta_n(p_c, q)$ . The explicit formula obtained for  $\tau(q)$  is then proved to be strictly positive for  $q > 4$ . We should mention that this strategy is extensively used in the physics literature, in particular in the fundamental works of Baxter (again, we refer to [Baxter \[ibid.\]](#)).

In order to prove that the phase transition is continuous for  $q \leq 4$ , one may use the same strategy and try to prove that  $\tau(q)$  is equal to 0. Nevertheless, this does not seem so simple to do rigorously, so that we prefer an alternative approach. The fact that discrete contour integrals of the parafermionic observable vanish can be used for more than just identifying the critical point. For  $q \in [1, 4]$ , it in fact implies lower bounds on  $\theta_n(p_c, q)$ . These lower bounds decay at most polynomially fast, thus guaranteeing that the phase transition is continuous thanks to the dichotomy result of the previous section. This strategy was implemented in [Duminil-Copin, Sidoravicius, and Tassion \[2017\]](#) to complete the proof of Baxter's prediction regarding the continuity/discontinuity of the phase transition for the planar FK percolation with  $q \geq 1$ .

Let us conclude this short review by mentioning that for the special value of  $q = 2$ , the parafermionic observable satisfies stronger constraints. This was used to show that, for this value of  $q$ , the observable is conformally covariant in the scaling limit [Smirnov](#)

[2010a] (this paper by Smirnov had a resounding impact on our understanding of FK percolation with  $q = 2$ ), and that the strong RSW property is satisfied [Duminil-Copin, Hongler, and Nolin \[2011\]](#). This could be the object of a review on its own, especially regarding the conjectures generalizing these results to other values of  $q$ , but we reached the end of our allowed space. We refer to [Duminil-Copin and Smirnov \[2012a\]](#) for more details and relevant references.

**Remark 5.2.** *The strategy described above is very two-dimensional in nature since it relies on planarity in several occasions (crossing probabilities for the dichotomy result, parafermionic observables or transfer matrix formalism for deciding between continuity or discontinuity). In higher dimensions, the situation is more challenging. We have seen that even for Bernoulli percolation, continuity of  $\theta(p)$  had not yet been proved for dimensions  $3 \leq d \leq 10$ . Let us mention that for FK percolation, several results are nevertheless known. One can prove the continuity of  $\theta(p, 2)$  [Aizenman, Duminil-Copin, and Sidoravicius \[2015\]](#) using properties specific to the Ising model (which is associated with the FK percolation with cluster-weight  $q = 2$  via the Edwards-Sokal coupling). Using the mean-field approximation and Reflection-Positivity, one may also show that the phase transition is discontinuous if  $d$  is fixed and  $q \geq q_c(d) \gg 1$  [Kotecký and Shlosman \[1982\]](#), or if  $q \geq 3$  is fixed and  $d \geq d_c(q) \gg 1$  [Biskup and Chayes \[2003\]](#). The conjecture that  $q_c(d)$  is equal to 2 for any  $d \geq 3$  remains widely open and represents a beautiful challenge for future mathematical physicists.*

## References

- M. Aizenman, H. Kesten, and C. M. Newman (1987). “Uniqueness of the infinite cluster and continuity of connectivity functions for short and long range percolation”. *Comm. Math. Phys.* 111.4, pp. 505–531. MR: [901151](#) (cit. on p. 2850).
- Michael Aizenman and David J. Barsky (1987). “Sharpness of the phase transition in percolation models”. *Comm. Math. Phys.* 108.3, pp. 489–526. MR: [874906](#) (cit. on pp. 2849, 2862).
- Michael Aizenman, Hugo Duminil-Copin, and Vladas Sidoravicius (2015). “Random currents and continuity of Ising model’s spontaneous magnetization”. *Comm. Math. Phys.* 334.2, pp. 719–742. MR: [3306602](#) (cit. on p. 2871).
- David J. Barsky, Geoffrey R. Grimmett, and Charles M. Newman (1991). “Percolation in half-spaces: equality of critical densities and continuity of the percolation probability”. *Probab. Theory Related Fields* 90.1, pp. 111–148. MR: [1124831](#) (cit. on p. 2851).
- Rodney J Baxter (1973). “Potts model at the critical temperature”. *Journal of Physics C: Solid State Physics* 6.23, p. L445 (cit. on p. 2869).



- Rodney J. Baxter (1989). *Exactly solved models in statistical mechanics*. Reprint of the 1982 original. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], London, pp. xii+486. MR: [998375](#) (cit. on p. [2870](#)).
- William Beckner (1975). “[Inequalities in Fourier analysis](#)”. *Ann. of Math. (2)* 102.1, pp. 159–182. MR: [0385456](#) (cit. on p. [2859](#)).
- Vincent Beffara and Hugo Duminil-Copin (2012). “[The self-dual point of the two-dimensional random-cluster model is critical for  \$q \geq 1\$](#) ”. *Probab. Theory Related Fields* 153.3-4, pp. 511–542. MR: [2948685](#) (cit. on pp. [2860](#), [2864](#), [2867](#)).
- (2013). “[Planar percolation with a glimpse of Schramm-Loewner evolution](#)”. *Probab. Surv.* 10, pp. 1–50. MR: [3161674](#) (cit. on p. [2853](#)).
- A. A. Belavin, A. M. Polyakov, and A. B. Zamolodchikov (1984). “[Infinite conformal symmetry of critical fluctuations in two dimensions](#)”. *J. Statist. Phys.* 34.5-6, pp. 763–774. MR: [751712](#) (cit. on p. [2853](#)).
- Itai Benjamini, Gil Kalai, and Oded Schramm (1999). “[Noise sensitivity of Boolean functions and applications to percolation](#)”. *Inst. Hautes Études Sci. Publ. Math.* 90, 5–43 (2001). MR: [1813223](#) (cit. on p. [2854](#)).
- Itai Benjamini and Oded Schramm (1996). “[Percolation beyond  \$\mathbb{Z}^d\$ , many questions and a few answers](#)”. *Electron. Comm. Probab.* 1, no. 8, 71–82. MR: [1423907](#) (cit. on p. [2851](#)).
- Marek Biskup and Lincoln Chayes (2003). “[Rigorous analysis of discontinuous phase transitions via mean-field bounds](#)”. *Comm. Math. Phys.* 238.1-2, pp. 53–93. MR: [1989669](#) (cit. on p. [2871](#)).
- Béla Bollobás and Oliver Riordan (2006a). “[A short proof of the Harris-Kesten theorem](#)”. *Bull. London Math. Soc.* 38.3, pp. 470–484. MR: [2239042](#) (cit. on p. [2860](#)).
- (2006b). “[The critical probability for random Voronoi percolation in the plane is  \$1/2\$](#) ”. *Probab. Theory Related Fields* 136.3, pp. 417–468. MR: [2257131](#) (cit. on pp. [2859](#), [2860](#)).
- (2010). “[Percolation on self-dual polygon configurations](#)”. In: *An irregular mind*. Vol. 21. Bolyai Soc. Math. Stud. János Bolyai Math. Soc., Budapest, pp. 131–217. MR: [2815602](#) (cit. on p. [2860](#)).
- Aline Bonami (1970). “[Étude des coefficients de Fourier des fonctions de  \$L^p\(G\)\$](#) ”. *Ann. Inst. Fourier (Grenoble)* 20.fasc. 2, 335–402 (1971). MR: [0283496](#) (cit. on p. [2859](#)).
- Jean Bourgain, Jeff Kahn, Gil Kalai, Yitzhak Katznelson, and Nathan Linial (1992). “[The influence of variables in product spaces](#)”. *Israel J. Math.* 77.1-2, pp. 55–64. MR: [1194785](#) (cit. on p. [2859](#)).
- S. R. Broadbent and J. M. Hammersley (1957). “Percolation processes. I. Crystals and mazes”. *Proc. Cambridge Philos. Soc.* 53, pp. 629–641. MR: [0091567](#) (cit. on p. [2848](#)).
- R. M. Burton and M. Keane (1989). “[Density and uniqueness in percolation](#)”. *Comm. Math. Phys.* 121.3, pp. 501–505. MR: [990777](#) (cit. on pp. [2850](#), [2851](#)).

- Federico Camia and Charles M. Newman (2006). “Two-dimensional critical percolation: the full scaling limit”. *Comm. Math. Phys.* 268.1, pp. 1–38. MR: [2249794](#) (cit. on p. [2853](#)).
- John L. Cardy (1992). “Critical percolation in finite geometries”. *J. Phys. A* 25.4, pp. 201–206. MR: [1151081](#) (cit. on p. [2853](#)).
- L Chayes and J Machta (1997). “Graphical representations and cluster algorithms I. Discrete spin systems”. *Physica A: Statistical Mechanics and its Applications* 239.4, pp. 542–601 (cit. on p. [2856](#)).
- Dmitry Chelkak and Stanislav Smirnov (2011). “Discrete complex analysis on isoradial graphs”. *Adv. Math.* 228.3, pp. 1590–1630. MR: [2824564](#) (cit. on p. [2867](#)).
- (2012). “Universality in the 2D Ising model and conformal invariance of fermionic observables”. *Invent. Math.* 189.3, pp. 515–580. MR: [2957303](#) (cit. on p. [2867](#)).
- Computation of the critical point for random-cluster models via the parafermionic observable* (n.d.). Preprint (cit. on p. [2867](#)).
- Eytan Domany, D Mukamel, Bernard Nienhuis, and A Schwimmer (1981). “Duality relations and equivalences for models with  $O(N)$  and cubic symmetry”. *Nuclear Physics B* 190.2, pp. 279–287 (cit. on p. [2856](#)).
- H Duminil-Copin, A Raoufi, and V Tassion (2017a). “Subcritical phase of  $d$ -dimensional Poissonboolean percolation and its vacant set”. preprint (cit. on p. [2863](#)).
- H Duminil-Copin and V Tassion (n.d.). “Renormalization of crossings in planar dependent percolation models”. Preprint (cit. on p. [2869](#)).
- Hugo Duminil-Copin (July 2017). “Lectures on the Ising and Potts models on the hypercubic lattice”. arXiv: [1707.00520](#) (cit. on pp. [2855](#), [2864](#)).
- Hugo Duminil-Copin, Maxime Gagnebin, Matan Harel, Ioan Manolescu, and Vincent Tassion (Nov. 2016). “Discontinuity of the phase transition for the planar random-cluster and Potts models with  $q > 4$ ”. arXiv: [1611.09877](#) (cit. on p. [2869](#)).
- Hugo Duminil-Copin, Alexander Glazman, Ron Peled, and Yinon Spinka (July 2017). “Macroscopic loops in the loop  $O(n)$  model at Nienhuis’ critical point”. arXiv: [1707.09335](#) (cit. on pp. [2867](#), [2869](#)).
- Hugo Duminil-Copin, Clément Hongler, and Pierre Nolin (2011). “Connection probabilities and RSW-type bounds for the two-dimensional FK Ising model”. *Comm. Pure Appl. Math.* 64.9, pp. 1165–1198. MR: [2839298](#) (cit. on p. [2871](#)).
- Hugo Duminil-Copin and Ioan Manolescu (2016). “The phase transitions of the planar random-cluster and Potts models with  $q \geq 1$  are sharp”. *Probab. Theory Related Fields* 164.3–4, pp. 865–892. MR: [3477782](#) (cit. on p. [2860](#)).
- Hugo Duminil-Copin, Aran Raoufi, and Vincent Tassion (May 2017b). “Exponential decay of connection probabilities for subcritical Voronoi percolation in  $\mathbb{R}^d$ ”. arXiv: [1705.07978](#) (cit. on p. [2863](#)).

- Hugo Duminil-Copin, Aran Raoufi, and Vincent Tassion (May 2017c). “Sharp phase transition for the random-cluster and Potts models via decision trees”. arXiv: 1705.03104 (cit. on p. 2863).
- (2018). “A new computation of the critical point for the planar random-cluster model with  $q \geq 1$ ”. *Ann. Inst. Henri Poincaré Probab. Stat.* 54.1, pp. 422–436. arXiv: 1604.03702. MR: 3765895 (cit. on p. 2860).
- Hugo Duminil-Copin, Vladas Sidoravicius, and Vincent Tassion (2017). “Continuity of the phase transition for planar random-cluster and Potts models with  $1 \leq q \leq 4$ ”. *Comm. Math. Phys.* 349.1, pp. 47–107. MR: 3592746 (cit. on pp. 2868–2870).
- Hugo Duminil-Copin and Stanislav Smirnov (2012a). “Conformal invariance of lattice models”. In: *Probability and statistical physics in two and more dimensions*. Vol. 15. Clay Math. Proc. Amer. Math. Soc., Providence, RI, pp. 213–276. MR: 3025392 (cit. on pp. 2867, 2871).
- (2012b). “The connective constant of the honeycomb lattice equals  $\sqrt{2 + \sqrt{2}}$ ”. *Ann. of Math. (2)* 175.3, pp. 1653–1665. MR: 2912714 (cit. on p. 2867).
- Hugo Duminil-Copin and Vincent Tassion (2016). “A new proof of the sharpness of the phase transition for Bernoulli percolation and the Ising model”. *Comm. Math. Phys.* 343.2, pp. 725–745. MR: 3477351 (cit. on p. 2862).
- P. Erdős and A. Rényi (1959). “On random graphs. I”. *Publ. Math. Debrecen* 6, pp. 290–297. MR: 0120167 (cit. on p. 2859).
- Robert Fitzner and Remco van der Hofstad (June 2015). “Mean-field behavior for nearest-neighbor percolation in  $d > 10$ ”. arXiv: 1506.07977 (cit. on p. 2850).
- C. M. Fortuin and P. W. Kasteleyn (1972). “On the random-cluster model. I. Introduction and relation to other models”. *Physica* 57, pp. 536–564. MR: 0359655 (cit. on p. 2855).
- Damien Gaboriau and Russell Lyons (2009). “A measurable-group-theoretic solution to von Neumann’s problem”. *Invent. Math.* 177.3, pp. 533–540. MR: 2534099 (cit. on p. 2852).
- Christophe Garban, Gábor Pete, and Oded Schramm (2011). “The Fourier spectrum of critical percolation [MR2736153]”. In: *Selected works of Oded Schramm. Volume 1*, 2. Sel. Works Probab. Stat. Springer, New York, pp. 445–530. MR: 2883380 (cit. on p. 2854).
- (2013). “Pivotal, cluster, and interface measures for critical planar percolation”. *J. Amer. Math. Soc.* 26.4, pp. 939–1024. MR: 3073882 (cit. on p. 2854).
- Christophe Garban, Gábor Pete, and Oded Schramm (2017). “The scaling limits of near-critical and dynamical percolation”. *Journal of European Math Society* (cit. on p. 2854).
- B. T. Graham and G. R. Grimmett (2006). “Influence and sharp-threshold theorems for monotonic measures”. *Ann. Probab.* 34.5, pp. 1726–1745. MR: 2271479 (cit. on pp. 2860, 2863).

- Geoffrey Grimmett (1999). *Percolation*. Second. Vol. 321. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, pp. xiv+444. MR: [1707339](#) (cit. on p. [2854](#)).
- (2006). *The random-cluster model*. Vol. 333. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, pp. xiv+377. MR: [2243761](#) (cit. on p. [2855](#)).
- Olle Häggström, Yuval Peres, and Jeffrey E. Steif (1997). “Dynamical percolation”. *Ann. Inst. H. Poincaré Probab. Statist.* 33.4, pp. 497–528. MR: [1465800](#) (cit. on p. [2853](#)).
- J. M. Hammersley (1959). “Bornes supérieures de la probabilité critique dans un processus de filtration”. In: *Le calcul des probabilités et ses applications. Paris, 15-20 juillet 1958*. Colloques Internationaux du Centre National de la Recherche Scientifique, LXXXVII. Centre National de la Recherche Scientifique, Paris, pp. 17–37. MR: [0105751](#) (cit. on p. [2848](#)).
- Takashi Hara and Gordon Slade (1990). “Mean-field critical behaviour for percolation in high dimensions”. *Comm. Math. Phys.* 128.2, pp. 333–391. MR: [1043524](#) (cit. on p. [2850](#)).
- T. E. Harris (1960). “A lower bound for the critical probability in a certain percolation process”. *Proc. Cambridge Philos. Soc.* 56, pp. 13–20. MR: [0115221](#) (cit. on pp. [2848](#), [2850](#), [2858](#)).
- Y. Ikhlef, R. Weston, M. Wheeler, and P. Zinn-Justin (2013). “Discrete holomorphicity and quantized affine algebras”. *J. Phys. A* 46.26, pp. 265205, 34. arXiv: [1302.4649](#). MR: [3070959](#) (cit. on p. [2867](#)).
- Yacine Ikhlef and John Cardy (2009). “Discretely holomorphic parafermions and integrable loop models”. *J. Phys. A* 42.10, pp. 102001, 11. MR: [2485852](#) (cit. on p. [2867](#)).
- Jeff Kahn, Gil Kalai, and Nathan Linial (1988). “The influence of variables on Boolean functions”. In: *Foundations of Computer Science, 1988., 29th Annual Symposium on*. IEEE, pp. 68–80 (cit. on p. [2859](#)).
- Richard Kenyon (2000). “Conformal invariance of domino tiling”. *Ann. Probab.* 28.2, pp. 759–795. MR: [1782431](#) (cit. on p. [2867](#)).
- Harry Kesten (1980). “The critical probability of bond percolation on the square lattice equals  $\frac{1}{2}$ ”. *Comm. Math. Phys.* 74.1, pp. 41–59. MR: [575895](#) (cit. on p. [2849](#)).
- (1982). *Percolation theory for mathematicians*. Vol. 2. Progress in Probability and Statistics. Birkhäuser, Boston, Mass., pp. iv+423. MR: [692943](#) (cit. on p. [2850](#)).
- R. Kotecký and S. B. Shlosman (1982). “First-order phase transitions in large entropy lattice models”. *Comm. Math. Phys.* 83.4, pp. 493–515. MR: [649814](#) (cit. on p. [2871](#)).
- Robert Langlands, Philippe Pouliot, and Yvan Saint-Aubin (1994). “Conformal invariance in two-dimensional percolation”. *Bull. Amer. Math. Soc. (N.S.)* 30.1, pp. 1–61. MR: [1230963](#) (cit. on p. [2853](#)).

- M. V. Menshikov (1986). “Coincidence of critical points in percolation problems”. *Dokl. Akad. Nauk SSSR* 288.6, pp. 1308–1311. MR: [852458](#) (cit. on p. [2849](#)).
- Bernard Nienhuis (1982). “Exact critical point and critical exponents of  $O(n)$  models in two dimensions”. *Phys. Rev. Lett.* 49.15, pp. 1062–1065. MR: [675241](#) (cit. on p. [2866](#)).
- Ryan O’Donnell, Michael Saks, Oded Schramm, and Rocco A. Servedio (2005). “Every decision tree has an influential variable”. In: *Foundations of Computer Science, 2005. FOCS 2005. 46th Annual IEEE Symposium on*. IEEE, pp. 31–39 (cit. on p. [2862](#)).
- A. Ju. Ol’sanskiĭ (1980). “On the question of the existence of an invariant mean on a group”. *Uspekhi Mat. Nauk* 35.4(214), pp. 199–200. MR: [586204](#) (cit. on p. [2852](#)).
- Igor Pak and Tatiana Smirnova-Nagnibeda (2000). “On non-uniqueness of percolation on nonamenable Cayley graphs”. *C. R. Acad. Sci. Paris Sér. I Math.* 330.6, pp. 495–500. MR: [1756965](#) (cit. on p. [2852](#)).
- C.-E. Pfister and Y. Velenik (1997). “Random-cluster representation of the Ashkin-Teller model”. *J. Statist. Phys.* 88.5-6, pp. 1295–1331. MR: [1478070](#) (cit. on p. [2856](#)).
- M. A. Rajabpour and J. Cardy (2007). “Discretely holomorphic parafermions in lattice  $Z_N$  models”. *J. Phys. A* 40.49, pp. 14703–14713. MR: [2441869](#) (cit. on p. [2867](#)).
- N. Reshetikhin (Oct. 2010). “Lectures on the integrability of the 6-vertex model”. arXiv: [1010.5031](#) (cit. on p. [2870](#)).
- Lucio Russo (1978). “A note on percolation”. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* 43.1, pp. 39–48 (cit. on pp. [2849](#), [2858](#), [2860](#)).
- (1982). “An approximate zero-one law”. *Z. Wahrsch. Verw. Gebiete* 61.1, pp. 129–139. MR: [671248](#) (cit. on p. [2859](#)).
- Oded Schramm (2000). “Scaling limits of loop-erased random walks and uniform spanning trees”. *Israel J. Math.* 118, pp. 221–288. MR: [1776084](#) (cit. on p. [2853](#)).
- P. D. Seymour and D. J. A. Welsh (1978). “Percolation probabilities on the square lattice”. *Ann. Discrete Math.* 3. Advances in graph theory (Cambridge Combinatorial Conf., Trinity College, Cambridge, 1977), pp. 227–245. MR: [0494572](#) (cit. on pp. [2849](#), [2858](#), [2860](#)).
- S. Smirnov (n.d.). *Private communications* (cit. on p. [2853](#)).
- Stanislav Smirnov (2001). “Critical percolation in the plane: conformal invariance, Cardy’s formula, scaling limits”. *C. R. Acad. Sci. Paris Sér. I Math.* 333.3, pp. 239–244. MR: [1851632](#) (cit. on p. [2853](#)).
- (2010a). “Conformal invariance in random cluster models. I. Holomorphic fermions in the Ising model”. *Ann. of Math. (2)* 172.2, pp. 1435–1467. MR: [2680496](#) (cit. on p. [2870](#)).
- (2010b). “Discrete complex analysis and probability”. In: *Proceedings of the International Congress of Mathematicians. Volume I*. Hindustan Book Agency, New Delhi, pp. 595–621. MR: [2827906](#) (cit. on p. [2865](#)).

- Michel Talagrand (1994). “On Russo’s approximate zero-one law”. *Ann. Probab.* 22.3, pp. 1576–1587. MR: [1303654](#) (cit. on p. [2859](#)).
- Vincent Tassion (2014). “Planarité et localité en percolation”. PhD thesis. Lyon, École normale supérieure (cit. on p. [2860](#)).
- (2016). “Crossing probabilities for Voronoi percolation”. *Ann. Probab.* 44.5, pp. 3385–3398. MR: [3551200](#) (cit. on p. [2860](#)).
- Kevin Whyte (1999). “Amenability, bi-Lipschitz equivalence, and the von Neumann conjecture”. *Duke Math. J.* 99.1, pp. 93–112. MR: [1700742](#) (cit. on p. [2852](#)).
- Andrew Chi Chih Yao (1977). “Probabilistic computations: toward a unified measure of complexity (extended abstract)”, pp. 222–227. MR: [0489016](#) (cit. on p. [2862](#)).

Received 2017-11-28.

HUGO DUMINIL-COPIN  
INSTITUT DES HAUTES ÉTUDES SCIENTIFIQUES AND UNIVERSITÉ DE GENÈVE  
[duminil@ihes.fr](mailto:duminil@ihes.fr)



# RANDOM MATRICES AND HIGH-DIMENSIONAL STATISTICS: BEYOND COVARIANCE MATRICES

NOUREDDINE EL KAROUI

## Abstract

The last twenty-or-so years have seen spectacular progress in our understanding of the fine spectral properties of large-dimensional random matrices. These results have also shown light on the behavior of various statistical estimators used in multivariate statistics. In this short note, we will describe new strands of results, which show that intuition and techniques built on the theory of random matrices and concentration of measure ideas shed new light and bring to the fore new ideas about an arguably even more important set of statistical tools, namely M-estimators and certain bootstrap methods. All the results are obtained in the large  $n$ , large  $p$  setting, where both the number of observations and the number of predictors go to infinity.

## 1 Introduction

Random matrices have a very long history in multivariate statistics, going as far back as [Wishart \[1928\]](#). Traditionally, they have been associated with problems arising from techniques such as Principal Components Analysis (PCA) [Pearson \[1901\]](#), [Hotelling \[1933\]](#), [Anderson \[1963\]](#), and [Jolliffe \[2002\]](#) or covariance matrix estimation where there is a natural focus on estimating spectral properties of large data matrices. We start by setting up precisely the problem and reviewing some of those important results before moving on to new statistical developments.

**1.1 Setup.** In most of this short review, we will be concerned with data stored in a matrix  $X$ , with  $n$  rows and  $p$  columns.  $n$  denotes the number of observations of  $p$  dimensional vectors available to the data analyst. The  $i$ -th row of  $X$  is denoted  $X_i'$  and  $X_i \in \mathbb{R}^p$  is referred to as the  $i$ -th vector of covariates.  $p$ , the dimension of  $X_i$ , is the number of measurements per observation. If one works with financial data for instance [Laloux, Cizeau,](#)

---

The author gratefully acknowledges the support of grant NSF DMS-1510172. He would also like to thank Peter Bickel and Elizabeth Purdom for numerous discussions on these and related topics over the years.

MSC2010: primary 62F12; secondary 60F99, 62F40.



Bouchaud, and M. Potters [1999],  $p$  may be the number of assets in one's portfolio,  $n$  the number of days where those assets are monitored and  $X_{i,j}$  may be the daily return of asset  $j$  on day  $i$ .

**Traditional asymptotics.** Traditionally, statistical theory has been concerned with studying the properties of estimators, i.e. functions of the data matrix  $X$  (and possibly other random variables), as  $n \rightarrow \infty$  while  $p$  stayed fixed Anderson [1984] and Huber [1972] or was growing slowly with  $n$  Portnoy [1984] and Mammen [1989]. While mathematically and statistically interesting at the time, these sorts of problems are now really well-understood and their asymptotic analysis essentially amounts to doing probabilistic perturbation analysis (see more generally van der Vaart [1998]).

**Modern developments.** However, in the last two decades, technological advances in data collection have made it possible to work with datasets where both  $n$  and  $p$  are large: in genomics,  $p$  may be of order tens of thousands or millions and hundreds of observations Ramaswamy et al. [2001], data collected from internet companies may have millions of predictors Criteo [n.d.] and billions of observations, whereas financial data collected daily on a few hundreds of companies would yield after a year a dataset with hundreds of observations and hundreds of predictors Laloux, Cizeau, Bouchaud, and M. Potters [1999].

**The case for “large  $p$ , large  $n$ ”.** It is therefore now natural to study the so called “large  $n$ , large  $p$ ” setting Johnstone [2001, 2007] where  $p$  and  $n$  grow to infinity but  $p/n \rightarrow \kappa \in (0, \infty)$ . On a more mathematical note, the ratio  $p/n$  can be somewhat informally seen as one measure of statistical difficulty of the problem. Fixing it amounts to doing asymptotics while the difficulty of the statistical problem stays constant and hence should (or at least could) yield asymptotic approximations of better quality than their traditional “fixed  $p$ , large  $n$ ” counterparts. This is what we will see in some of the results described below. Furthermore, in the “fixed  $p$ , large  $n$ ” settings, many asymptotic optimality results are meaningful only when it comes to relative errors, however absolute errors are typically infinitesimal and as such may not matter very much to applied statisticians and data analysts. By contrast, we will see that in the “large  $p$ , large  $n$ ” setting, analyses predict substantial absolute differences between methods and as such may inform practitioners in the decision of what methods to use.

**1.2 Modern random matrices.** A key tool in multivariate statistics is the so-called sample covariance matrix, usually denoted, for an  $n \times p$  data matrix  $X$ ,

$$\widehat{\Sigma} = \frac{1}{n-1} (X - \bar{X})(X - \bar{X})'.$$

Here  $\bar{X} = 1_n \widehat{\mu}'$ , where  $\widehat{\mu} \in \mathbb{R}^p$  is the sample mean of the columns, i.e.  $\widehat{\mu} = X'1_n/n$ . (We use  $'$  to denote transposition throughout the paper;  $1_n$  denotes the vector whose entries are all 1 in  $n$  dimension.). The  $p \times p$  matrix  $\widehat{\Sigma}$  therefore simply contains the empirical covariances between the various observed covariates.

This matrix is of course at the heart of much of multivariate statistics as it is the fundamental building block of principal components analysis (PCA) - probably the most widely used dimensionality reduction technique and the template for numerous modern variations - variants such as canonical correlation analysis [Anderson \[1984\]](#), and also plays a key role in the analysis of many supervised learning techniques.

To make things concrete, let us return to PCA. In that technique, practically speaking, the observations  $\{X_i\}_{i=1}^n$  are projected onto the eigenvectors of  $\widehat{\Sigma}$  to perform dimensionality reduction and allow for visualization; see [Hastie, R. Tibshirani, and Friedman \[2009\]](#) for a concrete introduction. A recurring question is how many dimensions should be used for this projection [Cattell \[1966\]](#). This in turn revolves around estimation of eigenvalues questions.

**Classical bulk results.** To get a sense of the utility of large  $n$ , large  $p$  asymptotics in this context, we can return to a classical result [Marčenko and L. A. Pastur \[1967\]](#), which of course was later extended [Wachter \[1978\]](#), [Silverstein \[1995\]](#), [Götze and Tikhomirov \[2004\]](#), [Pajor and L. Pastur \[2009\]](#), and [El Karoui \[2009\]](#) and says the following :

**Theorem 1.1** (Marchenko-Pastur). *Suppose  $X_i$ 's are independent and identically distributed (i.i.d) random variables with mean 0 and covariance identity, i.e.  $\text{cov}(X_i) = \mathbf{E}((X_i - \mathbf{E}(X_i))(X_i - \mathbf{E}(X_i))') = \text{Id}_p$  and mild concentration properties (see above references for details). Suppose further that  $p/n \rightarrow \kappa \in (0, 1)$ . Then the empirical distribution of the eigenvalues of  $\widehat{\Sigma}$  is asymptotically non-random and converges weakly almost surely to  $F_\kappa$ , a distribution whose density can be written*

$$(1) \quad f_\kappa(x) = \frac{\sqrt{(b_\kappa - x)(x - a_\kappa)}}{2\pi x \kappa} 1_{a_\kappa \leq x \leq b_\kappa},$$

where  $b_\kappa = (1 + \sqrt{\kappa})^2$  and  $a_\kappa = (1 - \sqrt{\kappa})^2$ .

This result already illustrates the great difference between modern (i.e. large  $n$ , large  $p$ ) asymptotics and the classical setting where  $p = o(n)$ . In this latter case, the empirical

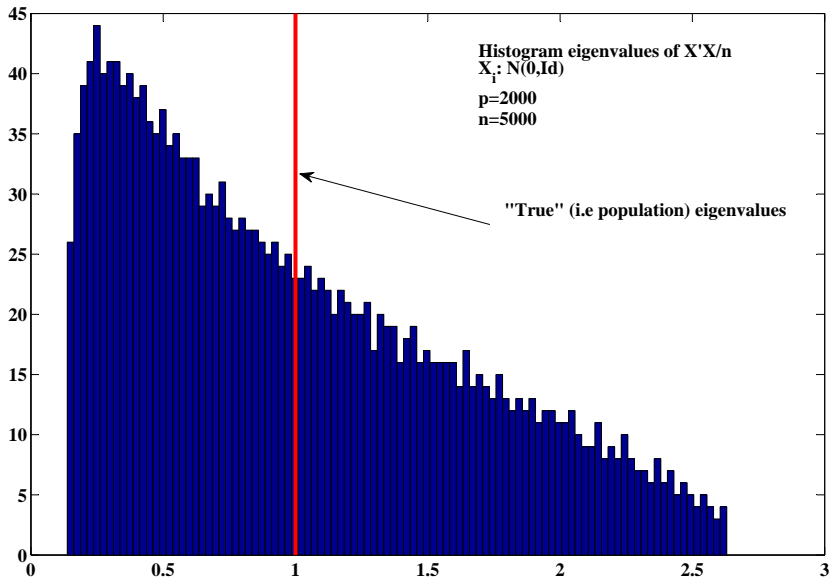


Figure 1: Illustration of Marchenko-Pastur law and high-dimensional estimation problem;  $n=500$ ,  $p=200$ ;  $X_i \sim \mathcal{N}(0, Id_p)$ , i.i.d

distribution of eigenvalues goes, under the assumption of the previous theorem, to a point mass at 1; informally speaking all eigenvalues are consistently (loosely speaking correctly) estimated. The above theorem clearly shows that it is not the case in the “large  $n$ , large  $p$ ” setting.

We can also illustrate the problem with a simple picture, comparing the histogram of observed eigenvalues of  $\hat{\Sigma}$  with the population eigenvalues, i.e. those of  $\text{cov}(X_i) = \Sigma$ . See Figure 1, p. 2878.

This picture clearly illustrates the issue that the new paradigm of high-dimensional statistics creates: even though elementary concentration bounds show that entry-per-entry, i.e. in  $\ell_\infty$  norm, estimation of  $\Sigma$  by e.g.  $\hat{\Sigma}$  is near trivial in the setup we consider, estimation of the spectrum of  $\Sigma$  may not be trivial. We refer the interested reader to El Karoui [2008] and Bickel and Levina [2008] (and Chaudhuri, Drton, and Richardson [2007] in the low-dimensional setting) for early work taking advantage of structure in the covariance matrix to improve estimation and to the recent Bun, Bouchaud, and Potters [2017] for a survey of applied random matrix theoretic work related to the questions we just discussed.

**Right edge results.** In the context of PCA, it is natural to ask questions about the largest eigenvalues of sample covariance matrices, as they could be used in a sequential testing fashion to determine how many components to keep in PCA.

A seminal result in this area in statistics is due to Johnstone who showed, building up on [Tracy and Widom \[1994b,a, 1996\]](#), the following remarkable result in [Johnstone \[2001\]](#).

**Theorem 1.2** (Johnstone). *Suppose  $X_i$ 's are i.i.d  $\mathcal{N}(0, \text{Id}_p)$  and denote by  $l_1$  the largest eigenvalue of  $(n-1)\widehat{\Sigma}$ . Then as  $p$  and  $n$  tend to infinity, while  $p/n \rightarrow \kappa \in (0, \infty)$ , we have*

$$(2) \quad \frac{l_1(\widehat{\Sigma}) - \mu_{n-2,p}}{\sigma_{n-2,p}} \Longrightarrow TW_1 ,$$

with

$$\mu_{n,p} = (\sqrt{n} + \sqrt{p})^2 \quad \text{and} \quad \sigma_{n,p} = (\sqrt{p} + \sqrt{n}) \left( \frac{1}{\sqrt{n}} + \frac{1}{\sqrt{p}} \right)^{1/3} .$$

Here  $TW_1$  is the Tracy-Widom distribution appearing in the study of the Gaussian Orthogonal Ensemble [Mehta \[1991\]](#) and [Deift \[1999\]](#) and  $\Longrightarrow$  denotes weak convergence.

In short, the largest eigenvalue of a sample covariance matrix computed from Gaussian data with identity covariance has fluctuations of size  $n^{-2/3}$  around the edge of the Marchenko-Pastur distribution and the law of these fluctuations is asymptotically Tracy-Widom. Despite the fact that a great deal had been analytically known by statisticians about these questions [James \[1964\]](#), [Constantine \[1963\]](#), and [Muirhead \[1982\]](#) for a number of years, both the scale and the nature of the fluctuations discovered by Johnstone in his breakthrough paper came as a great surprise to the statistics community.

Johnstone's work is also connected to [Forrester \[1993\]](#) and [Johansson \[2000\]](#). Later work extended Johnstone's result in many directions: to cite a few, see [Soshnikov \[2002\]](#) for results concerning the first  $k$  eigenvalues, for any fixed  $k$ , and relaxed distributional assumptions, [El Karoui \[2003\]](#) for the case  $p/n$  tends to 0 or infinity at any rate, [Baik, Ben Arous, and P      \[2005\]](#) for the discovery of very important phase transitions under low rank perturbation of  $\Sigma = \text{Id}_p$ , [El Karoui \[2007\]](#) for the first result on general  $\Sigma$  and [Lee and Schnelli \[2016\]](#) for recent and powerful extensions of great potential in statistics.

This line of research continues with deep and insightful papers [Bloemendal, Knowles, Yau, and Yin \[2016\]](#) and has also benefited from progress in proving universality results - see for instance [Erd  s and Yau \[2012\]](#) and [Tao and Vu \[2012\]](#).

One's enthusiasm for the broad applicability of such results in practice may nonetheless have been tempered by connections made with concentration of measure techniques [Ledoux \[2001\]](#) and [Boucheron, Lugosi, and Massart \[2013\]](#) for instance in [El Karoui and Koesters \[2011\]](#). Those results implied that most of the results above were intimately

linked to effectively geometric (and not probabilistic) assumptions made about the data and that when these easy-to-check-on-the-data assumptions were violated, the results mentioned above did not hold.

**Other directions.** The problems discussed above are of course very linear in nature. As such they have a broad reach beyond linear dimensionality reduction (see below and [El Karoui and Koesters \[2011\]](#) for an example of a dimension-adaptive improvement of linear classification methods). Naturally, the reach of random matrix methods has extended beyond the strictly linear setup. For instance, the beautiful paper [Koltchinskii and Giné \[2000\]](#) studied the spectrum of so-called kernel random matrices, i.e. matrices with entries  $K(i, j) = K(X_i, X_j)$  in the classical setting where  $p$  grows slowly with  $n$ . These results are important for understanding kernel methods in Statistics, which generalize standard methods to higher-dimensional spaces where the inner product between the de-facto observations is not the standard inner product anymore [Wahba \[1990\]](#) and [Schölkopf and Smola \[2002\]](#). These matrices have been well understood in the high-dimensional case for quite a few years now [El Karoui \[2010\]](#) and [Do and Vu \[2013\]](#). Random matrix results also have had interesting applications in randomized linear algebra and numerical optimization, and have been useful in speeding up various algorithms or allowing them to scale to very large data sizes - see for instance [Achlioptas and McSherry \[2007\]](#) and [Drineas, Kannan, and Mahoney \[2006\]](#) and follow-up results. These results typically use mathematically fairly coarse but very nice and broadly applicable bounds [Tropp \[2012\]](#) to prove the reliability of the algorithms under study, a function of the fact that they have to hold in a pretty general setting to be useful to practitioners.

## 2 Beyond covariance matrices: M-estimators

The previous section reviewed results in random matrix theory that could be useful for tasks in exploratory data analysis and generally unsupervised learning. However, much of statistics is concerned with the situation where one observes a scalar response, generically denoted  $Y_i \in \mathbb{R}$ , associated with the vector of predictors  $X_i \in \mathbb{R}^p$ . The simplest model of relationship between the two is the linear model where

$$(\text{linear-model}) \quad \forall i, 1 \leq i \leq n, \quad Y_i = X_i' \beta_0 + \epsilon_i.$$

Here the data  $\{Y_i, X_i\}_{i=1}^n$  are observed. The parameter of interest  $\beta_0 \in \mathbb{R}^p$  is unobserved and so are the errors  $\epsilon_i \in \mathbb{R}$ . Typically, and in this short review,  $\{\epsilon_i\}_{i=1}^n$  are assumed to be i.i.d from a certain distribution. The question the statistician faces is to estimate  $\beta_0$ . This is often done by solving an optimization problem, i.e. using a so-called M-estimator: for

a loss function  $\ell$  chosen by the user,  $\beta_0$  is estimated through

$$\widehat{\beta}_\ell = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \ell(Y_i; X_i' \beta) .$$

In the context of the linear model described above, one often uses the less general formulation

$$(3) \quad \widehat{\beta}_\rho = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^n \rho(Y_i - X_i' \beta) .$$

These estimators and the related family of generalized linear models [McCullagh and Nelder \[1989\]](#) are of fundamental importance in both theoretical and applied statistics and statistical learning, in academia and industry [Chapelle, Manavoglu, and Rosales \[2014\]](#) and [Wood, Goude, and Shaw \[2015\]](#).

**2.1 Classical results: large  $n$ , small  $p$ .** As such these estimators have received a great amount of attention [Relles \[1968\]](#) and [Huber \[1973, 1981\]](#). In the classical case, i.e.  $p$  fixed and  $n \rightarrow \infty$ , [Huber \[1973\]](#) showed, under mild conditions, that  $\widehat{\beta}_\rho$  is asymptotically normally distributed with mean 0 and covariance, if  $\epsilon$  is a random variable with the same distribution as  $\epsilon_i$ 's mentioned in Equation ([linear-model](#)),

$$\operatorname{cov}(\widehat{\beta}_\rho) = (X'X)^{-1} \frac{\mathbf{E}(\psi^2(\epsilon))}{[\mathbf{E}(\psi'(\epsilon))]^2} , \text{ where } \psi = \rho' .$$

This result is striking for at least two reasons : 1) the impact of the design matrix  $X$ , is decoupled from that of the error distribution  $\epsilon$ ; 2) finding the optimal estimator in this class is fairly simple as one just needs to find the function  $\psi$  that minimizes  $\frac{\mathbf{E}(\psi^2(\epsilon))}{[\mathbf{E}(\psi'(\epsilon))]^2}$ . In fact, Huber carried out this program and showed that in low-dimension, when  $\epsilon$  has a density  $f_\epsilon$ , the optimal loss function is

$$\rho_{\text{opt}} = -\log f_\epsilon .$$

In other words, the maximum likelihood estimator [Fisher \[1922\]](#) and [Lehmann and Casella \[1998\]](#) is optimal in this context, when one seeks to minimize the variability of the estimator.

Important work in the 70's, 80's and 90's extended some of these results to various situations where  $p$  was allowed to grow with  $n$  but  $p = o(n)$  - see for instance [Portnoy \[1984, 1985, 1986, 1987\]](#), [Mammen \[1989\]](#), and [Yohai \[1974\]](#). See also [Dümbgen,](#)

[Samworth, and Schuhmacher \[2011\]](#) for more recent results in the classical dimensional framework and very interesting connections with the field of shape restricted estimation [Groeneboom and Jongbloed \[2014\]](#).

**2.2 Modern high-dimensional results: large  $n$ , large  $p$ .** It is natural to ask similar questions to those raised above in the modern context of large  $n$ , large  $p$  asymptotics, as in fact was done as far back as [Huber \[1973\]](#).

Before we proceed, let us say that much effort was devoted in the last two decades in statistics and statistical learning to understanding the properties of the estimators of the form

$$\hat{\beta}_\lambda = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \frac{1}{n} \sum_{i=1}^n \rho(Y_i - X_i' \beta) + \lambda P(\beta) ,$$

where  $P$  is a penalty function, for instance  $P(\beta) = \|\beta\|_2^2$  or  $P(\beta) = \|\beta\|_1$ . However, works in this line of investigation put rather stringent conditions on  $\beta$ , such as dramatic sparsity (i.e. only a fixed number of coefficients of  $\beta_0$  are allowed to not be equal to zero as  $p \rightarrow \infty$ ), which essentially turns these problems into rather classical ones; their analysis depend essentially on well-understood methods, which nonetheless had to be adapted to these specific problems. See [Bühlmann and van de Geer \[2011\]](#) for a book-length survey of this line of work. Let us also note that in truly large case applications [Chapelle, Manavoglu, and Rosales \[2014\]](#), practitioners are not willing to make these stringent assumptions.

**2.2.1 Behavior of the estimator.** By contrast we make no such restrictions on  $\beta_0$ . We focus on the unpenalized case for ease of presentation. To get a sense of results in this context, let us recall the system obtained in [El Karoui, Bean, Bickel, Lim, and Yu \[2013\]](#). Let us consider  $\hat{\beta}$  as in Equation (3). Suppose  $p/n \rightarrow \kappa \in (0, 1)$ . For simplicity assume that are  $X_i \stackrel{iid}{\sim} (0, \operatorname{Id}_p)$ , with i.i.d entries and certain moment conditions - see [El Karoui \[2013, 2018\]](#) for technical details - we have

**Theorem 2.1.** *Under regularity conditions on  $\{\epsilon_i\}$  and  $\rho$  (convex),  $\|\hat{\beta}_\rho - \beta_0\|_2$  is asymptotically deterministic. Call  $r_\rho(\kappa)$  its limit and let  $\hat{z}_\epsilon$  be a random variable with  $\hat{z}_\epsilon = \epsilon + r_\rho(\kappa)Z$ , where  $Z \sim \mathfrak{N}(0, 1)$ , independent of  $\epsilon$ , where  $\epsilon$  has the same distribution as  $\epsilon_i$ 's. For  $c$  deterministic, we have*

$$(4) \quad \begin{cases} \mathbf{E} ([\operatorname{prox}(c\rho)]'(\hat{z}_\epsilon)) &= 1 - \kappa , \\ \kappa r_\rho^2(\kappa) &= \mathbf{E} ([\hat{z}_\epsilon - \operatorname{prox}(c\rho)(\hat{z}_\epsilon)]^2) . \end{cases}$$

where by definition (see [Moreau \[1965\]](#)) for a convex function  $f: \mathbb{R} \mapsto \mathbb{R}$ ,

$$\text{prox}(f)(x) = \operatorname{argmin}_{y \in \mathbb{R}} \left( f(y) + \frac{1}{2}(x - y)^2 \right).$$

We note that the system generalizes easily to much more general setups (involving penalization) - see [El Karoui \[2018\]](#). In particular, the system (4) is quite sensitive to the Euclidean geometry of the predictors,  $X_i$ 's. For instance, if we had  $X_i = \lambda_i Z_i$  where  $Z_i \sim \mathcal{N}(0, \text{Id}_p)$  and  $\lambda_i$  is an independent scalar “well-behaved” random variable with  $\mathbf{E}(\lambda_i^2) = 1$ , a similar type of result would hold, but it would depend on the distribution of  $\lambda_i$  and not only its second moment. In particular,  $r_\rho(\kappa)$  would change, despite the fact that in both models,  $\text{cov}(X_i) = \text{Id}_p$ . As such, one cannot hope for strong universality results in this context. See also [Donoho and Montanari \[2016\]](#) for another point of view on this system.

We also note that the previous result can be generalized to the case where  $\text{cov}(X_i) = \Sigma$  by simple and classical rotational invariance arguments - see [Eaton \[2007\]](#) and [El Karoui, Bean, Bickel, Lim, and Yu \[2013\]](#). In the case where  $X_i$ 's are Gaussian, [El Karoui, Bean, Bickel, Lim, and Yu \[2013\]](#) also uses those to characterize the distribution of  $\widehat{\beta}_\rho - \beta_0$  in a non-asymptotic fashion.

Finally, the behavior of the residuals  $e_i = Y_i - X_i' \widehat{\beta}_\rho$  is very different in high-dimension from what it is in low-dimension; see [El Karoui, Bean, Bickel, Lim, and Yu \[ibid.\]](#) and follow-up papers for characterization. In particular, the residuals are not close in our framework to the “true errors”,  $\epsilon_i$ 's, which is problematic as in many practical statistical methods - based on low-dimensional intuition - the residuals are used as proxies for those “true errors”.

**2.2.2 New loss functions.** In light of the system (4), it is natural to ask which function  $\rho$  minimizes  $r_\rho(\kappa)$ , which is one measure of the inaccuracy of  $\widehat{\beta}_\rho$  as an estimator of  $\beta_0$ . This question was investigated in [Bean, Bickel, El Karoui, and Yu \[2013\]](#). The following result is shown there.

**Theorem 2.2.** *Suppose that  $\epsilon$  has a log-concave density, i.e.  $-\log f_\epsilon$  is convex. Suppose  $r_\rho(\kappa)$  is the solution of (4). Then if  $p_2(x) = x^2/2$ , the optimal loss function that minimizes  $r_\rho(\kappa)$  over convex  $\rho$  functions is*

$$\rho_{opt} = (p_2 + r_{opt}^2 \log \phi_{r_{opt}} \star f_\epsilon)^* - p_2.$$

where  $r_{opt} = \min\{r : r^2 I_\epsilon(r) = p/n\}$ .



In the theorem above,  $\phi_r$  is the density of a mean 0 Gaussian random variable with variance  $r^2$ ,  $\star$  denotes convolution,  $I_\epsilon(r)$  is the Fisher information [Lehmann and Casella \[1998\]](#) of  $\phi_r \star f_\epsilon$  and  $g^*(x) = \sup_{y \in \mathbb{R}} (xy - g(y))$ , is the Fenchel-Legendre dual of  $g$  [Hiriart-Urruty and Lemaréchal \[2001\]](#).

The function  $\rho_{opt}$  can be shown to be convex under the hypotheses of the theorem. It depends of course on  $p/n$ , our proxy for the statistical difficulty of the problem. In other words, this function quantifies the intuitively compelling notion that the loss function we use in these M-estimation problems should be adapted to the statistical hardness of the problem. Interestingly, the function in question is not the maximum likelihood estimator, which is the usual method that is used to determine on statistical grounds the loss function that should be used for a particular problem. We present a (limited) numerical comparison of these new loss functions and the maximum likelihood estimator in [Figure 2](#).

Finally, it should be noted that the impact of choosing a better loss function is not limited to reducing uncertainty about the estimator. It also improves the quality of predictions, as the standard measure of expected prediction error [Hastie, R. Tibshirani, and Friedman \[2009\]](#) is closely tied to the size of  $\mathbb{E} \left( \|\hat{\beta}_\rho - \beta_0\|_2^2 \right)$  in the models we consider.

### 3 Bootstrap and resampling questions

Modern statistics is increasingly computational and as such many methods have been devised to try to assess sampling variability of estimators through the use of simulations and without relying on asymptotic analyses. In other words, there are numerical ways to try to get at results such as those obtained in [Theorems 1.2 and 2.1](#) for instance.

The most prominent of such methods is the bootstrap, proposed by Efron in the breakthrough paper [Efron \[1979\]](#). Numerous variants of the bootstrap have appeared since then, and the bootstrap created an entire field of research, both theoretical and applied. See for instance [Bickel and Freedman \[1981\]](#), [Efron \[1982\]](#), [Davison and Hinkley \[1997\]](#), [Hall \[1992\]](#), and [Efron and R. J. Tibshirani \[1993\]](#) for classic references.

It is therefore natural to ask how the bootstrap performs in the modern high-dimensional context. Before we present some results in this direction, let us give a very brief introduction to the non-parametric bootstrap.

**3.1 Non-parametric bootstrap and plug-in principle.** As a so-called resampling method, the bootstrap seeks to re-use the data to assess for instance the variability of an estimator. Concretely, suppose we have data  $\{X_i\}_{i=1}^n \in \mathbb{R}^p$ , assumed to be i.i.d. and we are interested in the fluctuation behavior of a statistic/function of the data  $\hat{\theta} = \theta(X_1, \dots, X_n)$ . For instance,  $\hat{\theta}$  could be the sample mean of the  $X_i$ 's or the largest eigenvalue of the sample covariance matrix of the  $X_i$ 's.

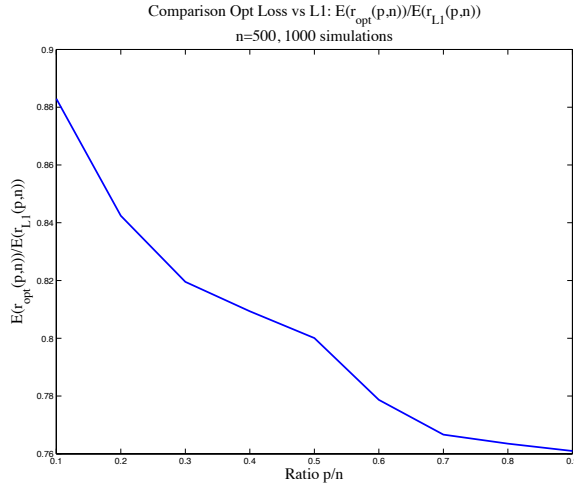


Figure 2: Numerical comparison of dimension-adaptive optimal loss and maximum likelihood loss: case where  $f_\epsilon(x) = e^{-|x|}/2$ , a.k.a. double exponential errors. We plot the ratio  $\mathbf{E}(\|\hat{\beta}_{opt} - \beta_0\|_2) / \mathbf{E}(\|\hat{\beta}_{l_1} - \beta_0\|_2)$  as a function of  $p/n$ . The ratio is always less than 1 :  $\rho_{opt}$ , which varies with  $p/n$  and is used to compute  $\hat{\beta}_{opt}$ , beats  $\ell_1$  loss, i.e.  $\rho(x) = |x|$ , the “optimal loss” in this context according to maximum likelihood theory. The curve is obtained by estimating the expectation through averaging over 1,000 independent simulations.

The non-parametric bootstrap uses the following algorithm :

- For  $b = 1, \dots, B$ , repeat:
- Sample  $n$  times with replacement from  $\{X_i\}_{i=1}^n$ , to get bootstrapped dataset  $D_b = \{X_{1,b}^*, \dots, X_{n,b}^*\}$ .
- Compute  $\hat{\theta}(X^*)_{n,b} = \theta(X_{1,b}^*, \dots, X_{n,b}^*)$ .

Then the empirical distribution of  $\{\hat{\theta}(X^*)_{n,b}\}_{b=1}^B$  is used to assess the sampling variability of the original statistic  $\hat{\theta} = \theta(X_1, \dots, X_n)$  for instance by computing the bootstrap estimate of variance (i.e. the empirical variance of  $\{\hat{\theta}(X^*)_{n,b}\}_{b=1}^B$  if the statistic is one-dimensional), or more sophisticated functions of the empirical distribution.

This is the so-called plug-in principle: one considers that the bootstrap data-generating process mimics the “true” (i.e. sampling from the population ) data-generating process and

proceeds with bootstrap data as one would do with data sampled from the population. As such the bootstrap offers the promise of uncertainty assessment for arbitrarily complicated statistics without much need for mathematical understanding.

One natural question is of course to know when the bootstrap works (and what it means for the bootstrap to work). The first such results appeared in the pioneering [Bickel and Freedman \[1981\]](#); nowadays, a common way to look at this problem is by looking at  $\theta$  as a function over probability distributions -  $\hat{\theta}$  being  $\theta$  applied to the empirical distribution of the data - and requiring  $\theta$  to be sufficiently smooth in an appropriate sense [van der Vaart \[1998\]](#).

**3.2 Bootstrapping regression M-estimates.** Because of the lack of closed formulae to characterize the behavior of estimators such as  $\hat{\beta}_\rho$  defined in Equation (3), the bootstrap became early on an appealing tool to use for this task [Shorack \[1982\]](#) and questions related to the ones we raise in the high-dimensional setting were addressed in setting where  $p/n \rightarrow 0$  in [Wu \[1986\]](#) and [Mammen \[1989, 1993\]](#).

In [El Karoui and Purdom \[n.d.\]](#), various results concerning the bootstrap in high-dimension regression are presented. Bootstrapping as described above the observations  $\{(Y_i, X_i)\}_{i=1}^n$  is called the pairs bootstrap in this setting. Elementary algebra shows that the pairs bootstrap amounts to fitting weighted regression models, i.e for bootstrap weights  $\{w_i^*\}$ ,

$$\hat{\beta}_{\rho,w}^* = \operatorname{argmin}_{\beta \in \mathbb{R}^p} \sum_{i=1}^n w_i^* \rho(Y_i - X_i' \beta).$$

For instance, it is shown that (for precise technical details see [El Karoui and Purdom \[ibid.\]](#)):

**Theorem 3.1.** *Suppose weights  $(w_i)_{i=1}^n$  are i.i.d.,  $\mathbf{E}(w_i) = 1$ , have sufficiently many moments and are bounded away from 0. Let  $X_i \stackrel{iid}{\sim} \mathfrak{N}(0, \operatorname{Id}_p)$  and let  $v$  be a (sequence of) deterministic unit vector.*

*Suppose  $\hat{\beta}$  is obtained by solving a least-squares problem, i.e  $\rho(x) = x^2/2$  and that the linear model holds. Let us call  $\operatorname{var}(\epsilon_i) = \sigma_\epsilon^2$  and corresponding bootstrapped estimates  $\hat{\beta}_w^*$ .*

*If  $\lim p/n = \kappa < 1$  then asymptotically as  $n \rightarrow \infty$*

$$p \mathbf{E} \left( \operatorname{var} \left( v' \hat{\beta}_w^* \right) \right) \rightarrow \sigma_\epsilon^2 \left[ \kappa \frac{1}{1 - \kappa - \mathbf{E} \left( \frac{1}{(1 + c w_i)^2} \right)} - \frac{1}{1 - \kappa} \right],$$

where  $c$  is the unique solution of

$$\mathbf{E} \left( \frac{1}{1 + cw_i} \right) = 1 - \kappa .$$

We note that in the previous context, it is not complicated to show that

$$p\text{var} \left( v' \widehat{\beta} \right) \rightarrow \sigma_\epsilon^2 \frac{\kappa}{1 - \kappa} .$$

Therefore the type of bootstraps described above fails at the very simple task of estimating the variance  $v' \widehat{\beta}$ , even for least squares. [Figure 3](#) on p. 2896 gives a graphical illustration of the problem, showing that the bootstrap overestimates the variance of our estimator.

[El Karoui and Purdom \[ibid.\]](#) contains many other results concerning other types of bootstraps and other resampling techniques, such as the jackknife. In general, the results show that even when classical bootstrap theory would suggest that the bootstrap should work (i.e. the statistics of interest are sufficiently “smooth”), it does not work in high-dimension, even when the statistician has very minimal requirements about what it means to work. Problematically, various bootstraps can fail in many ways, yielding confidence intervals with either too much or not enough coverage for instance. See [El Karoui and Purdom \[ibid.\]](#) for details and relations to relevant literature as well as [Bickel and Freedman \[1983\]](#) for an early take on closely related questions, with however different requirements concerning bootstrap performance and analysis of a different kind of bootstraps.

**3.3 Bootstrap and eigenvalues.** It is also natural to wonder whether the bootstrap would be able to “automatically discover” results such as [Theorem 1.2](#) and adapt to phase transitions such as the one discovered in [Baik, Ben Arous, and P      \[2005\]](#). Analysis of the bootstrap for eigenvalues in low-dimension goes as far back as [Beran and Srivastava \[1985\]](#) and [Eaton and Tyler \[1991\]](#). In [El Karoui and Purdom \[2016\]](#), questions of that type are investigated in high-dimension through a mix of theory and simulations, for various statistics related to eigenvalues of random matrices. Many mathematical questions remain open; however the results are generally negative, in that typically bootstrap confidence intervals do not have the right coverage probabilities. The only positive results about the bootstrap in that context are situations where the population covariance  $\Sigma$  has very isolated eigenvalues, and the problem is hence effectively low-dimensional and therefore of limited mathematical interest.

As such the bootstrap appears as of this writing to be a genuinely perturbation analytic technique and hence to be poorly suited to the kind of problems discussed in this short review.

## 4 Conclusions

We have presented a small overview of recent results in theoretical statistics focused on the high-dimensional case, where the number of measurements per observations grows with the number of observations.

Mathematical analysis in this setup reveals the breakdown of basic consistency results. Furthermore, classical optimality results (based essentially on the likelihood principle) do not hold, yielding results and methods that upended many practitioners' intuition.

Interestingly, the analyses summarized above led the way to the proposal of new loss functions outside of "standard" families and adapted to the statistical difficulty of the problem, as measured by  $p/n$ .

Finally, standard data-driven methods of uncertainty assessment such as the bootstrap seem to completely break down in this setup, where they are most needed by practitioners given the complexity of the problems.

As such the large  $n$ , large  $p$  setting is much more than just a technical hurdle for theoreticians but seems to call for a serious rethinking of tools used by statisticians, whether they be involved in theory, methodology or applications.

Much mathematically stimulating work remains to be done to be able to develop improved methods (both for estimation and uncertainty assessment) and improve our understanding of statistics in this still novel and challenging framework.

## References

- Dimitris Achlioptas and Frank McSherry (2007). "Fast computation of low-rank matrix approximations". *J. ACM* 54.2, Art. 9, 19. MR: [2295993](#) (cit. on p. [2880](#)).
- T. W. Anderson (1963). "Asymptotic theory for principal component analysis". *Ann. Math. Statist.* 34, pp. 122–148. MR: [0145620](#) (cit. on p. [2875](#)).
- (1984). *An introduction to multivariate statistical analysis*. Second. Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, pp. xviii+675. MR: [771294](#) (cit. on pp. [2876](#), [2877](#)).
- Jinho Baik, Gérard Ben Arous, and Sandrine Péché (2005). "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices". *Ann. Probab.* 33.5, pp. 1643–1697. MR: [2165575](#) (cit. on pp. [2879](#), [2887](#)).
- D. Bean, P. J. Bickel, Nouredine El Karoui, and B. Yu (2013). "Optimal m-estimation in high-dimensional regression". *Proceedings of the National Academy of Sciences* 110 (36), pp. 14563–14568 (cit. on p. [2883](#)).
- Rudolf Beran and Muni S. Srivastava (1985). "Bootstrap tests and confidence regions for functions of a covariance matrix". *Ann. Statist.* 13.1, pp. 95–115. MR: [773155](#) (cit. on p. [2887](#)).

- P. J. Bickel and D. A. Freedman (1983). “Bootstrapping regression models with many parameters”. In: *A Festschrift for Erich L. Lehmann*. Wadsworth Statist./Probab. Ser. Wadsworth, Belmont, Calif., pp. 28–48. MR: [689736](#) (cit. on p. [2887](#)).
- Peter J. Bickel and David A. Freedman (1981). “Some asymptotic theory for the bootstrap”. *Ann. Statist.* 9.6, pp. 1196–1217. MR: [630103](#) (cit. on pp. [2884](#), [2886](#)).
- Peter J. Bickel and Elizaveta Levina (2008). “Covariance regularization by thresholding”. *Ann. Statist.* 36.6, pp. 2577–2604. MR: [2485008](#) (cit. on p. [2878](#)).
- Alex Bloemendal, Antti Knowles, Horng-Tzer Yau, and Jun Yin (2016). “On the principal components of sample covariance matrices”. *Probab. Theory Related Fields* 164.1-2, pp. 459–552. MR: [3449395](#) (cit. on p. [2879](#)).
- Stéphane Boucheron, Gábor Lugosi, and Pascal Massart (2013). *Concentration inequalities*. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, pp. x+481. MR: [3185193](#) (cit. on p. [2879](#)).
- Peter Bühlmann and Sara van de Geer (2011). *Statistics for high-dimensional data*. Springer Series in Statistics. Methods, theory and applications. Springer, Heidelberg, pp. xviii+556. MR: [2807761](#) (cit. on p. [2882](#)).
- Joël Bun, Jean-Philippe Bouchaud, and Marc Potters (2017). “Cleaning large correlation matrices: tools from random matrix theory”. *Phys. Rep.* 666, pp. 1–109. MR: [3590056](#) (cit. on p. [2878](#)).
- R. Cattell (1966). “The scree test for the number of factors”. *Multivariate Behav. Res.* 1, pp. 245–276 (cit. on p. [2877](#)).
- O. Chapelle, E. Manavoglu, and R. Rosales (Dec. 2014). “Simple and scalable response prediction for display advertising”. *ACM Trans. Intell. Syst. Technol.* 5 (4), 61:1–61:34 (cit. on pp. [2881](#), [2882](#)).
- Sanjay Chaudhuri, Mathias Drton, and Thomas S. Richardson (2007). “Estimation of a covariance matrix with zeros”. *Biometrika* 94.1, pp. 199–216. MR: [2307904](#) (cit. on p. [2878](#)).
- A. G. Constantine (1963). “Some non-central distribution problems in multivariate analysis”. *Ann. Math. Statist.* 34, pp. 1270–1285. MR: [0181056](#) (cit. on p. [2879](#)).
- Criteo* (n.d.). Criteo public datasets (cit. on p. [2876](#)).
- A. C. Davison and D. V. Hinkley (1997). *Bootstrap methods and their application*. Vol. 1. Cambridge Series in Statistical and Probabilistic Mathematics. With 1 IBM-PC floppy disk (3.5 inch; HD). Cambridge University Press, Cambridge, pp. x+582. MR: [1478673](#) (cit. on p. [2884](#)).
- P. A. Deift (1999). *Orthogonal polynomials and random matrices: a Riemann-Hilbert approach*. Vol. 3. Courant Lecture Notes in Mathematics. New York University, Courant Institute of Mathematical Sciences, New York; American Mathematical Society, Providence, RI, pp. viii+273. MR: [1677884](#) (cit. on p. [2879](#)).

- Yen Do and Van Vu (2013). “[The spectrum of random kernel matrices: universality results for rough and varying kernels](#)”. *Random Matrices Theory Appl.* 2.3, pp. 1350005, 29. MR: [3109422](#) (cit. on p. [2880](#)).
- David Donoho and Andrea Montanari (2016). “[High dimensional robust M-estimation: asymptotic variance via approximate message passing](#)”. *Probab. Theory Related Fields* 166.3–4, pp. 935–969. MR: [3568043](#) (cit. on p. [2883](#)).
- Petros Drineas, Ravi Kannan, and Michael W. Mahoney (2006). “[Fast Monte Carlo algorithms for matrices. II. Computing a low-rank approximation to a matrix](#)”. *SIAM J. Comput.* 36.1, pp. 158–183. MR: [2231644](#) (cit. on p. [2880](#)).
- Lutz Dümbgen, Richard Samworth, and Dominic Schuhmacher (2011). “[Approximation by log-concave distributions, with applications to regression](#)”. *Ann. Statist.* 39.2, pp. 702–730. MR: [2816336](#) (cit. on p. [2881](#)).
- Morris L. Eaton (2007). *Multivariate statistics*. Vol. 53. Institute of Mathematical Statistics Lecture Notes—Monograph Series. A vector space approach, Reprint of the 1983 original [MR0716321]. Institute of Mathematical Statistics, Beachwood, OH, pp. viii+512. MR: [2431769](#) (cit. on p. [2883](#)).
- Morris L. Eaton and David E. Tyler (1991). “[On Wielandt’s inequality and its application to the asymptotic distribution of the eigenvalues of a random symmetric matrix](#)”. *Ann. Statist.* 19.1, pp. 260–271. MR: [1091849](#) (cit. on p. [2887](#)).
- B. Efron (1979). “[Bootstrap methods: another look at the jackknife](#)”. *Ann. Statist.* 7.1, pp. 1–26. MR: [515681](#) (cit. on p. [2884](#)).
- Bradley Efron (1982). *The jackknife, the bootstrap and other resampling plans*. Vol. 38. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, Pa., pp. vi+92. MR: [659849](#) (cit. on p. [2884](#)).
- Bradley Efron and Robert J. Tibshirani (1993). *An introduction to the bootstrap*. Vol. 57. Monographs on Statistics and Applied Probability. Chapman and Hall, New York, pp. xvi+436. MR: [1270903](#) (cit. on p. [2884](#)).
- Noureddine El Karoui (Sept. 2003). “[On the largest eigenvalue of Wishart matrices with identity covariance when  \$n\$ ,  \$p\$  and  \$p/n\$  tend to infinity](#)”. arXiv: [math/0309355](#) (cit. on p. [2879](#)).
- (2007). “[Tracy-Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices](#)”. *Ann. Probab.* 35.2, pp. 663–714. MR: [2308592](#) (cit. on p. [2879](#)).
  - (2008). “[Operator norm consistent estimation of large-dimensional sparse covariance matrices](#)”. *Ann. Statist.* 36.6, pp. 2717–2756. MR: [2485011](#) (cit. on p. [2878](#)).
  - (2009). “[Concentration of measure and spectra of random matrices: applications to correlation matrices, elliptical distributions and beyond](#)”. *Ann. Appl. Probab.* 19.6, pp. 2362–2405. MR: [2588248](#) (cit. on p. [2877](#)).

- (2010). “[The spectrum of kernel random matrices](#)”. *Ann. Statist.* 38.1, pp. 1–50. MR: [2589315](#) (cit. on p. 2880).
- (Nov. 2013). “[Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators : rigorous results](#)”. arXiv: [1311.2445](#) (cit. on p. 2882).
- (2018). “[On the impact of predictor geometry on the performance on high-dimensional ridge-regularized generalized robust regression estimators](#)”. *Probab. Theory Related Fields* 170.1-2, pp. 95–175. MR: [3748322](#) (cit. on pp. 2882, 2883).
- Noureddine El Karoui, D. Bean, P. J. Bickel, C. Lim, and B. Yu (2013). *On robust regression with high-dimensional predictors* (cit. on pp. 2882, 2883).
- Noureddine El Karoui and Holger Koesters (May 2011). “[Geometric sensitivity of random matrix results: consequences for shrinkage estimators of covariance and related statistical methods](#)”. arXiv: [1105.1404](#) (cit. on pp. 2879, 2880).
- Noureddine El Karoui and E. Purdom (n.d.). “Can we trust the bootstrap in high-dimension?” Technical Report 824, UC Berkeley, Department of Statistics, February 2015 (cit. on pp. 2886, 2887).
- Noureddine El Karoui and Elizabeth Purdom (Aug. 2016). “[The bootstrap, covariance matrices and PCA in moderate and high-dimensions](#)”. arXiv: [1608.00948](#) (cit. on p. 2887).
- László Erdős and Horng-Tzer Yau (2012). “[Universality of local spectral statistics of random matrices](#)”. *Bull. Amer. Math. Soc. (N.S.)* 49.3, pp. 377–414. MR: [2917064](#) (cit. on p. 2879).
- R. A. Fisher (1922). “On the mathematical foundations of theoretical statistics”. *Philosophical Transactions of the Royal Society, A* 222, pp. 309–368 (cit. on p. 2881).
- P. J. Forrester (1993). “[The spectrum edge of random matrix ensembles](#)”. *Nuclear Phys. B* 402.3, pp. 709–728. MR: [1236195](#) (cit. on p. 2879).
- Friedrich Götze and Alexander Tikhomirov (2004). “[Rate of convergence in probability to the Marchenko-Pastur law](#)”. *Bernoulli* 10.3, pp. 503–548. MR: [2061442](#) (cit. on p. 2877).
- Piet Groeneboom and Geurt Jongbloed (2014). *Nonparametric estimation under shape constraints*. Vol. 38. Cambridge Series in Statistical and Probabilistic Mathematics. Estimators, algorithms and asymptotics. Cambridge University Press, New York, pp. xi+416. MR: [3445293](#) (cit. on p. 2882).
- Peter Hall (1992). *The bootstrap and Edgeworth expansion*. Springer Series in Statistics. Springer-Verlag, New York, pp. xiv+352. MR: [1145237](#) (cit. on p. 2884).
- Trevor Hastie, Robert Tibshirani, and Jerome Friedman (2009). *The elements of statistical learning*. Second. Springer Series in Statistics. Data mining, inference, and prediction. Springer, New York, pp. xxii+745. MR: [2722294](#) (cit. on pp. 2877, 2884).



- Jean-Baptiste Hiriart-Urruty and Claude Lemaréchal (2001). *Fundamentals of convex analysis*. Grundlehren Text Editions. Abridged version of it Convex analysis and minimization algorithms. I [Springer, Berlin, 1993; MR1261420 (95m:90001)] and it II [ibid.; MR1295240 (95m:90002)]. Springer, Berlin, pp. x+259. MR: [1865628](#) (cit. on p. [2884](#)).
- H. Hotelling (1933). “Analysis of a complex of statistical variables into principal components”. *Journal of Educational Psychology* 24, pp. 417–441 (cit. on p. [2875](#)).
- Peter J. Huber (1972). “The 1972 Wald lecture. Robust statistics: A review”. *Ann. Math. Statist.* 43, pp. 1041–1067. MR: [0314180](#) (cit. on p. [2876](#)).
- (1973). “Robust regression: asymptotics, conjectures and Monte Carlo”. *Ann. Statist.* 1, pp. 799–821. MR: [0356373](#) (cit. on pp. [2881](#), [2882](#)).
  - (1981). *Robust statistics*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, pp. ix+308. MR: [606374](#) (cit. on p. [2881](#)).
- Alan T. James (1964). “Distributions of matrix variates and latent roots derived from normal samples”. *Ann. Math. Statist.* 35, pp. 475–501. MR: [0181057](#) (cit. on p. [2879](#)).
- Kurt Johansson (2000). “Shape fluctuations and random matrices”. *Comm. Math. Phys.* 209.2, pp. 437–476. MR: [1737991](#) (cit. on p. [2879](#)).
- Iain M. Johnstone (2001). “On the distribution of the largest eigenvalue in principal components analysis”. *Ann. Statist.* 29.2, pp. 295–327. MR: [1863961](#) (cit. on pp. [2876](#), [2879](#)).
- (2007). “High dimensional statistical inference and random matrices”. In: *International Congress of Mathematicians. Vol. I*. Eur. Math. Soc., Zürich, pp. 307–333. MR: [2334195](#) (cit. on p. [2876](#)).
- I. T. Jolliffe (2002). *Principal component analysis*. Second. Springer Series in Statistics. Springer-Verlag, New York, pp. xxx+487. MR: [2036084](#) (cit. on p. [2875](#)).
- Vladimir Koltchinskii and Evarist Giné (2000). “Random matrix approximation of spectra of integral operators”. *Bernoulli* 6.1, pp. 113–167. MR: [1781185](#) (cit. on p. [2880](#)).
- L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters (1999). “Noise dressing of financial correlation matrices”. *Phys. Rev. Lett.* 83 (7), pp. 1467–1470 (cit. on pp. [2875](#), [2876](#)).
- Michel Ledoux (2001). *The concentration of measure phenomenon*. Vol. 89. Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI, pp. x+181. MR: [1849347](#) (cit. on p. [2879](#)).
- Ji Oon Lee and Kevin Schnelli (2016). “Tracy–Widom distribution for the largest eigenvalue of real sample covariance matrices with general population”. *Ann. Appl. Probab.* 26.6, pp. 3786–3839. arXiv: [1409.4979](#). MR: [3582818](#) (cit. on p. [2879](#)).
- E. L. Lehmann and George Casella (1998). *Theory of point estimation*. Second. Springer Texts in Statistics. Springer-Verlag, New York, pp. xxvi+589. MR: [1639875](#) (cit. on pp. [2881](#), [2884](#)).

- Enno Mammen (1989). “Asymptotics with increasing dimension for robust regression with applications to the bootstrap”. *Ann. Statist.* 17.1, pp. 382–400. MR: [981457](#) (cit. on pp. [2876](#), [2881](#), [2886](#)).
- (1993). “Bootstrap and wild bootstrap for high-dimensional linear models”. *Ann. Statist.* 21.1, pp. 255–285. MR: [1212176](#) (cit. on p. [2886](#)).
- V. A. Marčenko and L. A. Pastur (1967). “Distribution of eigenvalues in certain sets of random matrices”. *Mat. Sb. (N.S.)* 72 (114), pp. 507–536. MR: [0208649](#) (cit. on p. [2877](#)).
- P. McCullagh and J. A. Nelder (1989). *Generalized linear models*. Monographs on Statistics and Applied Probability. Second edition [of MR0727836]. Chapman & Hall, London, pp. xix+511. MR: [3223057](#) (cit. on p. [2881](#)).
- Madan Lal Mehta (1991). *Random matrices*. Second. Academic Press, Inc., Boston, MA, pp. xviii+562. MR: [1083764](#) (cit. on p. [2879](#)).
- Jean-Jacques Moreau (1965). “Proximité et dualité dans un espace hilbertien”. *Bull. Soc. Math. France* 93, pp. 273–299. MR: [0201952](#) (cit. on p. [2883](#)).
- Robb J. Muirhead (1982). *Aspects of multivariate statistical theory*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., New York, p. 673. MR: [652932](#) (cit. on p. [2879](#)).
- A. Pajor and L. Pastur (2009). “On the limiting empirical measure of eigenvalues of the sum of rank one matrices with log-concave distribution”. *Studia Math.* 195.1, pp. 11–29. MR: [2539559](#) (cit. on p. [2877](#)).
- K. Pearson (1901). “On lines and planes of closest fit to systems of points in space”. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* 2, pp. 559–572 (cit. on p. [2875](#)).
- Stephen Portnoy (1984). “Asymptotic behavior of  $M$ -estimators of  $p$  regression parameters when  $p^2/n$  is large. I. Consistency”. *Ann. Statist.* 12.4, pp. 1298–1309. MR: [760690](#) (cit. on pp. [2876](#), [2881](#)).
- (1985). “Asymptotic behavior of  $M$  estimators of  $p$  regression parameters when  $p^2/n$  is large. II. Normal approximation”. *Ann. Statist.* 13.4, pp. 1403–1417. MR: [811499](#) (cit. on p. [2881](#)).
- (1986). “Asymptotic behavior of the empiric distribution of  $M$ -estimated residuals from a regression model with many parameters”. *Ann. Statist.* 14.3, pp. 1152–1170. MR: [856812](#) (cit. on p. [2881](#)).
- (1987). “A central limit theorem applicable to robust regression estimators”. *J. Multivariate Anal.* 22.1, pp. 24–50. MR: [890880](#) (cit. on p. [2881](#)).
- S. Ramaswamy et al. (2001). “Multiclass cancer diagnosis using tumor gene expression signatures”. 98, pp. 15149–15154 (cit. on p. [2876](#)).
- Daniel Arthur Relles (1968). *Robust Regression by Modified Least Squares*. Thesis (Ph.D.)–Yale University. ProQuest LLC, Ann Arbor, MI, p. 135. MR: [2617863](#) (cit. on p. [2881](#)).

- B. Schölkopf and A. J. Smola (2002). *Learning with kernels*. Cambridge, MA: The MIT Press (cit. on p. [2880](#)).
- Galen R. Shorack (1982). “Bootstrapping robust regression”. *Comm. Statist. A—Theory Methods* 11.9, pp. 961–972. MR: [655465](#) (cit. on p. [2886](#)).
- Jack W. Silverstein (1995). “Strong convergence of the empirical distribution of eigenvalues of large-dimensional random matrices”. *J. Multivariate Anal.* 55.2, pp. 331–339. MR: [1370408](#) (cit. on p. [2877](#)).
- Alexander Soshnikov (2002). “A note on universality of the distribution of the largest eigenvalues in certain sample covariance matrices”. *J. Statist. Phys.* 108.5-6. Dedicated to David Ruelle and Yasha Sinai on the occasion of their 65th birthdays, pp. 1033–1056. MR: [1933444](#) (cit. on p. [2879](#)).
- Terence Tao and Van Vu (2012). “Random covariance matrices: universality of local statistics of eigenvalues”. *Ann. Probab.* 40.3, pp. 1285–1315. MR: [2962092](#) (cit. on p. [2879](#)).
- Craig A. Tracy and Harold Widom (1994a). “Fredholm determinants, differential equations and matrix models”. *Comm. Math. Phys.* 163.1, pp. 33–72. MR: [1277933](#) (cit. on p. [2879](#)).
- (1994b). “Level-spacing distributions and the Airy kernel”. *Comm. Math. Phys.* 159.1, pp. 151–174. MR: [1257246](#) (cit. on p. [2879](#)).
- (1996). “On orthogonal and symplectic matrix ensembles”. *Comm. Math. Phys.* 177.3, pp. 727–754. MR: [1385083](#) (cit. on p. [2879](#)).
- Joel A. Tropp (2012). “User-friendly tail bounds for sums of random matrices”. *Found. Comput. Math.* 12.4, pp. 389–434. MR: [2946459](#) (cit. on p. [2880](#)).
- A. W. van der Vaart (1998). *Asymptotic statistics*. Vol. 3. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge, pp. xvi+443. MR: [1652247](#) (cit. on pp. [2876](#), [2886](#)).
- Kenneth W. Wachter (1978). “The strong limits of random matrix spectra for sample matrices of independent elements”. *Ann. Probability* 6.1, pp. 1–18. MR: [0467894](#) (cit. on p. [2877](#)).
- Grace Wahba (1990). *Spline models for observational data*. Vol. 59. CBMS-NSF Regional Conference Series in Applied Mathematics. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, pp. xii+169. MR: [1045442](#) (cit. on p. [2880](#)).
- J. Wishart (1928). “The generalised product moment distribution in samples from a normal multivariate population”. *Biometrika* 20 (A), pp. 32–52 (cit. on p. [2875](#)).
- Simon N. Wood, Yannig Goude, and Simon Shaw (2015). “Generalized additive models for large data sets”. *J. R. Stat. Soc. Ser. C. Appl. Stat.* 64.1, pp. 139–155. MR: [3293922](#) (cit. on p. [2881](#)).
- C.-F. J. Wu (1986). “Jackknife, bootstrap and other resampling methods in regression analysis”. *Ann. Statist.* 14.4. With discussion and a rejoinder by the author, pp. 1261–1350. MR: [868303](#) (cit. on p. [2886](#)).

Víctor J. Yohai (1974). “[Robust estimation in the linear model](#)”. *Ann. Statist.* 2. Collection of articles dedicated to Jerzy Neyman on his 80th birthday, pp. 562–567. MR: [0365875](#) (cit. on p. [2881](#)).

Received 2017-12-07.

NOUREDDINE EL KAROUI

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA AT BERKELEY

and

CRITEO RESEARCH

[nkaroui@berkeley.edu](mailto:nkaroui@berkeley.edu)

[n.elkaroui@criteo.com](mailto:n.elkaroui@criteo.com)

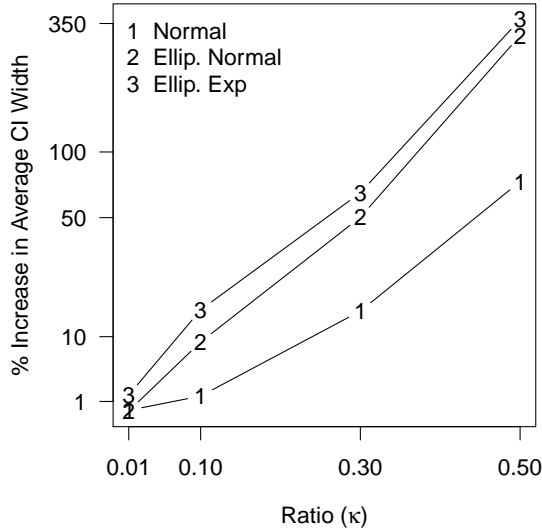


Figure 3: **Comparison of width of 95% confidence intervals of  $e_1' \hat{\beta}_\rho$  for  $L_2$  loss:**  $\rho(x) = x^2/2$ ;  $e_1$  is the first canonical basis vector in  $\mathbb{R}^p$ ; y-axis is the percent increase of the average confidence interval width based on simulation ( $n = 500$ ), as compared to exact theoretical result for least squares; the percent increase is plotted against the ratio  $\kappa = p/n$  (x-axis). Shown are three different choices in simulating the entries of the design matrix  $X$ : (1) Normal:  $X_i \stackrel{iid}{\sim} \mathcal{N}(0, \text{Id}_p)$  (2) Ellip. Normal:  $X_i = \lambda_i Z_i$  with  $\lambda_i \stackrel{iid}{\sim} N(0, 1)$  and independently  $Z_i \stackrel{iid}{\sim} \mathcal{N}(0, \text{Id}_p)$  and (3) Ellip. Exp:  $X_i = \lambda_i Z_i$  with  $\lambda_i \stackrel{iid}{\sim} \text{Exp}(\sqrt{2})$ . The errors  $\epsilon_i$ 's are i.i.d  $\mathcal{N}(0, 1)$

# MULTISCALE ANALYSIS OF WAVE PROPAGATION IN RANDOM MEDIA

JOSSELIN GARNIER

## Abstract

Wave propagation in random media can be studied by multiscale and stochastic analysis. We review some recent advances and their applications. In particular, in a physically relevant regime of separation of scales, wave propagation is governed by a Schrödinger-type equation driven by a Brownian field. We study the associated moment equations and describe the propagation of coherent and incoherent waves. We quantify the scintillation of the wave and the fluctuations of the Wigner distribution. These results make it possible to introduce and characterize correlation-based imaging methods.

## 1 Wave propagation in random media

In many wave propagation scenarios the medium is not constant, but varies in a complicated fashion on a spatial scale that is small compared to the total propagation distance. This is the case for wave propagation through the turbulent atmosphere, the Earth's crust, the ocean, and complex biological tissue for instance. If one aims to use transmitted or reflected waves for communication or imaging purposes it is important to characterize how such microstructure affects and corrupts the wave.

Motivated by the situation described above we consider wave propagation through time-independent complex media with a spatially varying index of refraction. Typically we cannot expect to know the index of refraction pointwise so we model it as a realization of a random process. When the index of refraction is a random process, the wave field is itself a random process and we are interested in how the statistics of the random medium affects the statistics of the wave field. The analysis of wave propagation in random media has a long history. It was first dealt with phenomenological models such as the radiative transfer theory. The first mathematical papers were written in the 60's by Keller [1964] who connected radiative transport theory and random wave equations. In the review presented at

the ICM by Papanicolaou [1998], the main focus was on wave transport and localization (when random inhomogeneities are strong enough to trap wave energy in a bounded region). The book Fouque, Garnier, Papanicolaou, and Sølna [2007] gives the state of the art about wave propagation in randomly layered media, that is a situation mathematically tractable and physically relevant (especially in geophysics). In the recent years several new features have emerged.

1) First statistical stability has become a central issue. Indeed the modeling of a complex medium by a random medium involves ensemble averages. In some circumstances, such as the turbulent atmosphere or the ocean, the medium may change (slowly) in time so that ensemble averages can be experimentally achieved. This is not the case in other configurations, such as seismology, in which the Earth is not moving although it can be considered as a realization of a random medium to model uncertainty and lack of information. It is then important to look for statistically stable quantities, that is to say, quantities that depend on the statistics of the random medium that can be known or estimated, but not on the particular realization that is inaccessible. To quantify statistical stability, variance calculations are required, which are based on high-order moment analysis.

2) Motivated by statistical stability analysis and time-reversal experiments for waves in random media, new methods for communication and imaging have been introduced that are based on wave field correlations. The understanding and analysis of these methods again require high-order moments calculations.

3) Following the analysis of wave field correlations, it was understood that information about the medium could be extracted from wave correlations even when the illumination is provided not by controlled sources, but by uncontrolled, opportunistic or ambient noise sources.

A common feature of these recent developments is the analysis and use of wave field correlations that reveal very rich information. Modern imaging techniques such as seismic interferometry (see N. M. Shapiro, Campillo, Stehly, and Ritzwoller [2005], Schuster [2009], and Wapenaar, Slob, Snieder, and Curtis [2010]) or coherent interferometric imaging (see Borcea, Papanicolaou, and Tsogka [2005, 2006]) correlate wave field traces that have been corrupted by the microstructure of the medium and use their space-time correlation function for imaging.

**1.1 Multiscale analysis.** In its most common form, the analysis of wave propagation in random media consists in studying the wave field solution of the scalar time-harmonic wave equation (called the Helmholtz equation) with a randomly heterogeneous index of refraction. Even though the scalar wave equation is simple and linear, the relation between the statistics of the index of refraction and the statistics of the wave field is highly non-trivial. In order to simplify and understand this relation, one can carry out a multiscale

analysis that will transform the random Helmholtz equation into a mathematically tractable yet physically relevant problem. This analysis is based on a separation of scales technique and limit theorems (homogenization, diffusion-approximation, ...), in the framework set forth by [Asch, Kohler, Papanicolaou, Postel, and White \[1991\]](#). The wave propagation problem can, indeed, be characterized by several length scales: the typical wavelength (which depends on the source), the correlation radius of the medium, the typical propagation distance. The bandwidth of the source and the relative amplitude of the medium fluctuations may also play a role. Different scaling regimes (corresponding to different physical configurations) can be analyzed when certain ratios between these length scales go to zero or infinity. They lead to tractable and relatively easy to interpret asymptotic results. Typically, the solution of the random Helmholtz equation can be shown to converge to the solution of a deterministic or stochastic partial differential equation driven by a Brownian field as in the situation addressed in [Section 2](#). Stochastic calculus can then be used to compute quantities of interest.

In the random travel time model, which is a special high-frequency regime in which the wavelength is much smaller than the correlation radius of the medium, the fluctuations of the medium affect only the phase of the wave, which satisfies a random eikonal equation [Borcea, Garnier, Papanicolaou, and Tsogka \[2011\]](#) and [Tatarskii \[1961\]](#). In the random paraxial model in which the wavelength is smaller than the correlation radius of the medium, backscattering can be neglected but there is significant lateral scattering as the wave advances over long propagation distances and the wave field satisfies a random Schrödinger-type equation [Tappert \[1977\]](#) and [Uscinski \[1977\]](#). In the randomly layered regime, the medium is only varying along the longitudinal direction (along the propagation axis), there is significant backscattering and the plane wave mode amplitudes satisfy a system of ordinary random differential equations [Fouque, Garnier, Papanicolaou, and Sølna \[2007\]](#). In the radiative transport regime, in which the wavelength is of the same order as the correlation radius of the medium, the angular distribution of the mean wave energy satisfies a transport equation [Ryzhik, Papanicolaou, and Keller \[1996\]](#) and [Sato and Fehler \[1998\]](#).

In this review paper we consider a scaling regime corresponding to long-range high-frequency beam propagation and small-scale medium fluctuations giving negligible backscattering. This is the so-called white-noise paraxial regime, as described by the Itô–Schrödinger model, which is presented in [Section 2](#). This model is a simplification of the random Helmholtz equation since it corresponds to an evolution problem, but yet in the regime that we consider it describes the propagated field in a weak sense in that it gives the correct statistical structure of the wave field. The Itô–Schrödinger model can be derived rigorously from the random Helmholtz equation by a separation of scales technique in the high-frequency regime (see [Bailly, Clouet, and Fouque \[1996\]](#) in the case of a randomly layered medium and [Garnier and Sølna \[2008, 2009a,b\]](#) in the case of a three-dimensional



random medium). It models many situations, for instance laser beam propagation [Strohbehn \[1978\]](#), underwater acoustics [Tappert \[1977\]](#), or migration problems in geophysics [Claerbout \[1985\]](#). The Itô–Schrödinger model makes it possible to use Itô’s stochastic calculus, which in turn enables the closure of the hierarchy of moment equations [Fouque, Papanicolaou, and Samuelides \[1998\]](#) and [Ishimaru \[1997\]](#). The equations for the first-order and second-order moments are easy to solve. The equation for the fourth-order moments is difficult and only approximations or numerical solutions have been available for a long time (see [Fante \[1975\]](#), [Tatarskii \[1971\]](#), and [Uscinski \[1985\]](#) and [Ishimaru \[1997, Sec. 20.18\]](#)). In a special scaling regime, it is, however, possible to derive expressions for the fourth-order moments as presented in [Section 3](#).

The results on the higher-order statistics of the wave field open the mathematical analysis of important problems. Below we discuss a few applications, from the understanding of physical conjectures such as star scintillation to the design of efficient correlation-based imaging schemes in complex media. We believe that many more problems than those mentioned here will benefit from the results regarding the statistics of the wave field. In fact, enhanced transducer technology and sampling schemes allow for using finer aspects of the wave field involving second- and fourth-order moments and in such complex cases a rigorous mathematical analysis is important to support, complement, or sometimes disprove, statements based on physical intuition alone.

**1.2 A few conjectures.** Star scintillation is a well-known paradigm, related to the observation that the irradiance of a star fluctuates due to interaction of the light with the turbulent atmosphere. Experimental observations indicate that the statistical distribution of the irradiance is exponential, with the irradiance being the square magnitude of the complex wave field. In the physical literature it is a well-accepted conjecture that the statistics of the complex wave field becomes circularly symmetric complex Gaussian when the wave propagates through the turbulent atmosphere [Valley and Knepp \[1976\]](#) and [Yakushkin \[1978\]](#), so that the irradiance is the sum of the squares of two independent real Gaussian random variables, which has chi-square distribution with two degrees of freedom, that is an exponential distribution. The mathematical proof of this conjecture has been obtained in randomly layered media [Fouque, Garnier, Papanicolaou, and Sølna \[2007, Chapter 9\]](#) but is still incomplete in three-dimensional random media [Fouque, Papanicolaou, and Samuelides \[1998\]](#) and [Garnier and Sølna \[2014, 2016\]](#). In [Section 4](#) we report results for the fourth-order moments that are consistent with the Gaussian conjecture.

Certain functionals of the wave field carry information about the medium and can be characterized in some specific regimes [Bal \[2004\]](#), [Bal and Pinaud \[2010\]](#), [Fannjiang \[2006\]](#), and [Papanicolaou, Ryzhik, and Sølna \[2007\]](#). For instance, the Wigner distribution (a transform in time-frequency analysis) is known to be convenient to study the

solution of Schrödinger equation [Gérard, Markowich, Mauser, and Poupaud \[1997\]](#) and [Ryzhik, Papanicolaou, and Keller \[1996\]](#). An important issue is the so-called statistical stability property: we look for functionals that become deterministic in the considered scaling regime and that depend only on the statistics of the random medium and not on the particular realization. The conjecture is that this can happen for well chosen functionals in the limit of rapid decorrelation of the medium fluctuations. In [Komorowski, Peszat, and Ryzhik \[2009\]](#) and [Komorowski and Ryzhik \[2012\]](#) the authors consider a situation with rapidly fluctuating random medium fluctuations in which the Wigner distribution is statistically stable. As shown in [Bal \[2004\]](#), however, the statistical stability also depends on the initial data and can be lost for very rough initial data even with a high lateral diversity as considered there. In [Section 5](#) we present a detailed and quantitative analysis of the stability of the Wigner distribution and derive an explicit expression of the coefficient of variation of the smoothed Wigner distribution as a function of the smoothing parameters, in the general situation in which the standard deviation can be of the same order as the mean. This is a realistic scenario, which is not too deep into a statistical stabilization situation, but which gives partly coherent but fluctuating wave functionals. These results make it possible to quantify such fluctuations and how their magnitudes can be controlled by optimal smoothing of the Wigner distribution.

**1.3 Applications to communication and imaging.** The understanding of the statistics of the wave field is very important for applications to communication and imaging. Many studies are driven by practical considerations or experimental observations.

1) The time-reversal experiments carried out by M. Fink and his group have motivated many theoretical developments [Fink, Cassereau, Derode, Prada, Roux, Tanter, Thomas, and Wu \[2000\]](#). These experiments are based on the use of a special device called a time-reversal mirror (TRM). A TRM is an array of transducers, that is to say, an array of sensors that can be used as sources and as receivers. A time-reversal experiment has two steps. In the first step, the TRM is used as an array of receivers. A wave is transmitted by a point source far away from the TRM and is recorded by the TRM. In the second step, the TRM is used as an array of sources. It transmits the time-reversed recorded signals. The main observations are that i) the wave refocuses on the original source location, ii) refocusing is enhanced when the medium is randomly scattering, and iii) the time-reversed refocused wave is statistically stable, in the sense that the profile of the focal spot depends on the statistical properties of the random medium, but not on its particular realization. The phenomenon of focusing enhancement has been analyzed quantitatively [Blomgren, Papanicolaou, and Zhao \[2002\]](#), [Papanicolaou, Ryzhik, and Sølna \[2004\]](#), and [Fouque, Garnier, Papanicolaou, and Sølna \[2007\]](#). Statistical stability of time-reversal refocusing for broadband pulses is usually qualitatively proved by the fact that the time-reversed

refocused wave is the superposition of many frequency-dependent components that are uncorrelated, which gives the self-averaging property by the law of large numbers [Blomgren, Papanicolaou, and Zhao \[2002\]](#) and [Papanicolaou, Ryzhik, and Sølna \[2004\]](#). For narrow-band pulses the analysis of the statistical stability phenomenon involves the evaluation of a fourth-order moment of the Green's function of the random wave equation. This problem has been addressed in [Ishimaru, Jaruwatanadilok, and Kuga \[2007\]](#) by using the circular complex Gaussian assumption without rigorous justification. It is, however, possible to prove that the fourth-order moments satisfy Isserlis formula (i.e. they can be expressed in terms of sums of products of second-order moments) and to give a detailed analysis of the statistical stability of the time-reversed refocused wave [Garnier and Sølna \[2016\]](#).

2) Wavefront-shaping-based schemes [Vellekoop and Mosk \[2007\]](#) have attracted a lot of attention in recent years. The primary goal is to focus monochromatic light through a layer of strongly scattering material. This is a challenging problem as multiple scattering of waves scrambles the transmitted light into random interference patterns called speckle patterns [Goodman \[2000\]](#). By using a spatial light modulator (SLM) before the scattering medium (a device that can modulate the intensity and/or the phase profile of the light beam), it is possible to focus light as first demonstrated in [Vellekoop and Mosk \[2007\]](#). Indeed, the elements of the SLM can impose prescribed phase shifts, and an optimization scheme makes it possible to choose the phase shifts so as to maximize the intensity transmitted at one target point behind the scattering medium. The optimal phase shifts are the opposite phases of the field emitted by a point source at the target point and recorded in the plane of the SLM [Mosk, Lagendijk, Lerosey, and Fink \[2012\]](#). In other words, the wavefront-shaping optimization procedure is equivalent to phase conjugation or time reversal. Moreover, it has been shown that the speckle memory effect [Feng, Kane, Lee, and Stone \[1988\]](#) and [Freund, Rosenbluh, and Feng \[1988\]](#) allows to focus on a neighboring point close to the original target point [Vellekoop and Mosk \[2007\]](#), which opens the way to the transmission of spatial patterns [Popoff, Lerosey, Fink, Boccaro, and Gigan \[2010\]](#). This phenomenon can be quantified by fourth-order moments calculations as shown in [Garnier and Sølna \[2017\]](#).

3) In imaging in complex media, when the propagation distance is large enough, the cross correlations of the recorded signals play an important role. This is because the mean (or coherent) field vanishes while the correlations carry information about the medium through which the waves propagate (see [Section 3](#) in the random paraxial regime). The mathematical analysis aims at two goals. First one needs to understand how information about the medium is encoded in the wave field correlations. Second one needs to determine how this information can be extracted in a statistically stable way. This requires detailed fourth-order moment calculations. These calculations help determining imaging functions that can operate in scattering media (see [Borcea, Garnier, Papanicolaou, and Tsogka \[2011\]](#) and [Garnier \[2016\]](#) for instance).

4) Intensity correlations is a recently proposed scheme for communication and imaging in the optical regime. Indeed, in optics only intensities (i.e. the square moduli of the complex envelopes of the wave fields) can be recorded. Intensity correlation-based imaging is a promising scheme for communication and imaging through relatively strong clutter. By using the correlation of the intensity or speckle for different incoming angles, or different positions of the source, or different output angles, one can get spatial information about the source [Newman and Webb \[2012\]](#). The idea of using the information about the statistical structure of speckle to enhance signaling is very interesting and corroborates the idea that modern schemes for communication and imaging require a mathematical theory for analysis of high-order moments.

5) In conventional imaging the waves are generated by an active array of sources and after propagation through the medium they are recorded by an array of receivers (that can be collocated or not with the array of sources). In passive imaging, only receiver arrays are used and the illumination is provided by unknown, uncontrolled, asynchronous, or opportunistic sources. From a theoretical point of view the cross correlations between the recorded signals carry information about the medium through which the waves propagate, as we explain in [Section 6](#). Therefore they play an important role in passive imaging and they can be used for travel-time tomography and reflector imaging [Garnier and Panicolaou \[2016\]](#). From an applied point of view the emergence of correlation-based imaging using ambient seismic noise has had a profound impact in seismology, as can be seen in the work of M. Campillo and co-workers [N. M. Shapiro, Campillo, Stehly, and Ritzwoller \[2005\]](#). The use of seismograms generated by earthquakes was previously the only way to image the Earth. With correlation-based imaging, the seismic noise recorded by a distributed network of sensors on the surface of the Earth can provide a lot of information about its structure. Beyond seismology, there are many new, emerging areas for correlation-based imaging methods, in passive synthetic aperture radar or in optical speckle intensity correlations for communications and imaging. We introduce and explain one of these methods called ghost imaging in [Section 7](#).

## 2 The white-noise paraxial model

In this section we describe how to derive the mathematically tractable Itô-Schrödinger model from the wave equation in a random medium. We consider the three-dimensional scalar wave equation:

$$(1) \quad \frac{1}{c^2(\vec{x})} \frac{\partial^2 u}{\partial t^2}(t, \vec{x}) - \Delta_{\vec{x}} u(t, \vec{x}) = F(t, \vec{x}), \quad t \in \mathbb{R}, \quad \vec{x} \in \mathbb{R}^3.$$

Here the source emits a time-harmonic signal with frequency  $\omega$  and it is localized in the plane  $z = 0$ :

$$(2) \quad F(t, \vec{x}) = \delta(z) f(\mathbf{x}) e^{-i\omega t}, \text{ with } \vec{x} = (\mathbf{x}, z) \in \mathbb{R}^2 \times \mathbb{R},$$

and the speed of propagation is spatially heterogeneous

$$(3) \quad \frac{1}{c^2(\vec{x})} = \frac{1}{c_o^2} (1 + \mu(\vec{x})),$$

where  $\mu$  is a zero-mean stationary random process with ergodic properties.

The time-harmonic field  $\hat{u}$  such that  $u(t, \vec{x}) = \hat{u}(\vec{x}) e^{-i\omega t}$  is solution of the random Helmholtz equation

$$(\partial_z^2 + \Delta_\perp) \hat{u} + \frac{\omega^2}{c_o^2} (1 + \mu(\mathbf{x}, z)) \hat{u} = -\delta(z) f(\mathbf{x}),$$

where  $\Delta_{\vec{x}} = \Delta_\perp + \partial_z^2$ . The function  $\hat{\phi}$  (slowly-varying envelope of a plane wave going along the  $z$ -axis) defined by

$$(4) \quad \hat{u}(\mathbf{x}, z) = \frac{i c_o}{2\omega} e^{i \frac{\omega z}{c_o}} \hat{\phi}(\mathbf{x}, z)$$

satisfies

$$(5) \quad \partial_z^2 \hat{\phi} + \left( 2i \frac{\omega}{c_o} \partial_z \hat{\phi} + \Delta_\perp \hat{\phi} + \frac{\omega^2}{c_o^2} \mu(\mathbf{x}, z) \hat{\phi} \right) = 2i \frac{\omega}{c_o} \delta(z) f(\mathbf{x}).$$

In the paraxial regime “ $\lambda \ll \ell_c, r_o \ll z$ ” (which means, the wavelength  $\lambda = 2\pi c_o/\omega$  is much smaller than the correlation radius  $\ell_c$  of the medium and the radius  $r_o$  of the source, which are themselves much smaller than the propagation distance  $z$ ) the forward-scattering approximation in direction  $z$  is valid and  $\hat{\phi}$  satisfies the Itô-Schrödinger equation [Garnier and Sølna \[2009a\]](#)

$$(6) \quad d_z \hat{\phi} = \frac{i c_o}{2\omega} \Delta_\perp \hat{\phi} dz + \frac{i\omega}{2c_o} \hat{\phi} \circ dB(\mathbf{x}, z), \quad \hat{\phi}(z = 0, \mathbf{x}) = f(\mathbf{x}),$$

where  $\circ$  stands for the Stratonovich integral,  $B(\mathbf{x}, z)$  is a Brownian field, that is a Gaussian process with mean zero and covariance

$$(7) \quad \mathbb{E}[B(\mathbf{x}, z) B(\mathbf{x}', z')] = \gamma(\mathbf{x} - \mathbf{x}') \min(z, z') \text{ with } \gamma(\mathbf{x}) = \int_{\mathbb{R}} \mathbb{E}[\mu(\mathbf{0}, 0) \mu(\mathbf{x}, z)] dz.$$

Remark: Existence, uniqueness and continuity of solutions of the Itô-Schrödinger model (6) are established in [Dawson and Papanicolaou \[1984\]](#). The proof of the convergence of the solution to (5) to the solution to (6) is in [Garnier and Sølna \[2009a\]](#). A first

proof in the case of layered media (i.e. when  $\mu \equiv \mu(z)$ ) can be found in [Bailly, Clouet, and Fouque \[1996\]](#). More precisely, the paraxial regime “ $\lambda \ll \ell_c, r_o \ll z$ ” corresponds to the scaled regime

$$\omega \rightarrow \frac{\omega}{\varepsilon^4}, \quad \mu(\mathbf{x}, z) \rightarrow \varepsilon^3 \mu\left(\frac{\mathbf{x}}{\varepsilon^2}, \frac{z}{\varepsilon^2}\right), \quad f(\mathbf{x}) \rightarrow f\left(\frac{\mathbf{x}}{\varepsilon^2}\right),$$

where  $\varepsilon$  is a small dimensionless parameter, and the convergence in  $\mathcal{C}([0, \infty), L^2(\mathbb{R}^2))$  (or in  $\mathcal{C}([0, \infty), H^k(\mathbb{R}^2))$ ) of the solution of the scaled version of (5) to the solution of the Itô-Schrödinger Equation (6) holds in distribution when  $\varepsilon \rightarrow 0$ . This result requires strong mixing properties for the random process  $\mu$ . If, however, the process  $\mu$  has long-range properties, in the sense that  $\mathbb{E}[\mu(\mathbf{x}, z)\mu(\mathbf{x}', z')] = r(z-z')\gamma(\mathbf{x}-\mathbf{x}')$  with  $r(z) \sim c_\alpha|z|^{-\alpha}$  as  $|z| \rightarrow +\infty$  and  $\alpha \in (0, 1)$ , then, under appropriate technical and scaling assumptions [Gomez and Pinaud \[2017\]](#), the limiting equation is the fractional Itô-Schrödinger model (6) in which  $B$  is a fractional Brownian field with Hurst index  $H = 1 - \alpha/2 \in (1/2, 1)$ , i.e. a Gaussian field with mean zero and covariance

$$\mathbb{E}[B(\mathbf{x}, z)B(\mathbf{x}', z')] = \gamma(\mathbf{x} - \mathbf{x}') \frac{c_\alpha}{2H(2H-1)} (z^{2H} + z'^{2H} - |z - z'|^{2H}).$$

In this case the stochastic integral in (6) can be understood as a generalized Stieljes integral (as  $H > 1/2$ ).

### 3 Statistics of the wave field

In this section we describe how to compute the moments of the wave field. By Itô’s formula and (6), the coherent (or mean) wave satisfies the Schrödinger equation with homogeneous damping (for  $z > 0$ ):

$$(8) \quad \partial_z \mathbb{E}[\hat{\phi}] = \frac{ic_o}{2\omega} \Delta_\perp \mathbb{E}[\hat{\phi}] - \frac{\omega^2 \gamma(\mathbf{0})}{8c_o^2} \mathbb{E}[\hat{\phi}],$$

and therefore  $\mathbb{E}[\hat{\phi}(\mathbf{x}, z)] = \hat{\phi}_0(\mathbf{x}, z) \exp(-z/Z_{\text{sca}})$ , where  $\hat{\phi}_0$  is the solution in the homogeneous medium. The coherent wave amplitude decays exponentially with the propagation distance and the characteristic decay length is the *scattering mean free path*  $Z_{\text{sca}}$ :

$$(9) \quad Z_{\text{sca}} = \frac{8c_o^2}{\gamma(\mathbf{0})\omega^2}.$$

This result shows that any coherent imaging or communication method

fails in random media when the propagation distance is larger than the scattering mean free path [Garnier and Papanicolaou \[2016\]](#).

The mean Wigner distribution defined by

$$(10) \quad W_m(x, \xi, z) = \int_{\mathbb{R}^2} \exp(-i\xi \cdot y) \mathbb{E} \left[ \hat{\phi}\left(x + \frac{y}{2}, z\right) \bar{\hat{\phi}}\left(x - \frac{y}{2}, z\right) \right] dy$$

is the angularly-resolved mean wave energy density (the bar stands for complex conjugation). By Itô's formula and (6), it solves the *radiative transport equation*

$$(11) \quad \partial_z W_m + \frac{c_o}{\omega} \xi \cdot \nabla_x W_m = \frac{\omega^2}{4(2\pi)^2 c_o^2} \int_{\mathbb{R}^2} \hat{\gamma}(\kappa) [W_m(\xi - \kappa) - W_m(\xi)] d\kappa,$$

starting from  $W_m(x, \xi, z=0) = W_0(x, \xi)$ , the Wigner distribution of the initial field  $f$ .  $\hat{\gamma}$  is the Fourier transform of  $\gamma$  and determines the scattering cross section of the radiative transport equation. This result shows that the fields observed at nearby points are correlated and their correlations contain information about the medium. Accordingly, one should use local cross correlations for imaging and communication in random media [Borcea, Papanicolaou, and Tsogka \[2005\]](#) and [Borcea, Garnier, Papanicolaou, and Tsogka \[2011\]](#).

In order to quantify the stability of correlation-based imaging methods, one needs to evaluate variances of empirical correlations, which involves the fourth-order moment:

$$(12) \quad M_4(q_1, q_2, r_1, r_2, z) = \mathbb{E} \left[ \hat{\phi}\left(\frac{r_1 + r_2 + q_1 + q_2}{2}, z\right) \hat{\phi}\left(\frac{r_1 - r_2 + q_1 - q_2}{2}, z\right) \right. \\ \left. \times \bar{\hat{\phi}}\left(\frac{r_1 + r_2 - q_1 - q_2}{2}, z\right) \bar{\hat{\phi}}\left(\frac{r_1 - r_2 - q_1 + q_2}{2}, z\right) \right].$$

By Itô's formula and (6), it satisfies the Schrödinger-type equation

$$(13) \quad \partial_z M_4 = \frac{ic_o}{\omega} (\nabla_{r_1} \cdot \nabla_{q_1} + \nabla_{r_2} \cdot \nabla_{q_2}) M_4 + \frac{\omega^2}{4c_o^2} U_4(q_1, q_2, r_1, r_2) M_4,$$

with the generalized potential

$$(14) \quad U_4(q_1, q_2, r_1, r_2) = \gamma(q_2 + q_1) + \gamma(q_2 - q_1) + \gamma(r_2 + q_1) + \gamma(r_2 - q_1) \\ - \gamma(q_2 + r_2) - \gamma(q_2 - r_2) - 2\gamma(\mathbf{0}).$$

These moment equations have been known for a long time [Ishimaru \[1997\]](#). Recently it was shown [Garnier and Sølna \[2016\]](#) that in the regime “ $\lambda \ll \ell_c \ll r_o \ll z$ ” the fourth-order moment can be expressed explicitly in terms of the function  $\gamma$ . These results can be used to address a wide range of applications in imaging and communication.

## 4 The scintillation index

In this section we study the intensity fluctuations of the wave field solution of (6) and characterize the scintillation index which quantifies the relative intensity fluctuations. It is a fundamental quantity associated for instance with light propagation through the atmosphere Ishimaru [1997]. It is defined as the square coefficient of variation of the intensity Ishimaru [ibid., Eq. (20.151)]:

$$(15) \quad S(\mathbf{x}, z) = \frac{\mathbb{E}[|\hat{\phi}(\mathbf{x}, z)|^4] - \mathbb{E}[|\hat{\phi}(\mathbf{x}, z)|^2]^2}{\mathbb{E}[|\hat{\phi}(\mathbf{x}, z)|^2]^2}.$$

When the spatial profile of the source (or the initial beam) has a Gaussian profile,

$$(16) \quad f(\mathbf{x}) = \exp\left(-\frac{|\mathbf{x}|^2}{r_o^2}\right),$$

and when “ $\lambda \ll \ell_c \ll r_o \ll z$ ”, the behavior of the scintillation index can be described as follows Garnier and Sølna [2016].

**Proposition 4.1.** *Let us consider the following form of the covariance function of the medium fluctuations:*

$$\gamma(\mathbf{x}) = \gamma(\mathbf{0})\tilde{\gamma}\left(\frac{\mathbf{x}}{\ell_c}\right),$$

with  $\tilde{\gamma}(\mathbf{0}) = 1$  and the width of the function  $\tilde{\gamma}$  is of order one. In the regime “ $\lambda \ll \ell_c \ll r_o \ll z$ ” the scintillation index (15) has the following expression:

$$(17) \quad S(\mathbf{x}, z) = 1 - \frac{\exp\left(-\frac{2|\mathbf{x}|^2}{r_o^2}\right)}{\left|\frac{1}{4\pi} \int_{\mathbb{R}^2} \exp\left(\frac{2z}{Z_{\text{sca}}} \int_0^1 \tilde{\gamma}\left(\mathbf{v} \frac{z}{Z_c} s\right) ds - \frac{|\mathbf{v}|^2}{4} + i \mathbf{v} \cdot \frac{\mathbf{x}}{r_o}\right) d\mathbf{v}\right|^2}.$$

In fact this result follows from the complete expressions of the second moment of the intensity and the second moment of the field that are given in Garnier and Sølna [ibid.]. The scintillation index at the beam center  $\mathbf{x} = \mathbf{0}$  is a function of  $z/Z_{\text{sca}}$  and  $z/Z_c$  only, where  $Z_c = \omega r_o \ell_c / c_o$  is the typical propagation distance for which diffractive effects are of order one, as shown in Garnier and Sølna [2009a, Eq. (4.4)]. It is interesting to note that, even if the propagation distance is larger than the scattering mean free path, the scintillation index can be smaller than one if  $Z_c$  is small compared to  $Z_{\text{sca}}$ .

In order to get more explicit expressions that facilitate interpretation of the results let us assume that  $\gamma(\mathbf{x})$  is smooth and can be expanded as

$$(18) \quad \gamma(\mathbf{x}) = \gamma(\mathbf{0})\left(1 - \frac{|\mathbf{x}|^2}{\ell_c^2} + o\left(\frac{|\mathbf{x}|^2}{\ell_c^2}\right)\right), \quad \mathbf{x} \rightarrow \mathbf{0}.$$



When scattering is strong in the sense that the propagation distance is larger than the scattering mean free path  $z \gg Z_{\text{sca}}$ , the expressions of the second moments of the field and of the intensity can be simplified:

$$\begin{aligned}\Gamma^{(2)}(\mathbf{x}, \mathbf{y}, z) &:= \mathbb{E} \left[ \hat{\phi} \left( \mathbf{x} + \frac{\mathbf{y}}{2}, z \right) \bar{\hat{\phi}} \left( \mathbf{x} - \frac{\mathbf{y}}{2}, z \right) \right] = \frac{r_o^2}{R_z^2} \exp \left( -\frac{|\mathbf{x}|^2}{R_z^2} - \frac{|\mathbf{y}|^2}{\rho_z^2} + i \frac{\omega \gamma(\mathbf{0}) z^2 \mathbf{x} \cdot \mathbf{y}}{2c_o \ell_c^2 R_z^2} \right), \\ \Gamma^{(4)}(\mathbf{x}, \mathbf{y}, z) &:= \mathbb{E} \left[ \left| \hat{\phi} \left( \mathbf{x} + \frac{\mathbf{y}}{2}, z \right) \right|^2 \left| \hat{\phi} \left( \mathbf{x} - \frac{\mathbf{y}}{2}, z \right) \right|^2 \right] = |\Gamma^{(2)}(\mathbf{x}, \mathbf{0}, z)|^2 + |\Gamma^{(2)}(\mathbf{x}, \mathbf{y}, z)|^2,\end{aligned}$$

where the beam radius  $R_z$  is

$$(19) \quad R_z^2 = r_o^2 + \frac{\gamma(\mathbf{0})z^3}{3\ell_c^2}$$

and the correlation radius of the beam  $\rho_z$  is

$$(20) \quad \rho_z^2 = \frac{4c_o^2 \ell_c^2}{\omega^2 \gamma(\mathbf{0}) z} \frac{r_o^2 + \frac{\gamma(\mathbf{0})z^3}{3\ell_c^2}}{r_o^2 + \frac{\gamma(\mathbf{0})z^3}{12\ell_c^2}}.$$

Note that the fourth-order moments satisfy the Isserlis formula (i.e. they can be expressed in terms of sums of products of second-order moments), and therefore the scintillation index  $S(\mathbf{x}, z)$  is equal to one. This observation is consistent with the physical intuition that, in the strongly scattering regime  $z/Z_{\text{sca}} \gg 1$ , the wave field is conjectured to have zero-mean complex circularly symmetric Gaussian statistics, and therefore the intensity is expected to have exponential (or Rayleigh) distribution [Fante \[1975\]](#) and [Ishimaru \[1997\]](#).

## 5 Fluctuations of the Wigner distribution

In this section we give an explicit characterization of the signal-to-noise ratio of the Wigner distribution. The Wigner distribution of the wave field is defined by

$$(21) \quad W(\mathbf{x}, \boldsymbol{\xi}, z) = \int_{\mathbb{R}^2} \exp(-i \boldsymbol{\xi} \cdot \mathbf{y}) \hat{\phi} \left( \mathbf{x} + \frac{\mathbf{y}}{2}, z \right) \bar{\hat{\phi}} \left( \mathbf{x} - \frac{\mathbf{y}}{2}, z \right) d\mathbf{y}.$$

It can be interpreted as the angularly-resolved wave energy density (note, however, that it is real-valued but not always non-negative valued). Its expectation satisfies the radiative transport [Equation \(11\)](#). It is known that the Wigner distribution is not statistically stable, and that it is necessary to smooth it (that is to say, to convolve it with a smoothing kernel) to get a quantity that can be measured in a statistically stable way (that is to say, the smoothed Wigner distribution for one typical realization is approximately equal to its expected value)

Bal [2004] and Papanicolaou, Ryzhik, and Sølna [2007]. Our goal in this section is to quantify this statistical stability.

Let us consider two positive parameters  $r_s$  and  $\xi_s$  and define the smoothed Wigner distribution:

$$(22) \quad W_s(\mathbf{x}, \boldsymbol{\xi}, z) = \frac{1}{(2\pi)^2 r_s^2 \xi_s^2} \iint_{\mathbb{R}^4} W(\mathbf{x} - \mathbf{x}', \boldsymbol{\xi} - \boldsymbol{\xi}', z) \exp\left(-\frac{|\mathbf{x}'|^2}{2r_s^2} - \frac{|\boldsymbol{\xi}'|^2}{2\xi_s^2}\right) d\mathbf{x}' d\boldsymbol{\xi}'.$$

If we denote by  $\rho_z$  the correlation radius of the field (given by (20) in the strongly scattering regime), we may anticipate that  $r_s$  and  $1/\xi_s$  should be of the order of  $\rho_z$  to ensure averaging. The coefficient of variation  $C_s$  of the smoothed Wigner distribution, which characterizes its statistical stability, is defined by:

$$(23) \quad C_s(\mathbf{x}, \boldsymbol{\xi}, z) = \frac{\sqrt{\mathbb{E}[W_s(\mathbf{x}, \boldsymbol{\xi}, z)^2] - \mathbb{E}[W_s(\mathbf{x}, \boldsymbol{\xi}, z)]^2}}{\mathbb{E}[W_s(\mathbf{x}, \boldsymbol{\xi}, z)]}.$$

An exact expression of the coefficient of variation of the smoothed Wigner distribution can be derived in the regime “ $\lambda \ll \ell_c \ll r_o \ll z$ ” Garnier and Sølna [2016]. It involves four-dimensional integrals and it is complicated to interpret it. This expression becomes simple in the strongly scattering regime  $z \gg Z_{\text{sca}}$ . We then get the following expression for the coefficient of variation Garnier and Sølna [ibid.].

**Proposition 5.1.** *In the regime “ $\lambda \ll \ell_c \ll r_o \ll z$ ”, if additionally  $z \gg Z_{\text{sca}}$  and  $\gamma$  can be expanded as (18), then the coefficient of variation of the smoothed Wigner distribution (22) satisfies:*

$$(24) \quad C_s(\mathbf{x}, \boldsymbol{\xi}, z)^2 = \frac{\frac{1}{\xi_s^2 \rho_z^2} + 1}{\frac{4r_s^2}{\rho_z^2} + 1},$$

where  $\rho_z$  is the correlation radius (20).

Note that the coefficient of variation becomes independent of  $\mathbf{x}$  and  $\boldsymbol{\xi}$ . Equation (24) is a simple enough formula to help determining the smoothing parameters  $\xi_s$  and  $r_s$  that are needed to reach a given value for the coefficient of variation:

- For  $2\xi_s r_s = 1$ , we have  $C_s(\mathbf{x}, \boldsymbol{\xi}, z) = 1$ .
- For  $2\xi_s r_s < 1$  (resp.  $> 1$ ) we have  $C_s(\mathbf{x}, \boldsymbol{\xi}, z) > 1$  (resp.  $< 1$ ); in other words, the smoothed Wigner transform can be considered as statistically stable as soon as  $2\xi_s r_s > 1$ .

The critical value  $r_s = 1/(2\xi_s)$  is indeed special. In this case, the smoothed Wigner distribution (22) can be written as the double convolution of the Wigner distribution  $W$  of the random field  $\hat{\phi}(\cdot, z)$  (defined by (21)) with the Wigner distribution

$$(25) \quad W_g(\mathbf{x}, \boldsymbol{\xi}) = \int_{\mathbb{R}^2} \exp(-i \boldsymbol{\xi} \cdot \mathbf{y}) \hat{\phi}_g(\mathbf{x} + \frac{\mathbf{y}}{2}) \overline{\hat{\phi}_g(\mathbf{x} - \frac{\mathbf{y}}{2})} d\mathbf{y}$$

of the Gaussian state

$$(26) \quad \hat{\phi}_g(\mathbf{x}) = \exp(-\xi_s^2 |\mathbf{x}|^2),$$

since we have

$$W_g(\mathbf{x}, \boldsymbol{\xi}) = \frac{2\pi}{\xi_s^2} \exp\left(-2\xi_s^2 |\mathbf{x}|^2 - \frac{|\boldsymbol{\xi}|^2}{2\xi_s^2}\right),$$

and therefore

$$(27) \quad W_s(\mathbf{x}, \boldsymbol{\xi}, z) = \frac{4\xi_s^2}{(2\pi)^3} \iint_{\mathbb{R}^4} W(\mathbf{x} - \mathbf{x}', \boldsymbol{\xi} - \boldsymbol{\xi}', z) W_g(\mathbf{x}', \boldsymbol{\xi}') d\mathbf{x}' d\boldsymbol{\xi}',$$

for  $r_s = 1/(2\xi_s)$ . It is known that the convolution of a Wigner distribution with a kernel that is itself the Wigner distribution of a function (such as  $W_g$ ) is nonnegative real valued (the smoothed Wigner distribution obtained with the Gaussian  $W_g$  is called Husimi function) Cartwright [1976] and Manfredi and Feix [2000]. This can be shown easily in our case as the smoothed Wigner distribution can be written as

$$(28) \quad W_s(\mathbf{x}, \boldsymbol{\xi}, z) = \frac{2\xi_s^2}{\pi} \left| \int_{\mathbb{R}^2} \exp(i \boldsymbol{\xi} \cdot \mathbf{x}') \overline{\hat{\phi}_g(\mathbf{x}') \hat{\phi}(\mathbf{x} - \mathbf{x}', z)} d\mathbf{x}' \right|^2,$$

for  $r_s = 1/(2\xi_s)$ . From this representation formula of  $W_s$  valid for  $r_s = 1/(2\xi_s)$ , we can see that it is the square modulus of a linear functional of  $\hat{\phi}(\cdot, z)$ . The physical intuition that  $\hat{\phi}(\cdot, z)$  has circularly symmetric complex Gaussian statistics in strongly scattering media then predicts that  $W_s(\mathbf{x}, \boldsymbol{\xi}, z)$  should have an exponential distribution, because the sum of the squares of two independent real-valued Gaussian random variables has an exponential distribution. This is indeed consistent with our theoretical finding that  $C_s = 1$  for  $r_s = 1/(2\xi_s)$ .

If  $r_s > 1/(2\xi_s)$ , by observing that

$$\exp\left(-\frac{|\mathbf{x}|^2}{2r_s^2}\right) = \int_{\mathbb{R}^2} \Psi(\mathbf{x} - \mathbf{x}') \exp(-2\xi_s^2 |\mathbf{x}'|^2) d\mathbf{x}',$$

where the function  $\Psi$  is defined by

$$(29) \quad \Psi(\mathbf{x}) = \frac{8\xi_s^4 r_s^2}{\pi(4\xi_s^2 r_s^2 - 1)} \exp\left(-\frac{2\xi_s^2 |\mathbf{x}|^2}{(4\xi_s^2 r_s^2 - 1)}\right),$$

we find that the smoothed Wigner distribution (22) can be expressed as:

(30)

$$W_s(z, \mathbf{x}, \boldsymbol{\xi}) = \int_{\mathbb{R}^2} \Psi(\mathbf{x} - \mathbf{x}') \left( \frac{2\xi_s^2}{\pi} \left| \int_{\mathbb{R}^2} \exp(i \boldsymbol{\xi} \cdot \mathbf{x}'') \overline{\hat{\phi}_g(\mathbf{x}'')} \hat{\phi}(\mathbf{x}' - \mathbf{x}'', z) d\mathbf{x}'' \right|^2 \right) d\mathbf{x}',$$

for  $r_s > 1/(2\xi_s)$ . From this representation formula for  $W_s$  valid for  $r_s > 1/(2\xi_s)$ , we can see that it is nonnegative valued and that it is a local average of (28), which has a unit coefficient of variation in the strongly scattering regime. That is why the coefficient of variation of the smoothed Wigner distribution is smaller than one when  $r_s > 1/(2\xi_s)$ .

## 6 Green's function estimation with noise sources

The previous section has explained that the cross correlation of the wave field (or equivalently the Wigner distribution which is its local Fourier transform) is a convenient way to look at waves propagating in complex media. In this section we show that it is also useful to study waves emitted by unknown noise sources. In Section 7 we will combine both situations and show that incoherent illumination can provide new paradigms and offer new opportunities for imaging in complex media.

We aim to exhibit a relation between the cross correlation of the signals emitted by noise sources and recorded by two sensors and the Green's function between the sensors. The Green's function is the signal recorded by the second sensor when the first one transmits a short (Dirac) pulse. The relation between the cross correlation of the recorded noise signals and the Green's function is very important. Indeed, in standard imaging, an array of sources transmits waves that are recorded by an array of receivers, that can be collocated with the array of sources (called an active array). One then gets the matrix of Green's functions from the sources to the receivers, that can be processed for imaging purposes. In ambient noise imaging, an array of receivers (called a passive array) records the waves emitted by ambient noise sources. Under favorable circumstances, the matrix of cross correlations of the signals recorded by the receivers gives the matrix of Green's functions between the receivers. In other words, the passive array data have been transformed into active array data ! The fact that ambient noise illumination allows for Green's function estimation and subsequent imaging opens fascinating perspectives. In seismology, it means that information is present in the seismic noise recorded by networks of seismometers and that it can be extracted by computing cross correlations.

We consider the solution  $u$  of the scalar wave equation in a three-dimensional inhomogeneous medium with propagation speed  $c(\vec{x})$ :

$$(31) \quad \frac{1}{c^2(\vec{x})} \frac{\partial^2 u}{\partial t^2} - \Delta_{\vec{x}} u = n(t, \vec{x}).$$

The term  $n(t, \vec{x})$  models noise sources. It is a zero-mean stationary (in time) random process with autocorrelation function

$$(32) \quad \langle n(t_1, \vec{y}_1) n(t_2, \vec{y}_2) \rangle = F(t_2 - t_1) \Gamma(\vec{y}_1, \vec{y}_2).$$

Here  $\langle \cdot \rangle$  stands for statistical average with respect to the distribution of the noise sources.

The time distribution of the noise sources is characterized by the correlation function  $F(t_2 - t_1)$ , which is a function of  $t_2 - t_1$  only because of time stationarity. The Fourier transform  $\hat{F}(\omega)$  of the time correlation function  $F(t)$  is a nonnegative, even, real-valued function proportional to the power spectral density of the sources:

$$(33) \quad \hat{F}(\omega) = \int_{\mathbb{R}} F(t) e^{i\omega t} dt.$$

The spatial distribution of the noise sources is characterized by the autocovariance function  $\Gamma(\vec{y}_1, \vec{y}_2)$ . In this review paper we assume that the random process  $n$  is delta-correlated in space:

$$(34) \quad \Gamma(\vec{y}_1, \vec{y}_2) = K(\vec{y}_1) \delta(\vec{y}_1 - \vec{y}_2),$$

although it is possible to address correlated noise sources, as shown in [Bardos, Garnier, and Papanicolaou \[2008\]](#) and [Garnier and Papanicolaou \[2010\]](#). The nonnegative valued function  $K$  characterizes the spatial support of the sources.

The solution of the wave [Equation \(31\)](#) has the integral representation:

$$u(t, \vec{x}) = \int_{\mathbb{R}^3} \int_{\mathbb{R}} n(t - s, \vec{y}) G(s, \vec{x}, \vec{y}) ds d\vec{y},$$

where  $G(t, \vec{x}, \vec{y})$  is the time-dependent Green's function, that is to say, the fundamental solution of the three-dimensional scalar wave equation:

$$\frac{1}{c^2(\vec{x})} \frac{\partial^2 G}{\partial t^2} - \Delta_{\vec{x}} G = \delta(t) \delta(\vec{x} - \vec{y}).$$

The empirical cross correlation of the signals recorded at  $\vec{x}_1$  and  $\vec{x}_2$  for an integration time  $T$  is

$$(35) \quad C_T(\tau, \vec{x}_1, \vec{x}_2) = \frac{1}{T} \int_0^T u(t, \vec{x}_1) u(t + \tau, \vec{x}_2) dt.$$

It is a statistically stable quantity, in the sense that for a large integration time  $T$ , the empirical cross correlation  $C_T$  is independent of the realization of the noise sources and it is equal to its expectation. This is stated in the following proposition [Garnier and Papanicolaou \[2009\]](#).

**Proposition 6.1.** 1. *The expectation of the empirical cross correlation  $C_T$  (with respect to the statistics or distribution of the sources) is independent of  $T$ :*

$$(36) \quad \langle C_T(\tau, \vec{x}_1, \vec{x}_2) \rangle = C^{(1)}(\tau, \vec{x}_1, \vec{x}_2),$$

where the statistical cross correlation  $C^{(1)}$  is given by

$$(37) \quad C^{(1)}(\tau, \vec{x}_1, \vec{x}_2) = \frac{1}{2\pi} \int_{\mathbb{R}^3} \int_{\mathbb{R}} \hat{F}(\omega) K(\vec{y}) \overline{\hat{G}(\omega, \vec{x}_1, \vec{y})} \hat{G}(\omega, \vec{x}_2, \vec{y}) e^{-i\omega\tau} d\omega d\vec{y},$$

and  $\hat{G}(\omega, \vec{x}, \vec{y})$  is the time-harmonic Green's function (i.e. the Fourier transform of  $G(t, \vec{x}, \vec{y})$ ).

2. *The empirical cross correlation  $C_T$  is a self-averaging quantity:*

$$(38) \quad C_T(\tau, \vec{x}_1, \vec{x}_2) \xrightarrow{T \rightarrow \infty} C^{(1)}(\tau, \vec{x}_1, \vec{x}_2),$$

in probability (with respect to the distribution of the sources).

Equation (37) holds whatever the spatial support of the sources but it does not give a simple relation between the cross correlation of the recorded noise signals and the Green's function. Such a relation emerges when the spatial support of the sources is extended. We give below a simple statement when the noise sources are located on the surface of a ball that encloses both the inhomogeneous region and the sensors, located at  $\vec{x}_1$  and  $\vec{x}_2$ .

**Proposition 6.2.** *We assume that the medium is homogeneous outside the ball  $B(\mathbf{0}, D)$  with center  $\mathbf{0}$  and radius  $D$ , and that the sources are localized with a uniform density on the surface of the sphere  $\partial B(\mathbf{0}, L)$  with center  $\mathbf{0}$  and radius  $L$ . If  $L \gg D$ , then for any  $\vec{x}_1, \vec{x}_2 \in B(\mathbf{0}, D)$ , we have*

$$(39) \quad \frac{\partial}{\partial \tau} C^{(1)}(\tau, \vec{x}_1, \vec{x}_2) = -\frac{c_o}{2} [F * G(\tau, \vec{x}_1, \vec{x}_2) - F * G(-\tau, \vec{x}_1, \vec{x}_2)].$$

Proposition 6.2 can be found in Schuster [2009] and Wapenaar, Slob, Snieder, and Curtis [2010]. The proof is based on the Helmholtz-Kirchhoff identity, which results from the second Green's identity and the Sommerfeld radiation condition. It is simple and gives the desired result quickly, but it requires a full aperture illumination. Multiscale analysis reveals that full aperture illumination is sufficient but not necessary for the cross correlation to be related to the Green's function Garnier and Papanicolaou [2016].

Proposition 6.2 shows that, when the noise sources surround the region of interest, then the lag-time derivative of the cross correlation of the signals recorded at two observation points is the Green's function between these two points, up to a convolution (in time) with the time covariance function of the noise sources and a symmetrization (which means

that we get in fact the causal and the anti-causal Green's functions). We can present an illustration of this result when the medium is homogeneous with background velocity  $c_o$  and the Green's function is  $G(t, \vec{x}, \vec{y}) = \frac{1}{4\pi|\vec{x}-\vec{y}|} \delta(t - \frac{|\vec{x}-\vec{y}|}{c_o})$ . If  $F(t) = -g''(t)$  with  $g(t) = \exp(-t^2/4)$  (the prime stands for derivative), so that  $\hat{F}(\omega) = 2\sqrt{\pi}\omega^2 \exp(-\omega^2)$ , then the cross correlation is by (39):

$$C^{(1)}(\tau, \vec{x}_1, \vec{x}_2) = \frac{c_o}{8\pi|\vec{x}_1 - \vec{x}_2|} \left[ g' \left( \tau - \frac{|\vec{x}_1 - \vec{x}_2|}{c_o} \right) - g' \left( \tau + \frac{|\vec{x}_1 - \vec{x}_2|}{c_o} \right) \right].$$

The autocorrelation function is

$$C^{(1)}(\tau, \vec{x}_1, \vec{x}_1) = -\frac{1}{4\pi} g''(\tau).$$

These two functions can be seen in [Figure 1](#): the autocorrelation function  $C^{(1)}(\tau, \vec{x}_1, \vec{x}_1)$  has the form of the second derivative of a Gaussian function, and the cross correlation  $C^{(1)}(\tau, \vec{x}_1, \vec{x}_j)$ ,  $j \geq 2$ , has two symmetric peaks with the form of the first derivative of a Gaussian function and centered at the travel times  $\pm|\vec{x}_1 - \vec{x}_j|/c_o$ . From the imaging point of view, this means that the travel times between the sensors can be estimated from the cross correlations of the noise signals and subsequently background velocity estimation can be carried out tomographically.

## 7 An example of incoherent wave imaging in complex media: Ghost imaging

In this section we study an imaging method called ghost imaging introduced in the optics literature. It illustrates the fact that incoherent illumination can be beneficial for correlation-based imaging in complex media. The experimental set-up proposed in [Valencia, Scarcelli, D'Angelo, and Shih \[2005\]](#) and [J. H. Shapiro and Boyd \[2012\]](#) is plotted in [Figure 2](#). The waves are emitted by a noise (or partially coherent) source. A beam splitter is used to generate two wave beams from this source:

- the “reference beam”, labeled ①, propagates through a homogeneous or scattering medium up to a high-resolution detector that measures the spatially resolved transmitted intensity.
- the “signal beam”, labeled ②, propagates through a homogeneous or scattering medium and interacts with an object to be imaged. The total transmitted intensity is measured by a bucket detector that measures the spatially integrated transmitted intensity.

This method is called ghost imaging because the high-resolution detector does not see the object to be imaged, and nevertheless a high-resolution image of the object is obtained by cross-correlating the two measured intensity signals. From the previous section we can anticipate that something may indeed happen when cross correlating these signals because

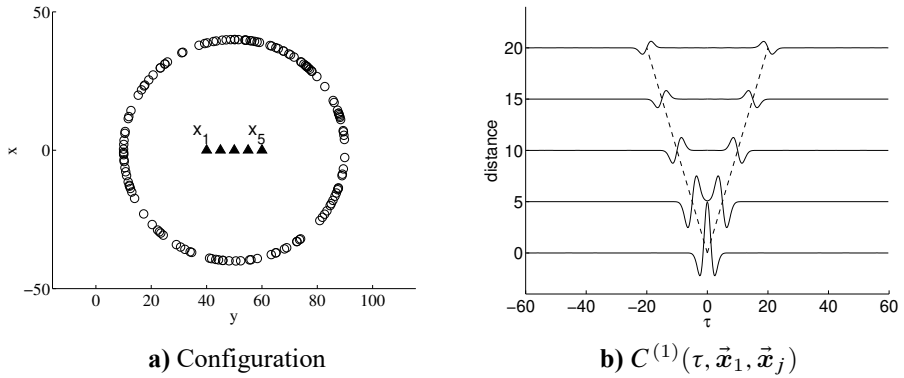


Figure 1: The configuration is shown in the plane  $(xy)$  in Figure a: the circles are the noise sources and the triangles are the sensors (the distance between two successive sensors is 5). The sources are randomly distributed on the surface of the three-dimensional sphere with center at  $(0, 50, 0)$  and radius 40. Figure b shows the cross correlation  $\tau \rightarrow C^{(1)}(\tau, \vec{x}_1, \vec{x}_j)$  between the pairs of sensors  $(\vec{x}_1, \vec{x}_j)$ ,  $j = 1, \dots, 5$ , versus the distance  $|\vec{x}_j - \vec{x}_1|$ . For  $j \geq 2$  the values have been multiplied by 6 as the autocorrelation function ( $j = 1$ ) takes larger values than the cross correlation functions ( $j \geq 2$ ). Here  $c_o = 1$ .



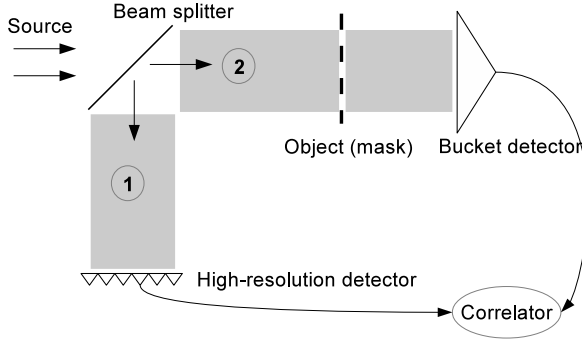


Figure 2: The ghost imaging setup. A partially coherent source is split into two beams by a beam splitter. The reference beam (labeled ①) does not interact with the object and its intensity is measured by a high-resolution detector. The signal beam (labeled ②) interacts with the object to be imaged and its intensity is measured by a bucket (single-pixel) detector. From [Garnier \[2016\]](#).

the cross correlation should be related to the Green's function between the plane of the high-resolution detector in reference path ① (or the corresponding plane just before the object in the signal path ②) and the plane of the bucket detector in ②. The relation is not, however, clear when only spatially-integrated intensities are measured, which requires a detailed analysis.

The object to be imaged is a mask modeled by a transmission function  $\mathcal{T}(x)$ . In the experiments, the object is typically a double slit [J. H. Shapiro and Boyd \[2012\]](#). The source is located in the plane  $z = 0$ . The propagation distance from the source to the high-resolution detector in the reference path ① is  $L$ . The propagation distance from the source to the object in the signal path ② is  $L$  as well, and the propagation distance from the object to the bucket detector is  $L_0$ . In each path the scalar wave  $(t, \vec{x}) \mapsto u_j(t, \vec{x})$ ,  $j = 1, 2$ , satisfies the scalar wave equation:

$$(40) \quad \frac{1}{c_j(\vec{x})^2} \frac{\partial^2 u_j}{\partial t^2} - \Delta_{\vec{x}} u_j = n(t, \mathbf{x}) \delta(z),$$

where  $c_j(\vec{x})$  is the speed of propagation in the medium corresponding to the  $j$ th path and the forcing term  $(t, \mathbf{x}) \mapsto n(t, \mathbf{x})$  models the source (identical for the two waves).

In the ghost experiment the source is typically a laser beam passed through a rotating glass diffuser [Valencia, Scarcelli, D'Angelo, and Shih \[2005\]](#) and [J. H. Shapiro and Boyd](#)

[2012]. We model it as

$$(41) \quad n(t, \mathbf{x}) = f(t, \mathbf{x})e^{-i\omega_o t} + c.c.,$$

where *c.c.* stands for complex conjugate,  $\omega_o$  is the carrier frequency, and  $f(t, \mathbf{x})$  is the complex-valued slowly varying envelope, whose Fourier transform (in time) has a typical width that is much smaller than  $\omega_o$ . It is assumed to be a complex-valued, zero-mean stationary Gaussian process with the covariance function:

$$(42) \quad \langle f(t, \mathbf{x}) \overline{f(t', \mathbf{x}')} \rangle = F(t - t')\Gamma(\mathbf{x}, \mathbf{x}'),$$

with  $F(0) = 1$  (with real-valued functions  $F$  and  $\Gamma$ ). The width of the Fourier transform of  $F$  is much smaller than  $\omega_o$ . Note that the modeling is similar to the one used for correlation-based imaging using ambient noise sources as discussed in the previous section. In this framework the scalar wave fields  $u_j$ ,  $j = 1, 2$ , can be written in the form

$$u_j(t, \vec{\mathbf{x}}) = v_j(t, \vec{\mathbf{x}})e^{-i\omega_o t} + c.c..$$

The detectors measure the intensities, i.e. the square moduli of the slowly varying envelopes  $v_j$ ,  $j = 1, 2$ . More exactly, the quantity that is measured by the high-resolution detector is the spatially-resolved intensity in the plane  $z = L$  of the reference path ①:

$$(43) \quad I_1(t, \mathbf{x}) = |v_1(t, (\mathbf{x}, L))|^2.$$

The quantity that is measured by the bucket detector is the spatially-integrated intensity in the plane  $z = L + L_0$  of the signal path ②:

$$(44) \quad I_2(t) = \int_{\mathbb{R}^2} |v_2(t, (\mathbf{x}, L + L_0))|^2 d\mathbf{x}.$$

These two quantities are correlated and their cross correlation defines the ghost imaging function:

$$(45) \quad \mathbb{C}_T(\mathbf{x}) = \frac{1}{T} \int_0^T I_1(t, \mathbf{x}) I_2(t) dt - \left[ \frac{1}{T} \int_0^T I_1(t, \mathbf{x}) dt \right] \left[ \frac{1}{T} \int_0^T I_2(t) dt \right].$$

We consider the partially coherent case:

$$(46) \quad \Gamma(\mathbf{x}, \mathbf{x}') = I_o \exp \left( -\frac{|\mathbf{x} + \mathbf{x}'|^2}{2r_o^2} - \frac{|\mathbf{x} - \mathbf{x}'|^2}{2\rho_o^2} \right),$$

in which the source is assumed to have a Gaussian spatial profile with radius  $r_o$  and a local Gaussian correlation function with radius  $\rho_o$ . This model is called Gaussian-Schell in the

optics literature [Mandel and Wolf \[1995\]](#). Note that we always have  $r_o \geq \rho_o$  (because  $\Gamma$  is a positive kernel). The limit case  $\rho_o \rightarrow 0$  corresponds to fully incoherent illumination: the field is delta-correlated in space. The limit case  $\rho_o = r_o$  in which

$$\Gamma(\mathbf{x}, \mathbf{x}') = I_o \exp\left(-\frac{|\mathbf{x}|^2}{r_o^2} - \frac{|\mathbf{x}'|^2}{r_o^2}\right)$$

corresponds to fully coherent illumination: the spatial profile of the field is deterministic (up to a random multiplicative factor) and has a Gaussian form with radius  $r_o$ . The following proposition shows that the ghost imaging function gives an image of the square transmission function  $\mathcal{T}^2$  up to an integral operator whose kernel can be identified [Garnier \[2016\]](#).

**Proposition 7.1.** *If  $T \rightarrow \infty$ , then the ghost imaging function converges in probability to*

$$(47) \quad \mathcal{C}(\mathbf{x}) = \int_{\mathbb{R}^2} H(\mathbf{x}, \mathbf{y}) \mathcal{T}(\mathbf{y})^2 d\mathbf{y},$$

with the kernel given by

$$(48) \quad \begin{aligned} H(\mathbf{x}, \mathbf{y}) = & \frac{I_o^2 \rho_o^4 r_o^4}{2^8 \pi^2 L^4} \int_{\mathbb{R}^2} d\boldsymbol{\alpha} \int_{\mathbb{R}^2} d\boldsymbol{\beta} \exp\left(-(|\boldsymbol{\alpha}|^2 + |\boldsymbol{\beta}|^2)\left(1 + \frac{\omega_o^2 r_o^2 \rho_o^2}{4c_o^2 L^2}\right)\right) \\ & \times \exp\left(-i \frac{\omega_o}{c_o L} (\rho_o(\mathbf{x} + \mathbf{y}) \cdot \boldsymbol{\alpha} + r_o(\mathbf{x} - \mathbf{y}) \cdot \boldsymbol{\beta})\right) \\ & \times \exp\left(-\frac{\omega_o^2 L}{4c_o^2} \int_0^1 2\gamma(\mathbf{0}) - \gamma((\rho_o \boldsymbol{\alpha} + r_o \boldsymbol{\beta})s) - \gamma((\rho_o \boldsymbol{\alpha} - r_o \boldsymbol{\beta})s) ds\right). \end{aligned}$$

If the medium is homogeneous along the two paths  $\gamma = 0$ , then the kernel is

$$(49) \quad H(\mathbf{x}, \mathbf{y}) = \frac{I_o^2 \rho_o^2 r_o^2 c_o^4}{64 \omega_o^4 \rho_{\text{gi}0}^2 R_{\text{gi}0}^2} \exp\left(-\frac{|\mathbf{x} - \mathbf{y}|^2}{2\rho_{\text{gi}0}^2} - \frac{|\mathbf{x} + \mathbf{y}|^2}{2R_{\text{gi}0}^2}\right),$$

with

$$(50) \quad \rho_{\text{gi}0}^2 = \frac{2c_o^2 L^2}{\omega_o^2 r_o^2} + \frac{\rho_o^2}{2}, \quad R_{\text{gi}0}^2 = \frac{2c_o^2 L^2}{\omega_o^2 \rho_o^2} + \frac{r_o^2}{2}.$$

[Equation \(50\)](#) shows that the resolution of the ghost imaging function is improved when the source becomes less coherent (i.e., when  $\rho_o$  decreases, the radius of the convolution kernel  $\rho_{\text{gi}0}$  decreases). It also shows that imaging is possible provided the object to be imaged (i.e. the support of the transmission function) is within the disk with radius  $R_{\text{gi}0}$ . This radius increases when the source becomes less coherent (i.e., when  $\rho_o$  decreases,  $R_{\text{gi}0}$  increases).

If the medium is random along the two paths and scattering is strong, in the sense that the propagation distance is larger than the scattering mean free path  $L/Z_{\text{sca}} \gg 1$ , then

$$(51) \quad H(\mathbf{x}, \mathbf{y}) = \frac{I_o^2 \rho_o^2 r_o^4 c_o^4}{64 \omega_o^4 \rho_{\text{gi1}}^2 R_{\text{gi1}}^2} \exp \left( -\frac{|\mathbf{x} - \mathbf{y}|^2}{2 \rho_{\text{gi1}}^2} - \frac{|\mathbf{x} + \mathbf{y}|^2}{2 R_{\text{gi1}}^2} \right),$$

with

$$(52) \quad \rho_{\text{gi1}}^2 = \frac{2c_o^2 L^2}{\omega_o^2 r_o^2} + \frac{\rho_o^2}{2} + \frac{8c_o^2 L^3}{3\omega_o^2 Z_{\text{sca}} \ell_c^2}, \quad R_{\text{gi1}}^2 = \frac{2c_o^2 L^2}{\omega_o^2 \rho_o^2} + \frac{r_o^2}{2} + \frac{8c_o^2 L^3}{3\omega_o^2 Z_{\text{sca}} \ell_c^2},$$

and the correlation radius of the medium  $\ell_c$  is defined as in (18).

In the partially coherent case  $\rho_o \leq r_o$ , formula (52) shows that the resolution is improved when the source becomes less coherent but it is degraded by scattering. Moreover, the radius of the region that can be imaged increases when the source becomes less coherent and when scattering becomes stronger. In other words, scattering degrades the resolution of the ghost imaging function, but it enhances the region that can be imaged.

In the limit case of a fully incoherent source  $\rho_o \rightarrow 0$  we have  $\rho_{\text{gi1}}^2 \rightarrow \rho_{\text{gi}}^2 := \frac{2c_o^2 L^2}{\omega_o^2 r_o^2} + \frac{8c_o^2 L^3}{3\omega_o^2 Z_{\text{sca}} \ell_c^2}$  and  $R_{\text{gi1}}^2 \rightarrow +\infty$ , which shows that the integral operator is then a convolution with a Gaussian kernel with radius  $\rho_{\text{gi}}$ .

In the limit case of a fully coherent source  $\rho_o = r_o$ , then  $\rho_{\text{gi}}^2 = R_{\text{gi}}^2$  and

$$H(\mathbf{x}, \mathbf{y}) = \frac{I_o^2 r_o^4 c_o^4}{64 \omega_o^4 R_{\text{gi}}^4} \exp \left( -\frac{|\mathbf{x}|^2}{R_{\text{gi}}^2} - \frac{|\mathbf{y}|^2}{R_{\text{gi}}^2} \right),$$

with  $R_{\text{gi}}^2 = \frac{2c_o^2 L^2}{\omega_o^2 r_o^2} + \frac{r_o^2}{2} + \frac{8c_o^2 L^3}{3\omega_o^2 Z_{\text{sca}} \ell_c^2}$ , which has a separable form. In this case we do not get any image of the transmission function and the imaging function has a Gaussian form with width  $R_{\text{gi}}$  whatever the form of the transmission function. This confirms that the incoherence (or partial coherence) of the source is the key ingredient for ghost imaging.

**Acknowledgments.** I am very grateful to L. Borcea, J.-P. Fouque, G. Papanicolaou, K. Sølna, and C. Tsogka for longstanding collaborations over the years on the subjects discussed here.

## References

Mark Asch, Werner Kohler, George Papanicolaou, Marie Postel, and Benjamin White (1991). “Frequency content of randomly scattered signals”. *SIAM Rev.* 33.4, pp. 519–625. MR: [1137513](#) (cit. on p. 2897).

- Florence Bailly, Jean-François Clouet, and Jean-Pierre Fouque (1996). “Parabolic and Gaussian white noise approximation for wave propagation in random media”. *SIAM J. Appl. Math.* 56.5, pp. 1445–1470. MR: [1409128](#) (cit. on pp. [2897](#), [2903](#)).
- Guillaume Bal (2004). “On the self-averaging of wave energy in random media”. *Multi-scale Model. Simul.* 2.3, pp. 398–420. MR: [2111700](#) (cit. on pp. [2898](#), [2899](#), [2907](#)).
- Guillaume Bal and Olivier Pinaud (2010). “Dynamics of wave scintillation in random media”. *Comm. Partial Differential Equations* 35.7, pp. 1176–1235. MR: [2753633](#) (cit. on p. [2898](#)).
- Claude Bardos, Josselin Garnier, and George Papanicolaou (2008). “Identification of Green’s functions singularities by cross correlation of noisy signals”. *Inverse Problems* 24.1, pp. 015011, 26. MR: [2384770](#) (cit. on p. [2910](#)).
- Peter Blomgren, George Papanicolaou, and Hongkai Zhao (2002). “Super-resolution in time-reversal acoustics”. *The Journal of the Acoustical Society of America* 111.1, pp. 230–248 (cit. on pp. [2899](#), [2900](#)).
- Liliana Borcea, Josselin Garnier, George Papanicolaou, and Chrysoula Tsogka (2011). “Enhanced statistical stability in coherent interferometric imaging”. *Inverse Problems* 27.8, pp. 085004, 33. MR: [2819946](#) (cit. on pp. [2897](#), [2900](#), [2904](#)).
- Liliana Borcea, George Papanicolaou, and Chrysoula Tsogka (2005). “Interferometric array imaging in clutter”. *Inverse Problems* 21.4, pp. 1419–1460. MR: [2158118](#) (cit. on pp. [2896](#), [2904](#)).
- (2006). “Adaptive interferometric imaging in clutter and optimal illumination”. *Inverse Problems* 22.4, pp. 1405–1436. MR: [2249471](#) (cit. on p. [2896](#)).
- Nancy D. Cartwright (1976). “A non-negative Wigner-type distribution”. *Physica A: Statistical Mechanics and its Applications* 83.1, pp. 210–212 (cit. on p. [2908](#)).
- Jon F. Claerbout (1985). *Imaging the earth’s interior*. Vol. 1. Blackwell scientific publications Oxford, Palo Alto (cit. on p. [2898](#)).
- Donald Dawson and George Papanicolaou (1984). “A random wave process”. *Appl. Math. Optim.* 12.2, pp. 97–114. MR: [764811](#) (cit. on p. [2902](#)).
- Albert C. Fannjiang (2006). “Self-averaging radiative transfer for parabolic waves”. *C. R. Math. Acad. Sci. Paris* 342.2, pp. 109–114. MR: [2193656](#) (cit. on p. [2898](#)).
- Ronald L. Fante (1975). “Electromagnetic beam propagation in turbulent media”. *Proceedings of the IEEE* 63.12, pp. 1669–1692 (cit. on pp. [2898](#), [2906](#)).
- Shechao Feng, Charles Kane, Patrick A. Lee, and A. Douglas Stone (1988). “Correlations and fluctuations of coherent wave transmission through disordered media”. *Physical review letters* 61.7, pp. 834–837 (cit. on p. [2900](#)).
- Mathias Fink, Didier Cassereau, Arnaud Derode, Claire Prada, Philippe Roux, Mickael Tanter, Jean-Louis Thomas, and François Wu (2000). “Time-reversed acoustics”. *Reports on progress in Physics* 63.12, pp. 1933–1995 (cit. on p. [2899](#)).

- Jean-Pierre Fouque, Josselin Garnier, George Papanicolaou, and Knut Sølna (2007). *Wave propagation and time reversal in randomly layered media*. Vol. 56. Stochastic Modelling and Applied Probability. Springer, New York, pp. xx+612. MR: [2327824](#) (cit. on pp. [2896–2899](#)).
- Jean-Pierre Fouque, George Papanicolaou, and Yann Samuelides (1998). “Forward and Markov approximation: the strong-intensity-fluctuations regime revisited”. *Waves Random Media* 8.3, pp. 303–314. MR: [1633157](#) (cit. on p. [2898](#)).
- Isaac Freund, Michael Rosenbluh, and Shechao Feng (1988). “Memory effects in propagation of optical waves through disordered media”. *Physical review letters* 61.20, pp. 2328–2331 (cit. on p. [2900](#)).
- Josselin Garnier (2016). “Ghost imaging in the random paraxial regime”. *Inverse Probl. Imaging* 10.2, pp. 409–432. MR: [3507901](#) (cit. on pp. [2900](#), [2914](#), [2916](#)).
- Josselin Garnier and George Papanicolaou (2009). “Passive sensor imaging using cross correlations of noisy signals in a scattering medium”. *SIAM J. Imaging Sci.* 2.2, pp. 396–437. MR: [2496063](#) (cit. on p. [2910](#)).
- (2010). “Resolution analysis for imaging with noise”. *Inverse Prob.* 26.7, pp. 074001, 22. MR: [2608011](#) (cit. on p. [2910](#)).
  - (2016). *Passive imaging with ambient noise*. Cambridge University Press, Cambridge, pp. xii+294. MR: [3616275](#) (cit. on pp. [2901](#), [2903](#), [2911](#)).
- Josselin Garnier and Knut Sølna (2008). “Random backscattering in the parabolic scaling”. *J. Stat. Phys.* 131.3, pp. 445–486. MR: [2386572](#) (cit. on p. [2897](#)).
- (2009a). “Coupled paraxial wave equations in random media in the white-noise regime”. *Ann. Appl. Probab.* 19.1, pp. 318–346. MR: [2498680](#) (cit. on pp. [2897](#), [2902](#), [2905](#)).
  - (2009b). “Scaling limits for wave pulse transmission and reflection operators”. *Wave Motion* 46.2, pp. 122–143. MR: [2488979](#) (cit. on p. [2897](#)).
  - (2014). “Scintillation in the white-noise paraxial regime”. *Comm. Partial Differential Equations* 39.4, pp. 626–650. MR: [3178071](#) (cit. on p. [2898](#)).
  - (2016). “Fourth-moment analysis for wave propagation in the white-noise paraxial regime”. *Arch. Ration. Mech. Anal.* 220.1, pp. 37–81. MR: [3458158](#) (cit. on pp. [2898](#), [2900](#), [2904](#), [2905](#), [2907](#)).
  - (2017). “Focusing waves through a randomly scattering medium in the white-noise paraxial regime”. *SIAM J. Appl. Math.* 77.2, pp. 500–519. MR: [3629440](#) (cit. on p. [2900](#)).
- Patrick Gérard, Peter A. Markowich, Norbert J. Mauser, and Frédéric Poupaud (1997). “Homogenization limits and Wigner transforms”. *Comm. Pure Appl. Math.* 50.4, pp. 323–379. MR: [1438151](#) (cit. on p. [2899](#)).
- Christophe Gomez and Olivier Pinaud (2017). “Fractional white-noise limit and paraxial approximation for waves in random media”. *Arch. Ration. Mech. Anal.* 226.3, pp. 1061–1138. MR: [3712278](#) (cit. on p. [2903](#)).

- Joseph W. Goodman (2000). *Statistical optics*. John Wiley & Sons, New York (cit. on p. 2900).
- Akira Ishimaru (1997). *Wave propagation and scattering in random media*. IEEE/OUP Series on Electromagnetic Wave Theory. Reprint of the 1978 original, With a foreword by Gary S. Brown, An IEEE/OUP Classic Reissue. IEEE Press, New York, pp. xxvi+574. MR: [1626707](#) (cit. on pp. 2898, 2904–2906).
- Akira Ishimaru, Sermsak Jaruwatanadilok, and Yasuo Kuga (2007). “Time reversal effects in random scattering media on superresolution, shower curtain effects, and backscattering enhancement”. *Radio Science* 42.6 (cit. on p. 2900).
- Joseph B. Keller (1964). “Stochastic equations and wave propagation in random media”. In: *Proc. Sympos. Appl. Math., Vol. XVI*. Amer. Math. Soc., Providence, R.I., pp. 145–170. MR: [0178638](#) (cit. on p. 2895).
- Tomasz Komorowski, Szymon Peszat, and Leonid Ryzhik (2009). “Limit of fluctuations of solutions of Wigner equation”. *Comm. Math. Phys.* 292.2, pp. 479–510. MR: [2544740](#) (cit. on p. 2899).
- Tomasz Komorowski and Leonid Ryzhik (2012). “Fluctuations of solutions to Wigner equation with an Ornstein-Uhlenbeck potential”. *Discrete Contin. Dyn. Syst. Ser. B* 17.3, pp. 871–914. MR: [2873119](#) (cit. on p. 2899).
- Leonard Mandel and Emil Wolf (1995). *Optical coherence and quantum optics*. Cambridge university press (cit. on p. 2916).
- Giovanni Manfredi and Marc R. Feix (2000). “Entropy and Wigner functions”. *Physical Review E* 62.4, pp. 4665–4674 (cit. on p. 2908).
- Allard P. Mosk, Ad Lagendijk, Geoffroy Lerosey, and Mathias Fink (2012). “Controlling waves in space and time for imaging and focusing in complex media”. *Nature photonics* 6.5, pp. 283–292 (cit. on p. 2900).
- Jason A. Newman and Kevin J. Webb (2012). “Fourier magnitude of the field incident on a random scattering medium from spatial speckle intensity correlations”. *Optics letters* 37.7, pp. 1136–1138 (cit. on p. 2901).
- George Papanicolaou (1998). “Mathematical problems in geophysical wave propagation”. In: *Proceedings of the International Congress of Mathematicians, Vol. I (Berlin, 1998)*. Extra Vol. I, pp. 403–427. MR: [1648040](#) (cit. on p. 2896).
- George Papanicolaou, Leonid Ryzhik, and Knut Sølna (2004). “Statistical stability in time reversal”. *SIAM J. Appl. Math.* 64.4, pp. 1133–1155. MR: [2068663](#) (cit. on pp. 2899, 2900).
- (2007). “Self-averaging from lateral diversity in the Itô-Schrödinger equation”. *Multi-scale Model. Simul.* 6.2, pp. 468–492. MR: [2338491](#) (cit. on pp. 2898, 2907).
- Sébastien Popoff, Geoffroy Lerosey, Mathias Fink, Albert Claude Boccarda, and Sylvain Gigan (2010). “Image transmission through an opaque material”. *Nature Communications* 1.6, pp. 1–5 (cit. on p. 2900).

- Leonid Ryzhik, George Papanicolaou, and Joseph B. Keller (1996). “[Transport equations for elastic and other waves in random media](#)”. *Wave Motion* 24.4, pp. 327–370. MR: [1427483](#) (cit. on pp. [2897](#), [2899](#)).
- Haruo Sato and Michael C. Fehler (1998). *Seismic wave propagation and scattering in the heterogeneous earth*. AIP Series in Modern Acoustics and Signal Processing. American Institute of Physics, New York; Springer-Verlag, New York, pp. xiv+308. MR: [1488700](#) (cit. on p. [2897](#)).
- Gerard Thomas Schuster (2009). *Seismic interferometry*. Vol. 1. Cambridge University Press Cambridge (cit. on pp. [2896](#), [2911](#)).
- Jeffrey H. Shapiro and Robert W. Boyd (2012). “[The physics of ghost imaging](#)”. *Quantum Information Processing* 11.4, pp. 949–993 (cit. on pp. [2912](#), [2914](#)).
- Nikolai M. Shapiro, Michel Campillo, Laurent Stehly, and Michael H. Ritzwoller (2005). “[High-resolution surface-wave tomography from ambient seismic noise](#)”. *Science* 307.5715, pp. 1615–1618 (cit. on pp. [2896](#), [2901](#)).
- John W. Strohbehn (1978). *Laser beam propagation in the atmosphere*. Springer, Berlin (cit. on p. [2898](#)).
- Fred D. Tappert (1977). “The parabolic approximation method”, 224–287. Lecture Notes in Phys., Vol. 70. MR: [0475274](#) (cit. on pp. [2897](#), [2898](#)).
- Valerian I. Tatarskii (1961). *Wave propagation in a turbulent medium*. Translated from the Russian by R. A. Silverman. McGraw-Hill Book Co., Inc., New York-Toronto-London, pp. xiv+285. MR: [0127671](#) (cit. on p. [2897](#)).
- (1971). “The Effects of Turbulent Atmosphere on Wave Propagation”. In: *U.S. Department of Commerce, TT-68-50464, Springfield* (cit. on p. [2898](#)).
- Barry J. Uscinski (1977). *The elements of wave propagation in random media*. McGraw-Hill Companies (cit. on p. [2897](#)).
- (1985). “[Analytical solution of the fourth-moment equation and interpretation as a set of phase screens](#)”. *J. Opt. Soc. Amer. A* 2.12, pp. 2077–2091. MR: [823410](#) (cit. on p. [2898](#)).
- Alejandra Valencia, Giuliano Scarcelli, Milena D’Angelo, and Yanhua Shih (2005). “[Two-photon imaging with thermal light](#)”. *Physical review letters* 94.6, p. 063601 (cit. on pp. [2912](#), [2914](#)).
- George Valley and Dennis Knepp (1976). “[Application of joint Gaussian statistics to interplanetary scintillation](#)”. *Journal of Geophysical Research* 81.25, pp. 4723–4730 (cit. on p. [2898](#)).
- Ivo M. Vellekoop and Allard P. Mosk (2007). “[Focusing coherent light through opaque strongly scattering media](#)”. *Optics letters* 32.16, pp. 2309–2311 (cit. on p. [2900](#)).
- Kees Wapenaar, Evert Slob, Roel Snieder, and Andrew Curtis (2010). “[Tutorial on seismic interferometry: Part 2—Underlying theory and new advances](#)”. *Geophysics* 75.5, A211–A227 (cit. on pp. [2896](#), [2911](#)).



Ivan G. Yakushkin (1978). “[Moments of field propagating in randomly inhomogeneous medium in the limit of saturated fluctuations](#)”. *Radiophysics and Quantum Electronics* 21.8, pp. 835–840 (cit. on p. [2898](#)).

Received 2018-02-09.

JOSSELIN GARNIER  
CENTRE DE MATHÉMATIQUES APPLIQUÉES  
ÉCOLE POLYTECHNIQUE  
91128 PALAISEAU CEDEX  
FRANCE  
[josselin.garnier@polytechnique.edu](mailto:josselin.garnier@polytechnique.edu)

# ASYMPTOTIC EFFICIENCY IN HIGH-DIMENSIONAL COVARIANCE ESTIMATION

VLADIMIR KOLTCHINSKII

## Abstract

We discuss recent results on asymptotically efficient estimation of smooth functionals of covariance operator  $\Sigma$  of a mean zero Gaussian random vector  $X$  in a separable Hilbert space based on  $n$  i.i.d. observations of this vector. We are interested in functionals that are of importance in high-dimensional statistics such as linear forms of eigenvectors of  $\Sigma$  (principal components) as well as in more general functionals of the form  $\langle f(\Sigma), B \rangle$ , where  $f : \mathbb{R} \mapsto \mathbb{R}$  is a sufficiently smooth function and  $B$  is an operator with nuclear norm bounded by a constant. In the case when  $X$  takes values in a finite-dimensional space of dimension  $d \leq n^\alpha$  for some  $\alpha \in (0, 1)$  and  $f$  belongs to Besov space  $B_{\infty,1}^s(\mathbb{R})$  for  $s > \frac{1}{1-\alpha}$ , we develop asymptotically normal estimators of  $\langle f(\Sigma), B \rangle$  with  $\sqrt{n}$  convergence rate and prove asymptotic minimax lower bounds showing their asymptotic efficiency.

## 1 Introduction

Let  $X_1, \dots, X_n$  be i.i.d. random variables sampled from unknown distribution  $P_\theta, \theta \in \Theta$ . Assume that the parameter space  $\Theta$  is a subset of a linear normed space and the goal is to estimate  $f(\theta)$  for a smooth functional  $f : \Theta \mapsto \mathbb{R}$  based on observations  $X_1, \dots, X_n$ . Let  $\mathcal{L}$  be the set of loss functions  $\ell : \mathbb{R} \mapsto \mathbb{R}_+$  such that  $\ell(0) = 0, \ell(-t) = \ell(t), t \in \mathbb{R}$ ,  $\ell$  is convex and increasing on  $\mathbb{R}_+$  and for some  $c > 0, \ell(t) = O(e^{c|t|})$  as  $t \rightarrow \infty$ . Let  $Z$  be a standard normal random variable.

**Definition 1.** An estimator  $T_n = T_n(X_1, \dots, X_n)$  will be called asymptotically efficient with respect to  $\Theta_n \subset \Theta, n \geq 1$  with convergence rate  $\sqrt{n}$  and (limit) variance  $\sigma_f^2(\theta) > 0$

---

The author was supported in part by NSF grants DMS-1509739 and CCF-1523768.

MSC2010: primary 62H12; secondary 62G20, 62H25, 60B20.

Keywords: asymptotic efficiency, sample covariance, bootstrap, effective rank, concentration inequalities, normal approximation.

iff the following properties hold:

$$(1) \quad \sup_{\theta \in \Theta_n} \sup_{x \in \mathbb{R}} \left| \mathbb{P}_\theta \left\{ \frac{n^{1/2}(T_n(X_1, \dots, X_n) - f(\theta))}{\sigma_f(\theta)} \leq x \right\} - \mathbb{P}\{Z \leq x\} \right| \rightarrow 0,$$

for all  $\ell \in \mathfrak{L}$ ,

$$(2) \quad \sup_{\theta \in \Theta_n} \left| \mathbb{E}_\theta \ell \left( \frac{n^{1/2}(T_n(X_1, \dots, X_n) - f(\theta))}{\sigma_f(\theta)} \right) - \mathbb{E} \ell(Z) \right| \rightarrow 0 \text{ as } n \rightarrow \infty$$

and

$$(3) \quad \liminf_{n \rightarrow \infty} \inf_{\tilde{T}_n} \sup_{\theta \in \Theta_n} \frac{n \mathbb{E}_\theta (\tilde{T}_n(X_1, \dots, X_n) - f(\theta))^2}{\sigma_f^2(\theta)} \geq 1$$

with the infimum in (3) being over all estimators  $\tilde{T}_n$ .

A similar definition can be also used for more general models in which the data  $X^{(n)}$  is sampled from a distribution  $P_\theta^{(n)}$ ,  $\theta \in \Theta$  as well as in the case of a sequence of smooth functions  $f_n : \Theta_n \mapsto \mathbb{R}$ .

The idea of asymptotically efficient estimation (initially understood as asymptotically normal estimation with the smallest possible limit variance) goes back to Fisher [1922, 1925]. Fisher conjectured (“Fisher’s program”) that, under suitable regularity of statistical model, the maximal likelihood method would yield asymptotically efficient estimators with the optimal limit variance being the reciprocal of the Fisher information. The difficulties with implementing Fisher’s program became apparent in the early 50s when Hodges developed a well known counterexample of a superefficient estimator in a regular statistical model. The development of contemporary view of asymptotic efficiency is due to several authors, in particular, to Le Cam and Hájek (LeCam [1953] and Hájek [1972]). For regular finite-dimensional models, asymptotically efficient estimators of smooth functions  $f(\theta)$  could be obtained from the maximum likelihood estimator  $\hat{\theta}$  using the Delta Method: for a continuously differentiable function  $f$ ,  $f(\hat{\theta}) - f(\theta) = \langle f'(\theta), \hat{\theta} - \theta \rangle + o_{\mathbb{P}}(n^{-1/2})$ , implying that  $n^{1/2}(f(\hat{\theta}) - f(\theta))$  is asymptotically normal  $N(0; \sigma_f^2(\theta))$  with the limit variance  $\sigma_f^2(\theta) = \langle I(\theta)^{-1} f'(\theta), f'(\theta) \rangle$ ,  $I(\theta)$  being the Fisher information matrix. The optimality of the limit variance is usually proved using convolution and local asymptotic minimax theorems (Hájek, Le Cam). It could be also proved using van Trees inequality (see Gill and Levit [1995]) leading to bounds similar to (3).

Due to slow convergence rates of estimation of infinite-dimensional parameters in nonparametric statistics, it becomes important to identify low-dimensional features of these parameters that admit asymptotically efficient estimation with parametric  $\sqrt{n}$ -rate.

Such features are often represented by smooth functionals of infinite-dimensional parameters. Early references on asymptotically efficient estimation of smooth functionals include [Levit \[1975, 1978\]](#) and [Ibragimov and Khasminskii \[1981\]](#) with a number of further publications for the last decades on estimation of linear, quadratic and more general smooth functionals and with connections to extensive literature on efficiency in semiparametric estimation (see [Bickel, Klaassen, Ritov, and Wellner \[1993\]](#), [Giné and Nickl \[2016\]](#) and references therein). Ibragimov, Nemirovski and Khasminskii in [Ibragimov, Nemirovski, and Khasminskii \[1986\]](#) and Nemirovski in [Nemirovski \[1990, 2000\]](#) systematically studied the problem of estimation of general smooth functionals of unknown parameter of Gaussian shift model. In this model (also known as Gaussian sequence model), the parameter of interest is a “signal”  $\theta \in \Theta$ , where  $\Theta$  is a bounded subset of a separable Hilbert space. Given an orthonormal basis  $\{e_k : k \geq 1\}$  of  $\mathbb{H}$ , the data consists of observations  $X_k = \langle \theta, e_k \rangle + \sigma Z_k, k \geq 1$ , where  $\{Z_k\}$  are i.i.d.  $N(0, 1)$  r.v. and  $\sigma$  is a small parameter characterizing the level of the noise (we will set  $\sigma := n^{-1/2}$ ). In [Ibragimov, Nemirovski, and Khasminskii \[1986\]](#) and [Nemirovski \[1990, 2000\]](#), two different notions of smoothness of a functional  $f$  were used, with control of the derivatives either in the operator norm, or in the Hilbert–Schmidt norm (of multilinear forms). The complexity of estimation problem was characterized by the rate of decay of Kolmogorov diameters of set  $\Theta$  defined as  $d_m(\Theta) := \inf_{L \subset \mathbb{H}, \dim(L) \leq m} \sup_{\theta \in \Theta} \|\theta - P_L \theta\|, m \geq 1$ ,  $P_L$  being the orthogonal projection on subspace  $L$ . Assuming that  $d_m(\Theta) \lesssim m^{-\beta}, m \geq 1$  for some  $\beta > 0$ , it was proved that efficient estimation (with a somewhat different definition of efficiency than [Definition 1](#)) of a smooth functional  $f$  on  $\mathbb{H}$  is possible for smoothness parameter  $s > s(\beta)$ , where  $s(\beta)$  is a threshold depending on the rate of decay  $\beta$  of Kolmogorov diameters. The estimation method was based on Taylor expansions of  $f(\theta)$  around an estimator  $\hat{\theta}$  with an optimal nonparametric rate, which allowed to reduce the problem to estimation of polynomial functions on  $\mathbb{H}$ . [Nemirovski \[1990, 2000\]](#) also proved that efficient estimation is impossible for some functionals  $f$  of smoothness  $s < s(\beta)$ .

More recently, estimation problems for functionals of unknown parameters have been studied in various models of high-dimensional statistics, including semi-parametric efficiency of regularization-based estimators (such as LASSO) [van de Geer, Bühlmann, Ritov, and Dezeure \[2014\]](#), [Javanmard and Montanari \[2014\]](#), [C.-H. Zhang and S. S. Zhang \[2014\]](#), [Janková and van de Geer \[2016\]](#) as well as minimax optimal rates of estimation of special functionals (in particular, linear and quadratic) [Cai and Low \[2005b\]](#), [Cai and Low \[2005a\]](#), [Collier, Comminges, and Tsybakov \[2017\]](#).

In this paper, we are primarily interested in the problem of estimation of smooth functionals of unknown covariance operator  $\Sigma$  based on a sample of size  $n$  of i.i.d. mean zero Gaussian random variables with covariance  $\Sigma$ . In this problem, the maximum likelihood

estimator is the sample covariance  $\hat{\Sigma}$ . By standard Hájek–LeCam theory, plug-in estimator  $h(\hat{\Sigma})$  is an asymptotically efficient estimator of a smooth functional  $h(\Sigma)$  in the finite-dimensional case. The problem of asymptotically efficient estimation of general smooth functionals of covariance operators in high-dimensional setting (when the dimension  $d$  of the space is allowed to grow with the sample size  $n$ ) has not been systematically studied. However, there are many results on asymptotic normality in this type of problems. In the 80s–90s, Girko developed asymptotically normal (but hardly asymptotically efficient) estimators of many special functionals of covariance matrices in high-dimensional setting (see [Girko \[1987\]](#), [Girko \[1995\]](#) and references therein). Central limit theorems for so called linear spectral statistics  $\text{tr}(f(\hat{\Sigma}))$  have been studied in random matrix theory with a number of deep results both in the case of high-dimensional sample covariance (or Wishart matrices) and in other random matrix models such as Wigner matrices, see, e.g., [Bai and Silverstein \[2004\]](#), [Lytova and Pastur \[2009\]](#). However, these results do not have straightforward statistical implications since  $\text{tr}(f(\hat{\Sigma}))$  does not “concentrate” around the corresponding population parameter (with the exception of some special functionals of this form such as *log-determinant*  $\log \det(\Sigma) = \text{tr}(\log \Sigma)$ , for which  $\log \det(\hat{\Sigma})$  (with a simple bias correction) provides an asymptotically normal estimator (see [Girko \[1987\]](#) and [Cai, Liang, and Zhou \[2015\]](#))). More recent references include [Fan, Rigollet, and Wang \[2015\]](#) where optimal error rates in estimation of several special functionals of covariance under sparsity assumptions were studied and [Gao and Zhou \[2016\]](#) where Bernstein-von Mises type theorems for functionals of covariance were proved.

In what follows,  $\mathfrak{B}(\mathbb{H})$  is the space of bounded linear operators in a separable Hilbert space  $\mathbb{H}$ .  $\mathfrak{B}(\mathbb{H})$  is usually equipped with the operator norm denoted by  $\|\cdot\|$ . Let  $\mathfrak{B}_{sa}(\mathbb{H})$  be the subspace of bounded self-adjoint operators. Denote by  $\mathfrak{C}_+(\mathbb{H})$  the cone of self-adjoint positively semi-definite nuclear operators in  $\mathbb{H}$  (the covariance operators). We use notation  $A^*$  for the adjoint operator of  $A$ ,  $\text{rank}(A)$  for the rank of  $A$ ,  $\text{tr}(A)$  for the trace of a trace class operator  $A$ , and  $\|A\|_p$  for the Schatten  $p$ -norm of  $A$  :  $\|A\|_p^p := \text{tr}(|A|^p)$ ,  $|A| = (A^*A)^{1/2}$ ,  $p \in [1, \infty]$ . In particular,  $\|A\|_1$  is the nuclear norm,  $\|A\|_2$  is the Hilbert–Schmidt norm and  $\|A\|_\infty = \|A\|$  is the operator norm of  $A$ . The inner product notation  $\langle \cdot, \cdot \rangle$  is used for the inner product in the underlying Hilbert space  $\mathbb{H}$ , for the Hilbert–Schmidt inner product between the operators and also for linear functionals on the spaces of operators (for instance,  $\langle A, B \rangle$ , where  $A$  is a bounded operator and  $B$  is a nuclear operator, is a value of such a linear functional on the space of bounded operators). Given  $u, v \in \mathbb{H}$ ,  $u \otimes v$  denotes the tensor product of vectors  $u$  and  $v$  :  $(u \otimes v)x := u\langle v, x \rangle$ ,  $x \in \mathbb{H}$ . Notation  $A \leq B$  means that operator  $B - A$  is positively semi-definite.

We also use the following notations: given  $a, b \geq 0$ ,  $a \lesssim b$  means that  $a \leq cb$  for a numerical constant  $c > 0$ ;  $a \gtrsim b$  is equivalent to  $b \lesssim a$ ;  $a \asymp b$  is equivalent to  $a \lesssim b$  and  $b \lesssim a$ . Sometimes, constants in the above relationships depend on some parameter(s).

In such cases, the signs  $\lesssim$ ,  $\gtrsim$  and  $\asymp$  are provided with subscripts:  $a \lesssim_\gamma b$  means that  $a \leq c_\gamma b$  for a constant  $c_\gamma > 0$ .

## 2 Effective rank and estimation of linear functionals of principal components

Let  $X$  be a centered Gaussian random variable in a separable Hilbert space  $\mathbb{H}$  with covariance operator  $\Sigma = \mathbb{E}(X \otimes X)$  and let  $X_1, \dots, X_n$  be a sample of  $n$  independent observations of  $X$ . The sample covariance operator is defined as  $\hat{\Sigma} := n^{-1} \sum_{j=1}^n X_j \otimes X_j$ . In the finite-dimensional case, it is well known that the operator norm error  $\|\hat{\Sigma} - \Sigma\|$  could be controlled in terms of the dimension  $d = \dim(\mathbb{H})$  of the space  $\mathbb{H}$ . In particular (see, e.g., [Vershynin \[2012\]](#)), for all  $t \geq 1$  with probability at least  $1 - e^{-t}$

$$(4) \quad \|\hat{\Sigma} - \Sigma\| \lesssim \|\Sigma\| \left( \sqrt{\frac{d}{n}} \vee \frac{d}{n} \vee \sqrt{\frac{t}{n}} \vee \frac{t}{n} \right),$$

which implies  $\mathbb{E}\|\hat{\Sigma} - \Sigma\| \asymp \|\Sigma\| \left( \sqrt{\frac{d}{n}} \vee \frac{d}{n} \right)$ . These bounds are sharp if the covariance operator is *isotropic* ( $\Sigma = cI$  for a constant  $c > 0$ ), or, more generally, it is of *isotropic type*, meaning that  $c_1 I \leq \Sigma \leq c_2 I$  for some constants  $0 < c_1 \leq c_2 < \infty$  (it is assumed that  $c, c_1, c_2$  are dimension free). The last condition holds, for instance, for well known *spiked covariance model* introduced by [Johnstone \[2001\]](#) (see also [Johnstone and Lu \[2009\]](#) and [Paul \[2007\]](#)). If the space  $\mathbb{H}$  is infinite-dimensional (or it is finite-dimensional, but the covariance operator  $\Sigma$  is not of isotropic type), bound (4) is no longer sharp and other complexity parameters become relevant in covariance estimation problem. In particular, [Vershynin \[2012\]](#) suggested to use in such cases so called *effective rank*  $\mathbf{r}(\Sigma) := \frac{\text{tr}(\Sigma)}{\|\Sigma\|}$  instead of the dimension. Clearly,  $\mathbf{r}(\Sigma) \leq \text{rank}(\Sigma) \leq \dim(\mathbb{H})$ . The next result was proved by [Koltchinskii and Lounici \[2017a\]](#) and it shows that  $\mathbf{r}(\Sigma)$  is a natural complexity parameter in covariance estimation (at least, in the Gaussian case).

**Theorem 1.** *The following expectation bound holds:*

$$(5) \quad \mathbb{E}\|\hat{\Sigma} - \Sigma\| \asymp \|\Sigma\| \left( \sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee \frac{\mathbf{r}(\Sigma)}{n} \right).$$

Moreover, for all  $t \geq 1$ , the following concentration inequality holds with probability at least  $1 - e^{-t}$  :

$$(6) \quad \left| \|\hat{\Sigma} - \Sigma\| - \mathbb{E}\|\hat{\Sigma} - \Sigma\| \right| \lesssim \|\Sigma\| \left( \left( \sqrt{\frac{\mathbf{r}(\Sigma)}{n}} \vee 1 \right) \sqrt{\frac{t}{n}} \vee \frac{t}{n} \right).$$

Note that bounds (5) and (6) were proved in [Koltchinskii and Lounici \[2017a\]](#) in a more general setting of estimation of covariance operator of a Gaussian random variable in a separable Banach space with effective rank defined as  $\mathbf{r}(\Sigma) := \frac{(\mathbb{E}\|X\|)^2}{\|\Sigma\|}$ . These bounds show that the “relative operator norm error”  $\frac{\|\hat{\Sigma} - \Sigma\|}{\|\Sigma\|}$  is controlled by the ratio  $\frac{\mathbf{r}(\Sigma)}{n}$  and that condition  $\mathbf{r}(\Sigma) = o(n)$  is necessary and sufficient for the operator norm consistency of the sample covariance. In view of these results, it became natural to study concentration and normal approximation properties of various statistics represented by functionals of sample covariance in a dimension free framework in which the effective rank  $\mathbf{r}(\Sigma)$  is allowed to be large (although satisfying the condition  $\mathbf{r}(\Sigma) = o(n)$ , which ensures that  $\hat{\Sigma}$  is a small perturbation of  $\Sigma$ ). This was done in [Koltchinskii and Lounici \[2016\]](#) in the case of bilinear forms of spectral projection operators of  $\hat{\Sigma}$  (empirical spectral projections) and in [Koltchinskii and Lounici \[2017c,b\]](#) in the case of their squared Hilbert–Schmidt error. It turned out that naive plug-in estimators (such as bilinear forms of empirical spectral projections) are not  $\sqrt{n}$ -consistent (unless  $\mathbf{r}(\Sigma) = o(n)$ ) due to their substantial bias and bias reduction becomes crucial for asymptotically efficient estimation. We briefly discuss below the approach to this problem developed by [Koltchinskii and Lounici \[2016\]](#), [Koltchinskii, Löffler, and Nickl \[2017\]](#).

Let  $\sigma(\Sigma)$  be the spectrum of  $\Sigma$  and let  $\lambda(\Sigma) = \sup(\sigma(\Sigma)) = \|\Sigma\|$  be its largest eigenvalue. Let  $g(\Sigma) := \text{dist}(\lambda(\Sigma); \sigma(\Sigma) \setminus \{\lambda(\Sigma)\})$  be the gap between  $\lambda(\Sigma)$  and the rest of the spectrum. Suppose  $\lambda(\Sigma)$  has multiplicity 1 and let  $P(\Sigma) = \theta(\Sigma) \otimes \theta(\Sigma)$  be the corresponding one-dimensional spectral projection. Here  $\theta(\Sigma)$  is the unit eigenvector corresponding to  $\lambda(\Sigma)$  (defined up to its sign). Given  $u \in \mathbb{H}$ , our goal is to estimate the linear functional  $\langle \theta(\Sigma), u \rangle$  based on i.i.d. observations  $X_1, \dots, X_n$  sampled from  $N(0; \Sigma)$  (note that the value of this functional is also defined only up to its sign, so, essentially, we can estimate only its absolute value). If  $\theta(\hat{\Sigma})$  denotes a unit eigenvector of sample covariance  $\hat{\Sigma}$  that corresponds to its top eigenvalue  $\lambda(\hat{\Sigma}) = \|\hat{\Sigma}\|$ , then  $\langle \theta(\hat{\Sigma}), u \rangle$  is the plug-in estimator of  $\langle \theta(\Sigma), u \rangle$ . Without loss of generality, we assume in what follows that  $\theta(\hat{\Sigma})$  and  $\theta(\Sigma)$  are properly aligned in the sense that  $\langle \theta(\hat{\Sigma}), \theta(\Sigma) \rangle \geq 0$  (which allows us indeed to view  $\langle \theta(\hat{\Sigma}), u \rangle$  as an estimator of  $\langle \theta(\Sigma), u \rangle$ ). It was shown in [Koltchinskii and Lounici \[2016\]](#), that the quantity

$$b(\Sigma) = b_n(\Sigma) := \mathbb{E}_{\Sigma} \langle \theta(\hat{\Sigma}), \theta(\Sigma) \rangle^2 - 1 \in [-1, 0]$$

characterizes the size of the bias of estimator  $\langle \theta(\hat{\Sigma}), u \rangle$ . In particular, the results of [Koltchinskii and Lounici \[2016\]](#) and [Koltchinskii, Löffler, and Nickl \[2017\]](#) imply that  $\langle \theta(\hat{\Sigma}), u \rangle$  “concentrates” around the value  $\sqrt{1 + b(\Sigma)} \langle \theta(\Sigma), u \rangle$  rather than around the value of the functional  $\langle \theta(\Sigma), u \rangle$  itself. To state this result more precisely, consider the spectral representation  $\Sigma = \sum_{\lambda \in \sigma(\Sigma)} \lambda P_{\lambda}$  with eigenvalues  $\lambda$  and corresponding orthogonal spectral

projections  $P_\lambda$ . Define

$$C(\Sigma) := \sum_{\lambda \neq \lambda(\Sigma)} (\lambda(\Sigma) - \lambda)^{-1} P_\lambda \text{ and } \sigma^2(\Sigma; u) := \lambda(\Sigma) \langle \Sigma C(\Sigma) u, C(\Sigma) u \rangle.$$

For  $u \in \mathbb{H}$ ,  $r > 1$ ,  $a > 1$  and  $\sigma_0 > 0$ , define the following class of covariance operators in  $\mathbb{H}$ :  $\mathcal{S}(r, a, \sigma_0, u) := \left\{ \Sigma : \mathbf{r}(\Sigma) \leq r, \frac{\|\Sigma\|}{g(\Sigma)} \leq a, \sigma^2(\Sigma; u) \geq \sigma_0^2 \right\}$ . Note that additional conditions on  $r, a, \sigma_0, u$  might be needed for the class  $\mathcal{S}(r, a, \sigma_0, u)$  to be nonempty.

**Theorem 2.** *Let  $u \in \mathbb{H}$ ,  $a > 1$  and  $\sigma_0 > 0$ . Suppose that  $r_n > 1$  and  $r_n = o(n)$  as  $n \rightarrow \infty$ . Then*

$$\sup_{\Sigma \in \mathcal{S}(r_n, a, \sigma_0, u)} \sup_{x \in \mathbb{R}} \left| \mathbb{P}_\Sigma \left\{ \frac{\sqrt{n}(\langle \theta(\hat{\Sigma}), u \rangle - \sqrt{1 + b(\Sigma)} \langle \theta(\Sigma), u \rangle)}{\sigma(\Sigma; u)} \leq x \right\} - \mathbb{P}\{Z \leq x\} \right| \rightarrow 0$$

and, for all  $\ell \in \mathcal{L}$ ,

$$\sup_{\Sigma \in \mathcal{S}(r_n, a, \sigma_0, u)} \left| \mathbb{E}_\Sigma \ell \left( \frac{\sqrt{n}(\langle \theta(\hat{\Sigma}), u \rangle - \sqrt{1 + b(\Sigma)} \langle \theta(\Sigma), u \rangle)}{\sigma(\Sigma; u)} \right) - \mathbb{E} \ell(Z) \right| \rightarrow 0 \text{ as } n \rightarrow \infty.$$

It was also proved in [Koltchinskii, Löffler, and Nickl \[2017\]](#) that  $b(\Sigma) \asymp \frac{\mathbf{r}(\Sigma)}{n}$ . This implies that the “bias”  $(\sqrt{1 + b(\Sigma)} - 1) \langle \theta(\Sigma), u \rangle$  of estimator  $\langle \theta(\hat{\Sigma}), u \rangle$  is asymptotically negligible (of the order  $o(n^{-1/2})$ ) if  $\mathbf{r}(\Sigma) = o(\sqrt{n})$ , which yields the following result:

**Corollary 1.** *Let  $u \in \mathbb{H}$ ,  $a > 1$  and  $\sigma_0 > 0$ . Suppose that  $r_n > 1$  and  $r_n = o(\sqrt{n})$  as  $n \rightarrow \infty$ , and that  $\mathcal{S}(r, a', \sigma'_0, u) \neq \emptyset$  for some  $r > 1, a' < a, \sigma'_0 > \sigma_0$ . Then  $\langle \theta(\hat{\Sigma}), u \rangle$  is an asymptotically efficient estimator of  $\langle \theta(\Sigma), u \rangle$  with respect to  $\mathcal{S}(r_n, a, \sigma_0, u)$  with convergence rate  $\sqrt{n}$  and variance  $\sigma^2(\Sigma; u)$ .*

On the other hand, it was shown in [Koltchinskii, Löffler, and Nickl \[ibid.\]](#) that, under the assumptions  $r_n = o(n)$  and  $\frac{r_n}{n^{1/2}} \rightarrow \infty$  as  $n \rightarrow \infty$ ,

$$\lim_{n \rightarrow \infty} \sup_{\Sigma \in \mathcal{S}(r_n, a, \sigma_0, u)} \mathbb{P}_\Sigma \left\{ |\langle \theta(\hat{\Sigma}), u \rangle - \langle \theta(\Sigma), u \rangle| \geq c \|u\| \frac{r_n}{n} \right\} = 1$$

for some constant  $c = c(a; \sigma_0) > 0$ , implying that  $\langle \theta(\hat{\Sigma}), u \rangle$  is not even  $\sqrt{n}$ -consistent estimator of  $\langle \theta(\Sigma), u \rangle$  when effective rank is larger than  $\sqrt{n}$ . Clearly, slower convergence rate is due to a large bias of the estimator  $\langle \theta(\hat{\Sigma}), u \rangle$  when the complexity of the problem becomes large (the effective rank exceeds  $\sqrt{n}$ ), and bias reduction is crucial to construct a  $\sqrt{n}$ -consistent estimator in this case. In [Koltchinskii and Lounici \[2016\]](#), a method of bias reduction based on estimation of bias parameter  $b(\Sigma)$  was developed. For simplicity,



assume that the sample size is even  $n = 2n'$  and split the sample into two equal parts, each of size  $n'$ . Let  $\hat{\Sigma}^{(1)}, \hat{\Sigma}^{(2)}$  be the sample covariances based on these two subsamples and let  $\theta(\hat{\Sigma}^{(1)}), \theta(\hat{\Sigma}^{(2)})$  be their top principal components. Since for all  $u \in \mathbb{H}$  and for  $i = 1, 2$ ,  $\langle \theta(\hat{\Sigma}^{(i)}), u \rangle$  “concentrates” around  $\sqrt{1 + b_{n'}(\Sigma)} \langle \theta(\Sigma), u \rangle$  and  $\theta(\hat{\Sigma}^{(i)}), i = 1, 2$  are independent, it is not hard to check that  $\langle \theta(\hat{\Sigma}^{(1)}), \theta(\hat{\Sigma}^{(2)}) \rangle$  “concentrates” around  $1 + b_{n'}(\Sigma)$ . Thus,  $\hat{b} := \langle \theta(\hat{\Sigma}^{(1)}), \theta(\hat{\Sigma}^{(2)}) \rangle - 1$  could be used as an estimator of  $b_{n'}(\Sigma)$ . It was proved in [Koltchinskii and Lounici \[2016\]](#) that  $\hat{b} - b_{n'}(\Sigma) = o_{\mathbb{P}}(n^{-1/2})$  provided that  $\mathbf{r}(\Sigma) = o(n)$ , and this led to a bias corrected estimator  $(1 + \hat{b})^{-1/2} \langle \theta(\hat{\Sigma}^{(1)}), u \rangle$  of linear functional  $\langle \theta(\Sigma), u \rangle$ , which was proved to be asymptotically normal with convergence rate  $\sqrt{n}$ . This approach was further developed in [Koltchinskii, Löffler, and Nickl \[2017\]](#), where a more subtle version of sample split yielded an asymptotically efficient estimator of the functional  $\langle \theta(\Sigma), u \rangle$ . Let  $m = m_n = o(n)$  as  $n \rightarrow \infty$ ,  $m < n/3$ . Split the sample  $X_1, \dots, X_n$  into three disjoint subsamples, one of size  $n' = n'_n := n - 2m > n/3$  and two others of size  $m$ . Let  $\hat{\Sigma}^{(1)}, \hat{\Sigma}^{(2)}, \hat{\Sigma}^{(3)}$  be the sample covariances based on these three subsamples and let  $\theta(\hat{\Sigma}^{(j)}), j = 1, 2, 3$  be the corresponding top principal components. Denote

$$\hat{d} := \frac{|\langle \theta(\hat{\Sigma}^{(1)}), \theta(\hat{\Sigma}^{(2)}) \rangle|}{|\langle \theta(\hat{\Sigma}^{(2)}), \theta(\hat{\Sigma}^{(3)}) \rangle|^{1/2}} \text{ and } \hat{\theta} := \frac{\theta(\hat{\Sigma}^{(1)})}{\hat{d} \vee (1/2)}.$$

**Theorem 3.** *Let  $u \in \mathbb{H}$ ,  $a > 1$  and  $\sigma_0 > 0$ . Suppose that  $r_n > 1$  and  $r_n = o(n)$  as  $n \rightarrow \infty$ . Suppose also that  $\mathcal{S}(r, a', \sigma'_0, u) \neq \emptyset$  for some  $r > 1, a' < a, \sigma'_0 > \sigma_0$ . Take  $m = m_n$  such that  $m_n = o(n)$  and  $nr_n = o(m_n^2)$  as  $n \rightarrow \infty$ . Then  $\langle \hat{\theta}, u \rangle$  is an asymptotically efficient estimator of  $\langle \theta(\Sigma), u \rangle$  with respect to  $\mathcal{S}(r_n, a, \sigma_0, u)$  with convergence rate  $\sqrt{n}$  and variance  $\sigma^2(\Sigma; u)$ .*

The approach to bias reduction and efficient estimation described above is based on rather special structural properties of the bias of empirical spectral projections. In the following sections, we discuss a much more general approach applicable to broader classes of problems.

### 3 Normal approximation bounds for plug-in estimators of smooth functionals

Let  $f : \mathbb{R} \mapsto \mathbb{R}$  be a smooth function and let  $B$  be a nuclear operator in  $\mathbb{H}$ . The goal is to estimate functionals of the form  $\langle f(\Sigma), B \rangle, \Sigma \in \mathcal{C}_+(\mathbb{H})$  based on i.i.d. observations  $X_1, \dots, X_n$  sampled from the Gaussian distribution with mean zero and covariance operator  $\Sigma$ . We will first consider a simple plug-in estimator  $\langle f(\hat{\Sigma}), B \rangle$ . To study its properties, we will rely on several results on Fréchet differentiability of operator functions  $\mathcal{B}_{sa}(\mathbb{H}) \ni A \mapsto f(A) \in \mathcal{B}_{sa}(\mathbb{H})$  with respect to the operator norm as well as on bounds

on the remainders of their Taylor expansions. These results could be found in operator theory literature (see, in particular, [Aleksandrov and Peller \[2016\]](#)).

We will need the definition of Besov spaces (see [Triebel \[1983\]](#) for more details). Consider a  $C^\infty$  function  $w \geq 0$  in  $\mathbb{R}$  with  $\text{supp}(w) \subset [-2, 2]$ , satisfying the assumptions  $w(t) = 1, |t| \leq 1$  and  $w(-t) = w(t), t \in \mathbb{R}$ . Let  $w_0(t) := w(t/2) - w(t), t \in \mathbb{R}$  (implying that  $\text{supp}(w_0) \subset \{t : 1 \leq |t| \leq 4\}$ ). For  $w_j(t) := w_0(2^{-j}t), t \in \mathbb{R}$ , we have  $\text{supp}(w_j) \subset \{t : 2^j \leq |t| \leq 2^{j+2}\}, j = 0, 1, \dots$  and also  $w(t) + \sum_{j \geq 0} w_j(t) = 1, t \in \mathbb{R}$ . Let  $W, W_j, j \geq 1$  be functions in Schwartz space  $\mathcal{S}(\mathbb{R})$  defined by their Fourier transforms:  $w(t) = (\mathcal{F}W)(t), w_j(t) = (\mathcal{F}W_j)(t), t \in \mathbb{R}, j \geq 0$ . For a tempered distribution  $f \in \mathcal{S}'(\mathbb{R})$ , define its Littlewood-Paley decomposition as the set of functions  $f_0 := f * W, f_n := f * W_{n-1}, n \geq 1$  with compactly supported Fourier transforms. By Paley-Wiener Theorem,  $f_n$  can be extended to an entire function of exponential type  $2^{n+1}$  (for all  $n \geq 0$ ). It is also well known that  $\sum_{n \geq 0} f_n = f$  with convergence of the series in the space  $\mathcal{S}'(\mathbb{R})$ . Define  $B_{\infty,1}^s$ -Besov norm as

$$\|f\|_{B_{\infty,1}^s} := \sum_{n \geq 0} 2^{ns} \|f_n\|_{L_\infty(\mathbb{R})}, s \in \mathbb{R}$$

and let  $B_{\infty,1}^s(\mathbb{R}) := \{f \in \mathcal{S}'(\mathbb{R}) : \|f\|_{B_{\infty,1}^s} < +\infty\}$  be the corresponding (inhomogeneous) Besov space. It is easy to check that, for  $s \geq 0$ , the series  $\sum_{n \geq 0} f_n$  converges uniformly in  $\mathbb{R}$  and the space  $B_{\infty,1}^s(\mathbb{R})$  is continuously embedded in the space  $C_u(\mathbb{R})$  of all bounded uniformly continuous functions equipped with the uniform norm  $\|\cdot\|_{L_\infty(\mathbb{R})}$ .

It was proved by [Peller \[1985\]](#) that, for all  $f \in B_{\infty,1}^1(\mathbb{R})$ , the mapping  $\mathfrak{B}_{sa}(\mathbb{H}) \ni A \mapsto f(A) \in \mathfrak{B}_{sa}(\mathbb{H})$  is Fréchet differentiable with respect to the operator norm (in fact, Peller used homogeneous Besov spaces). Let  $Df(A; H) = Df(A)(H)$  denote its derivative at  $A$  in direction  $H$ . If  $A \in \mathfrak{B}_{sa}(\mathbb{H})$  is a compact operator with spectral representation  $A = \sum_{\lambda \in \sigma(A)} \lambda P_\lambda$  with eigenvalues  $\lambda$  and spectral projections  $P_\lambda$ , then  $Df(A; H) = \sum_{\lambda, \mu \in \sigma(A)} f^{[1]}(\lambda, \mu) P_\lambda H P_\mu$ , where  $f^{[1]}(\lambda, \mu) := \frac{f(\lambda) - f(\mu)}{\lambda - \mu}, \lambda \neq \mu, f^{[1]}(\lambda, \mu) := f'(\lambda), \lambda = \mu$  is Loewner kernel (there are also extensions of this formula for more general operators with continuous spectrum with double operator integrals instead of the sums [Peller \[2006\]](#) and [Aleksandrov and Peller \[2016\]](#)). If  $f \in B_{\infty,1}^s(\mathbb{R})$  for some  $s \in (1, 2]$ , then the first order Taylor expansion  $f(A + H) = f(A) + Df(A; H) + S_f(A; H)$  holds with the following bound on the remainder:  $\|S_f(A; H)\| \lesssim_s \|f\|_{B_{\infty,1}^s} \|H\|^s, H \in \mathfrak{B}_{sa}(\mathbb{H})$  (see [Koltchinskii \[2017\]](#) for the proof fully based on methods of [Aleksandrov and Peller \[2016\]](#)). Applying the Taylor expansion to  $f(\hat{\Sigma})$  and using the bound on the remainder along with [Theorem 1](#), we get that

$$f(\hat{\Sigma}) - f(\Sigma) = Df(\Sigma; \hat{\Sigma} - \Sigma) + S_f(\Sigma; \hat{\Sigma} - \Sigma)$$

with  $\|S_f(\Sigma; \hat{\Sigma} - \Sigma)\| = o_{\mathbb{P}}(n^{-1/2})$  provided that  $\mathbf{r}(\Sigma) = o(n^{1-1/s})$ . It is also easy to check that

$$\sqrt{n}\langle Df(\Sigma; \hat{\Sigma} - \Sigma), B \rangle = n^{-1/2} \sum_{j=1}^n \langle Df(\Sigma; X_j \otimes X_j - \Sigma), B \rangle$$

is asymptotically normal  $N(0; \sigma_f^2(\Sigma, B))$ ,  $\sigma_f^2(\Sigma; B) := 2\|\Sigma^{1/2} Df(\Sigma; B) \Sigma^{1/2}\|_2^2$ , which, along with asymptotic negligibility of the remainder, implies the asymptotic normality of  $\sqrt{n}(\langle f(\hat{\Sigma}), B \rangle - \langle f(\Sigma), B \rangle)$  with the same limit mean and variance. Similar rather standard perturbation analysis (most often, based on holomorphic functional calculus rather than on more sophisticated tools of [Aleksandrov and Peller \[2016\]](#)) has been commonly used, especially, in applications to PCA, in the case of finite-dimensional problems of bounded dimension, see, e.g., [Anderson \[2003\]](#). It, however, fails as soon as the effective rank is sufficiently large (above  $n^{1-1/s}$  for functions of smoothness  $s \in (1, 2]$ ) since the remainder  $S_f(\Sigma; \hat{\Sigma} - \Sigma)$  of Taylor expansion is not asymptotically negligible. It turns out, that in this case  $\langle f(\hat{\Sigma}), B \rangle$  is still a  $\sqrt{n}$ -consistent and asymptotically normal estimator of its own expectation  $\langle \mathbb{E}_{\Sigma} f(\hat{\Sigma}), B \rangle$ , but the bias  $\langle \mathbb{E}_{\Sigma} f(\hat{\Sigma}) - f(\Sigma), B \rangle$  is no longer asymptotically negligible. In fact, the bias is equal to  $\langle \mathbb{E}_{\Sigma} S_f(\Sigma; \hat{\Sigma} - \Sigma), B \rangle$ , which is upper bounded by  $\lesssim \|f\|_{B_{\infty,1}^s} \|B\|_1 (\frac{\mathbf{r}(\Sigma)}{n})^{s/2}$ . This bound is sharp for typical smooth functions. For instance, if  $f(x) = x^2$  and  $B = u \otimes u$ , it is easy to check that

$$\sup_{\|u\| \leq 1} |\langle \mathbb{E}_{\Sigma} f(\hat{\Sigma}) - f(\Sigma), u \otimes u \rangle| \asymp \|\Sigma\|^2 \frac{\mathbf{r}(\Sigma)}{n},$$

and the bias is not asymptotically negligible if  $\mathbf{r}(\Sigma) \geq n^{1/2}$ . Moreover, if  $\frac{\mathbf{r}(\Sigma)}{\sqrt{n}} \rightarrow \infty$ , the plug-in estimator  $\langle f(\hat{\Sigma}), u \otimes u \rangle$  of  $\langle f(\Sigma), u \otimes u \rangle$  is not  $\sqrt{n}$ -consistent (for some  $u$  with  $\|u\| \leq 1$ ).

The next result (see also [Koltchinskii \[2017\]](#)) shows asymptotic normality (with  $\sqrt{n}$ -rate) of  $\langle f(\hat{\Sigma}), B \rangle$  as an estimator of its own expectation. Define

$$\mathfrak{G}_{f,B}(r; a; \sigma_0) := \left\{ \Sigma : \mathbf{r}(\Sigma) \leq r, \|\Sigma\| \leq a, \sigma_f^2(\Sigma; B) \geq \sigma_0^2 \right\}, \quad r > 1, a > 0, \sigma_0^2 > 0.$$

**Theorem 4.** *Let  $f \in B_{\infty,1}^s(\mathbb{R})$  for some  $s \in (1, 2]$  and let  $B$  be a nuclear operator. For any  $a > 0$ ,  $\sigma_0^2 > 0$  and  $r_n > 1$  such that  $r_n = o(n)$  as  $n \rightarrow \infty$ ,*

(7)

$$\sup_{\Sigma \in \mathfrak{G}_{f,B}(r_n; a; \sigma_0)} \sup_{x \in \mathbb{R}} \left| \mathbb{P}_{\Sigma} \left\{ \frac{n^{1/2} \langle f(\hat{\Sigma}) - \mathbb{E}_{\Sigma} f(\hat{\Sigma}), B \rangle}{\sigma_f(\Sigma; B)} \leq x \right\} - \mathbb{P}\{Z \leq x\} \right| \rightarrow 0 \text{ as } n \rightarrow \infty.$$

The proof of this result is based on the following simple representation

$$\langle f(\hat{\Sigma}) - \mathbb{E}_{\Sigma} f(\hat{\Sigma}), B \rangle = \frac{1}{n} \sum_{j=1}^n \langle Df(\Sigma; X_j \otimes X_j - \Sigma), B \rangle + \langle S_f(\Sigma; \hat{\Sigma} - \Sigma) - \mathbb{E}_{\Sigma} S_f(\Sigma; \hat{\Sigma} - \Sigma), B \rangle.$$

For the first term in the right hand side, it is easy to prove a normal approximation bound based on Berry-Esseen inequality. The main part of the proof deals with the second term, the centered remainder of Taylor expansion  $\langle S_f(\Sigma; \hat{\Sigma} - \Sigma), B \rangle - \mathbb{E} \langle S_f(\Sigma; \hat{\Sigma} - \Sigma), B \rangle$ . For this term, the following bound was proved using Gaussian concentration inequality: for all  $t \geq 1$  with probability at least  $1 - e^{-t}$ ,

$$(8) \quad |\langle S_f(\Sigma; \hat{\Sigma} - \Sigma) - \mathbb{E} S_f(\Sigma; \hat{\Sigma} - \Sigma), B \rangle| \\ \lesssim_s \|f\|_{B_{\infty,1}^s} \|B\|_1 \|\Sigma\|^s \left( \left( \frac{\mathbf{r}(\Sigma)}{n} \right)^{(s-1)/2} \bigvee \left( \frac{\mathbf{r}(\Sigma)}{n} \right)^{s-1/2} \bigvee \left( \frac{t}{n} \right)^{(s-1)/2} \bigvee \left( \frac{t}{n} \right)^{s-1/2} \right) \sqrt{\frac{t}{n}}.$$

It implies that the centered remainder is of the order  $\left( \frac{\mathbf{r}(\Sigma)}{n} \right)^{(s-1)/2} \sqrt{\frac{1}{n}}$ , which is  $o(n^{-1/2})$  as soon as  $\mathbf{r}(\Sigma) = o(n)$ .

If  $\Sigma \in \mathfrak{g}_{f,B}(r_n; a; \sigma_0)$  with  $r_n = o(n^{1-1/s})$ , the bias  $\langle \mathbb{E}_\Sigma f(\hat{\Sigma}) - f(\Sigma), B \rangle$  is of the order  $o(n^{-1/2})$  and plug-in estimator  $\langle f(\hat{\Sigma}), B \rangle$  of  $\langle f(\Sigma), B \rangle$  is asymptotically normal with  $\sqrt{n}$ -rate. The following corollary of Theorem 4 holds.

**Corollary 2.** *Let  $f \in B_{\infty,1}^s(\mathbb{R})$  for some  $s \in (1, 2]$ , and let  $B$  be a nuclear operator. Let  $a > 0, \sigma_0^2 > 0$  and let  $r_n > 1$  be such that  $r_n \rightarrow \infty$  and  $r_n = o(n^{1-\frac{1}{s}})$  as  $n \rightarrow \infty$ . Suppose that  $\mathfrak{g}_{f,B}(r; a'; \sigma'_0) \neq \emptyset$  for some  $r > 1, a' < a, \sigma'_0 > \sigma_0$ . Then  $\langle f(\hat{\Sigma}), B \rangle$  is an asymptotically efficient estimator of  $\langle f(\Sigma), B \rangle$  with respect to  $\mathfrak{g}_{f,B}(r_n; a; \sigma_0)$  with convergence rate  $\sqrt{n}$  and variance  $\sigma_f^2(\Sigma; B)$ .*

Thus, as soon as  $r_n = o(n^{1/2})$  and  $f$  is sufficiently smooth, the plug-in estimator is asymptotically efficient. However, as we have already pointed out above, this conclusion is not true if  $r_n \geq n^{1/2}$  regardless of the degree of smoothness of  $f$  (even in the case of function  $f(x) = x^2$ ). Moreover, not only asymptotic efficiency, but even  $\sqrt{n}$ -consistency of the plug-in estimator does not hold in this case, and the problem of asymptotically efficient estimation of functionals  $\langle f(\Sigma), B \rangle$  becomes much more complicated. In the following sections, we outline a solution of this problem with the dimension of the space rather than the effective rank playing the role of complexity parameter. The idea of our approach is to try to find a function  $g$  on the space  $\mathfrak{B}_{sa}(\mathbb{H})$  of self-adjoint operators that solves approximately the equation  $\mathbb{E}_\Sigma g(\hat{\Sigma}) = f(\Sigma)$  with an error of the order  $o(n^{-1/2})$ . If such solution  $g$  is sufficiently smooth, it could be possible to prove an analog of normal approximation of Theorem 4 for estimator  $\langle g(\hat{\Sigma}), B \rangle$ . Since the bias of this estimator is asymptotically negligible, it would be possible to show asymptotic normality of  $\langle g(\hat{\Sigma}), B \rangle$  as an estimator of  $\langle f(\Sigma), B \rangle$ .

## 4 Bootstrap Chain bias reduction and asymptotically efficient estimation

Assume that  $d := \dim(\mathbb{H})$  is finite (in what follows,  $d = d_n$  could grow with  $n$ ). It also will be assumed that the covariance operator  $\Sigma$  is of isotropic type. The following integral operator on the cone  $\mathcal{C}_+(\mathbb{H})$  will be crucial in our approach:

$$\mathcal{T}g(\Sigma) := \mathbb{E}_{\Sigma} g(\hat{\Sigma}) = \int_{\mathcal{C}_+(\mathbb{H})} g(S) P(\Sigma; dS), \Sigma \in \mathcal{C}_+(\mathbb{H}).$$

Here  $P(\Sigma; \cdot)$  is the distribution of the sample covariance  $\hat{\Sigma}$  based on  $n$  i.i.d. observations sampled from  $N(0; \Sigma)$ . Clearly,  $P(\Sigma; \cdot)$  is a rescaled Wishart distribution and  $P(\cdot; \cdot)$  is a Markov kernel on the cone  $\mathcal{C}_+(\mathbb{H})$ . We will call  $\mathcal{T}$  *the Wishart operator* and view it as an operator acting on bounded measurable functions on the cone  $\mathcal{C}_+(\mathbb{H})$  with values either in  $\mathbb{R}$  or in  $\mathcal{B}_{sa}(\mathbb{H})$ . Such operators are well known in the theory of Wishart matrices (see, e.g., [James \[1961\]](#), [Letac and Massam \[2004\]](#)). To obtain an unbiased estimator  $g(\hat{\Sigma})$  of  $f(\Sigma)$ , one needs to solve the integral equation  $\mathcal{T}g(\Sigma) = f(\Sigma)$ ,  $\Sigma \in \mathcal{C}_+(\mathbb{H})$  (*the Wishart equation*). Denoting  $\mathcal{B} := \mathcal{T} - \mathbb{I}$ ,  $\mathbb{I}$  being the identity operator, one can write the solution of the Wishart equation as a formal Neumann series  $g(\Sigma) = (\mathbb{I} + \mathcal{B})^{-1} f(\Sigma) = \sum_{j=0}^{\infty} (-1)^j \mathcal{B}^j f(\Sigma)$ . We will use its partial sums to define approximate solutions of the Wishart equation:

$$f_k(\Sigma) := \sum_{j=0}^{\infty} (-1)^j \mathcal{B}^j f(\Sigma), \Sigma \in \mathcal{C}_+(\mathbb{H}), k \geq 0,$$

with  $f_k(\hat{\Sigma})$  for a properly chosen  $k$  being an estimator of  $f(\Sigma)$ . Note that its bias is

$$\mathbb{E}_{\Sigma} f_k(\hat{\Sigma}) - f(\Sigma) = (-1)^k \mathcal{B}^{k+1} f(\Sigma), \Sigma \in \mathcal{C}_+(\mathbb{H})$$

and, to justify this approach to bias reduction, one has to show that, for smooth enough functions  $f$  and large enough value of  $k$ ,  $\langle \mathcal{B}^{k+1} f(\Sigma), B \rangle$  is of the order  $o(n^{-1/2})$ . Note that a similar approach was recently discussed by [Jiao, Han, and Weissman \[2017\]](#) in the case of a problem of estimation of smooth function of parameter of binomial model  $B(n; \theta)$ ,  $\theta \in [0, 1]$ . If  $\hat{\theta}$  denotes the frequency, then  $Tg(\theta) = \mathbb{E}_{\theta} g(\hat{\theta})$  is a Bernstein polynomial approximation of function  $g$  and bounds on  $\mathcal{B}^{k+1} f(\theta)$  were deduced in [Jiao, Han, and Weissman \[ibid.\]](#) from some results of classical approximation theory (see, e.g., [Totik \[1994\]](#)).

We describe below our approach in [Koltchinskii \[2017\]](#) based on a Markov chain interpretation of the problem. To this end, consider a Markov chain  $\hat{\Sigma}^{(0)} = \Sigma \rightarrow \hat{\Sigma}^{(1)} = \hat{\Sigma} \rightarrow \hat{\Sigma}^{(2)} \rightarrow \dots$  in the cone  $\mathcal{C}_+(\mathbb{H})$  with transition probability kernel  $P(\cdot; \cdot)$ . Note that

for any  $t \geq 1$ ,  $\hat{\Sigma}^{(t)}$  can be viewed as the sample covariance based on  $n$  i.i.d. observations sampled from normal distribution  $N(0; \hat{\Sigma}^{(t-1)})$ , conditionally on  $\hat{\Sigma}^{(t-1)}$ . In other words, the Markov chain  $\hat{\Sigma}^{(t)}, t = 0, 1, 2, \dots$  is an outcome of iterative parametric bootstrap procedure and it will be called in what follows *the Bootstrap Chain*. By bound (4), conditionally on  $\hat{\Sigma}^{(t-1)}$ , with a high probability  $\|\hat{\Sigma}^{(t)} - \hat{\Sigma}^{(t-1)}\| \lesssim \|\hat{\Sigma}^{(t-1)}\| \sqrt{\frac{d}{n}}$ , implying that the Bootstrap Chain moves with “small steps”, provided that  $d = o(n)$ . Now observe that  $\mathcal{T}^k f(\Sigma) = \mathbb{E}_{\Sigma} f(\hat{\Sigma}^{(k)})$  and, by Newton’s binomial formula,

$$\begin{aligned} (9) \quad \mathcal{B}^k f(\Sigma) &= (\mathcal{T} - \mathbb{I})^k f(\Sigma) = \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} \mathcal{T}^j f(\Sigma) = \\ &= \mathbb{E}_{\Sigma} \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(\hat{\Sigma}^{(j)}). \end{aligned}$$

Note that to compute functions  $\mathcal{B}^k f(\hat{\Sigma})$ ,  $k \geq 1$  (which is needed to compute the estimator  $f_k(\hat{\Sigma})$ ) one can use bootstrap:  $\mathcal{B}^k f(\hat{\Sigma}) = \mathbb{E}_{\hat{\Sigma}} \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(\hat{\Sigma}^{(j+1)})$  since the Bootstrap Chain now starts with  $\hat{\Sigma}^{(0)} = \hat{\Sigma}$ , and it can be approximated by the average of Monte Carlo simulations of  $\sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(\hat{\Sigma}^{(j+1)})$ .

Denote  $F(j) := f(\hat{\Sigma}^{(j)}), j \geq 0$  and  $\Delta F(j) := F(j+1) - F(j), j \geq 0$ . Then  $\sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(\hat{\Sigma}^{(j)}) = \Delta^k F(0)$  is the  $k$ -th order difference of sequence  $F(j), j \geq 0$  at  $j = 0$  (in other words, the  $k$ -th order difference of function  $f$  on the Markov chain  $\{\hat{\Sigma}^{(t)}\}$ ). It is well known that, for a  $k$  times continuously differentiable function  $f$  in  $\mathbb{R}$  the  $k$ -th order difference  $\Delta_h^k f(x)$ , where  $\Delta_h f(x) := f(x+h) - f(x)$ , is of the order  $O(h^k)$  as  $h \rightarrow 0$ . Since the chain  $\{\hat{\Sigma}^{(t)}\}$  moves with steps  $\asymp \sqrt{\frac{d}{n}}$ , it becomes plausible that, on average,  $\sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(\hat{\Sigma}^{(j)})$  would be of the order  $O((\frac{d}{n})^{k/2})$  for functions of smoothness  $k$ . The justification of this heuristic will be discussed in some detail in the next section and it is based on the development of certain integral representations of functions  $\mathcal{B}^k f(\Sigma), k \geq 1$  that rely on properties of orthogonally invariant functions on the cone  $\mathcal{C}_+(\mathbb{H})$ . These representations are then used to obtain bounds on operators  $\mathcal{B}^k f(\Sigma)$  and on the bias  $\mathbb{E}_{\Sigma} f_k(\hat{\Sigma}) - f(\Sigma)$  of estimator  $f(\Sigma)$ , to study smoothness properties of functions  $\mathcal{B}^k f(\Sigma)$  and  $f_k(\Sigma)$  that allow us to prove concentration bounds on the remainder  $\langle S_{f_k}(\Sigma; \hat{\Sigma} - \Sigma), B \rangle$  of the first order Taylor expansion of  $\langle f_k(\hat{\Sigma}), B \rangle$  and, finally, to establish normal approximation bounds for  $\langle f_k(\hat{\Sigma}) - f(\Sigma), B \rangle$ . This leads to the following result.

For given  $d > 1, a > 0$  and  $\sigma_0^2 > 0$ , let  $\mathcal{S}_{f,B}(d; a; \sigma_0)$  be the set of all covariance operators in  $d$ -dimensional space  $\mathbb{H}$  such that  $\|\Sigma\| \leq a, \|\Sigma^{-1}\| \leq a$  and  $\sigma_f^2(\Sigma; B) \geq \sigma_0^2$ .

**Theorem 5.** Suppose that, for some  $\alpha \in (0, 1)$ ,  $1 \leq d_n \leq n^\alpha$ ,  $n \geq 1$ . Let  $B = B_n$  be a self-adjoint operator with  $\|B\|_1 \leq 1$ . Let  $f \in B_{\infty,1}^s(\mathbb{R})$  for some  $s > \frac{1}{1-\alpha}$ , and let  $k$  be an integer number such that, for some  $\beta \in (0, 1]$ ,  $\frac{1}{1-\alpha} < k + 1 + \beta \leq s$ . Finally, suppose that  $\mathfrak{S}_{f,B}(d_n; a'; \sigma'_0) \neq \emptyset$  for some  $a' < a$ ,  $\sigma'_0 > \sigma_0$  and for all large enough  $n$ . Then  $\langle f_k(\hat{\Sigma}), B \rangle$  is an asymptotically efficient estimator of  $\langle f(\Sigma), B \rangle$  with respect to  $\mathfrak{S}_{f,B}(d_n; a; \sigma_0)$  with convergence rate  $\sqrt{n}$  and variance  $\sigma_f^2(\Sigma; B)$ .

Note that for  $\alpha \in (0, 1/2)$  and  $s > \frac{1}{1-\alpha}$ , we can set  $k = 0$ . In this case,  $f_k(\hat{\Sigma}) = f(\hat{\Sigma})$  is a standard plug-in estimator (see also [Corollary 2](#)). For  $\alpha = \frac{1}{2}$ , the assumption  $s > 2$  is needed and, to satisfy the condition  $k + 1 + \beta > \frac{1}{1-\alpha} = 2$ , we should choose  $k = 1$ . The bias correction becomes nontrivial in this case. For larger values of  $\alpha$ , more smoothness of  $f$  and more iterations  $k$  in the bias reduction method are needed.

## 5 Wishart operators and orthogonally invariant functions

In this section, we outline our approach to the proof of [Theorem 5](#) (see [Koltchinskii \[2017\]](#) for further details). The idea is to represent function  $f$  in the form  $f(x) = x\psi'(x)$ ,  $x \in \mathbb{R}$ , where  $\psi$  is a smooth function in the real line. Consider now the functional  $g(\Sigma) := \text{tr}(\psi(\Sigma))$ . Then,  $g$  is Fréchet differentiable with derivative  $Dg(\Sigma) = \psi'(\Sigma)$  and

$$(10) \quad f(\Sigma) = \Sigma^{1/2} Dg(\Sigma) \Sigma^{1/2} =: \mathfrak{D}g(\Sigma).$$

The functional  $g(\Sigma)$  is orthogonally invariant which allowed us to develop integral representations of functions  $\mathfrak{B}^k \mathfrak{D}g(\Sigma)$  and use them to study analytic properties of functions  $\mathfrak{B}^k f(\Sigma)$  and  $f_k(\Sigma)$ .

As in the previous section, we assume that  $\mathbb{H}$  is a finite-dimensional inner product space of dimension  $\dim(\mathbb{H}) = d$ . Recall that  $\mathcal{T}g(\Sigma) = \mathbb{E}_\Sigma g(\hat{\Sigma})$ ,  $\Sigma \in \mathcal{C}_+(\mathbb{H})$ . We will view  $\mathcal{T}$  as an operator from the space  $L_\infty(\mathcal{C}_+(\mathbb{H}))$  into itself,  $L_\infty(\mathcal{C}_+(\mathbb{H}))$  being the space of uniformly bounded Borel measurable real valued functions on the cone  $\mathcal{C}_+(\mathbb{H})$ . Alternatively,  $\mathcal{T}$  can be viewed as an operator from  $L_\infty(\mathcal{C}_+(\mathbb{H}); \mathfrak{B}_{sa}(\mathbb{H}))$  into  $L_\infty(\mathcal{C}_+(\mathbb{H}); \mathfrak{B}_{sa}(\mathbb{H}))$  (the space of uniformly bounded Borel measurable functions from  $\mathcal{C}_+(\mathbb{H})$  into  $\mathfrak{B}_{sa}(\mathbb{H})$ ).

A function  $g \in L_\infty(\mathcal{C}_+(\mathbb{H}))$  is called *orthogonally invariant* iff, for all orthogonal transformations  $U$  of  $\mathbb{H}$ ,  $g(U\Sigma U^{-1}) = g(\Sigma)$ ,  $\Sigma \in \mathcal{C}_+(\mathbb{H})$ . Any such function  $g$  could be represented as a symmetric function  $\varphi$  of eigenvalues  $\lambda_1(\Sigma), \dots, \lambda_d(\Sigma)$  of  $\Sigma$ :  $g(\Sigma) = \varphi(\lambda_1(\Sigma), \dots, \lambda_d(\Sigma))$ . A typical example is orthogonally invariant function  $g(\Sigma) = \text{tr}(\psi(\Sigma)) = \sum_{j=1}^d \psi(\lambda_j(\Sigma))$  for a function of real variable  $\psi$ . Denote by  $L_\infty^O(\mathcal{C}_+(\mathbb{H}))$  the subspace of all orthogonally invariant functions from  $L_\infty(\mathcal{C}_+(\mathbb{H}))$ . Clearly,  $L_\infty^O(\mathcal{C}_+(\mathbb{H}))$  is an algebra. It is easy to see that operators  $\mathcal{T}$  and  $\mathfrak{B} = \mathcal{T} - \mathfrak{I}$  map the space  $L_\infty^O(\mathcal{C}_+(\mathbb{H}))$  into itself.

A function  $g \in L_\infty(\mathbb{C}_+(\mathbb{H}); \mathfrak{B}_{sa}(\mathbb{H}))$  is called *orthogonally equivariant* iff, for all orthogonal transformations  $U$ ,  $g(U\Sigma U^{-1}) = Ug(\Sigma)U^{-1}$ ,  $\Sigma \in \mathbb{C}_+(\mathbb{H})$ . A function  $g : \mathbb{C}_+(\mathbb{H}) \mapsto \mathfrak{B}_{sa}(\mathbb{H})$  will be called differentiable (continuously differentiable) in  $\mathbb{C}_+(\mathbb{H})$  with respect to the operator norm iff there exists a uniformly bounded, Lipschitz and differentiable (continuously differentiable) extension of  $g$  to an open set  $G$ ,  $\mathbb{C}_+(\mathbb{H}) \subset G \subset \mathfrak{B}_{sa}(\mathbb{H})$ . If  $g : \mathbb{C}_+(\mathbb{H}) \mapsto \mathbb{R}$  is orthogonally invariant and continuously differentiable in  $\mathbb{C}_+(\mathbb{H})$  with derivative  $Dg$ , then it is easy to check that  $Dg$  is orthogonally equivariant.

We will use some simple properties of operators  $\mathcal{T}$  and  $\mathfrak{B} = \mathcal{T} - \mathfrak{L}$  acting in the space  $L_\infty^O(\mathbb{C}_+(\mathbb{H}))$  of uniformly bounded orthogonally invariant functions (and its subspaces). These properties are well known in the literature on Wishart distribution (at least, in the case of orthogonally invariant polynomials, see, e.g., [Letac and Massam \[2004\]](#)). Define the following differential operator  $\mathfrak{D}g(\Sigma) := \Sigma^{1/2}Dg(\Sigma)\Sigma^{1/2}$  acting on continuously differentiable functions in  $\mathbb{C}_+(\mathbb{H})$ . It turns out that operators  $\mathcal{T}$  and  $\mathfrak{D}$  commute (and, as a consequence,  $\mathfrak{B}$  and  $\mathfrak{D}$  also commute).

**Proposition 1.** *If  $g \in L_\infty^O(\mathbb{C}_+(\mathbb{H}))$  is continuously differentiable in  $\mathbb{C}_+(\mathbb{H})$  with a uniformly bounded derivative  $Dg$ , then, for all  $\Sigma \in \mathbb{C}_+(\mathbb{H})$ ,  $\mathfrak{D}\mathcal{T}g(\Sigma) = \mathcal{T}\mathfrak{D}g(\Sigma)$  and  $\mathfrak{D}\mathfrak{B}g(\Sigma) = \mathfrak{B}\mathfrak{D}g(\Sigma)$ .*

Let  $W$  be the sample covariance based on i.i.d. standard normal random variables  $Z_1, \dots, Z_n$  in  $\mathbb{H}$  (in other words,  $nW$  has standard Wishart distribution) and let  $W_1, \dots, W_k, \dots$  be i.i.d. copies of  $W$ . The next proposition provides representations of operators  $\mathcal{T}^k$  and  $\mathfrak{B}^k$  that will be used in what follows.

**Proposition 2.** *For all  $g \in L_\infty^O(\mathbb{C}_+(\mathbb{H}))$  and for all  $k \geq 1$ ,*

$$(11) \quad \mathcal{T}^k g(\Sigma) = \mathbb{E}g(W_k^{1/2} \dots W_1^{1/2} \Sigma W_1^{1/2} \dots W_k^{1/2})$$

and

$$(12) \quad \mathfrak{B}^k g(\Sigma) = \mathbb{E} \sum_{I \subset \{1, \dots, k\}} (-1)^{k-|I|} g(A_I^* \Sigma A_I),$$

where  $A_I := \prod_{i \in I} W_i^{1/2}$ . If, in addition,  $g$  is continuously differentiable in  $\mathbb{C}_+(\mathbb{H})$  with a uniformly bounded derivative  $Dg$ , then

$$(13) \quad D\mathfrak{B}^k g(\Sigma) = \mathbb{E} \sum_{I \subset \{1, \dots, k\}} (-1)^{k-|I|} A_I Dg(A_I^* \Sigma A_I) A_I^*,$$

and, for all  $\Sigma \in \mathbb{C}_+(\mathbb{H})$ ,

$$(14) \quad \mathfrak{D}\mathcal{T}^k g(\Sigma) = \mathcal{T}^k \mathfrak{D}g(\Sigma) \text{ and } \mathfrak{D}\mathfrak{B}^k g(\Sigma) = \mathfrak{B}^k \mathfrak{D}g(\Sigma).$$



Finally,

$$(15) \quad \mathfrak{B}^k \mathfrak{D}g(\Sigma) = \mathfrak{D}\mathfrak{B}^k g(\Sigma) = \mathbb{E} \left( \sum_{I \subset \{1, \dots, k\}} (-1)^{k-|I|} \Sigma^{1/2} A_I Dg(A_I^* \Sigma A_I) A_I^* \Sigma^{1/2} \right).$$

*Proof.* Note that  $\hat{\Sigma} \stackrel{d}{=} \Sigma^{1/2} W \Sigma^{1/2}$ . It is easy to check that

$$W^{1/2} \Sigma W^{1/2} = U^{-1} \Sigma^{1/2} W \Sigma^{1/2} U$$

where  $U$  is an orthogonal operator. Since  $g$  is orthogonally invariant, we have

$$(16) \quad \mathcal{T}g(\Sigma) = \mathbb{E} g(\hat{\Sigma}) = \mathbb{E} g(W^{1/2} \Sigma W^{1/2}).$$

Recall that orthogonal invariance of  $g$  implies orthogonal invariance of  $\mathcal{T}g$  and, by induction, of  $\mathcal{T}^k g$  for all  $k \geq 1$ . Then, also by induction, (16) implies that

$$\mathcal{T}^k g(\Sigma) = \mathbb{E} g(W_k^{1/2} \dots W_1^{1/2} \Sigma W_1^{1/2} \dots W_k^{1/2}).$$

For  $I \subset \{1, \dots, k\}$  with  $|I| = \text{card}(I) = j$  and  $A_I = \prod_{i \in I} W_i^{1/2}$ , it follows that  $\mathcal{T}^j g(\Sigma) = \mathbb{E} g(A_I^* \Sigma A_I)$ . In view of (9), we easily get (12). If  $g$  is continuously differentiable in  $\mathcal{C}_+(\mathbb{H})$  with a uniformly bounded derivative  $Dg$ , then (12) implies (13). Finally, it follows from (13) that the derivatives  $D\mathfrak{B}^k g$ ,  $k \geq 1$  are continuous and uniformly bounded in  $\mathcal{C}_+(\mathbb{H})$ . Similar property holds for the derivatives  $D\mathcal{T}^k g$ ,  $k \geq 1$  (as a consequence of (11) and the properties of  $g$ ). Therefore, (14) follows from Proposition 1 by induction. Formula (15) follows from (14) and (13).

□

□

The following functions provide the linear interpolation between the identity operator  $I$  and operators  $W_1^{1/2}, \dots, W_k^{1/2}$ :

$$V_j(t_j) := I + t_j(W_j^{1/2} - I), t_j \in [0, 1], 1 \leq j \leq k.$$

Note that for all  $j = 1, \dots, k$ ,  $t_j \in [0, 1]$ ,  $V_j(t_j) \in \mathcal{C}_+(\mathbb{H})$ . Let

$$R = R(t_1, \dots, t_k) = V_1(t_1) \dots V_k(t_k), \quad L = L(t_1, \dots, t_k) = V_k(t_k) \dots V_1(t_1) = R^*$$

and define

$$S = S(t_1, \dots, t_k) = L(t_1, \dots, t_k) \Sigma R(t_1, \dots, t_k), (t_1, \dots, t_k) \in [0, 1]^k,$$

$$\varphi(t_1, \dots, t_k) := \Sigma^{1/2} R(t_1, \dots, t_k) Dg(S(t_1, \dots, t_k)) L(t_1, \dots, t_k) \Sigma^{1/2}, (t_1, \dots, t_k) \in [0, 1]^k.$$

The following representation is basic in the analysis of functions  $\mathfrak{B}^k \mathfrak{D}g(\Sigma)$ .

**Proposition 3.** Suppose  $g \in L_\infty^O(\mathbb{C}_+(\mathbb{H}))$  is  $k+1$  times continuously differentiable function with uniformly bounded derivatives  $D^j g$ ,  $j = 1, \dots, k+1$ . Then the function  $\varphi$  is  $k$  times continuously differentiable in  $[0, 1]^k$  and

$$(17) \quad \mathfrak{B}^k \mathfrak{D}g(\Sigma) = \mathbb{E} \int_0^1 \cdots \int_0^1 \frac{\partial^k \varphi(t_1, \dots, t_k)}{\partial t_1 \cdots \partial t_k} dt_1 \cdots dt_k.$$

*Proof.* For  $\phi : [0, 1]^k \mapsto \mathbb{R}$ , define finite difference operators

$$\Delta_i \phi(t_1, \dots, t_k) := \phi(t_1, \dots, t_{i-1}, 1, t_{i+1}, \dots, t_k) - \phi(t_1, \dots, t_{i-1}, 0, t_{i+1}, \dots, t_k).$$

Then  $\Delta_1 \dots \Delta_k \phi$  is given by the following formula

$$(18) \quad \Delta_1 \dots \Delta_k \phi = \sum_{(t_1, \dots, t_k) \in \{0, 1\}^k} (-1)^{k-(t_1+\dots+t_k)} \phi(t_1, \dots, t_k).$$

It is well known that, if  $\phi$  is  $k$  times continuously differentiable in  $[0, 1]^k$ , then

$$(19) \quad \Delta_1 \dots \Delta_k \phi = \int_0^1 \cdots \int_0^1 \frac{\partial^k \phi(t_1, \dots, t_k)}{\partial t_1 \cdots \partial t_k} dt_1 \cdots dt_k.$$

Formula (19) also holds for vector- and operator-valued functions  $\phi$ . Identities (15) and (18) imply that

$$(20) \quad \mathfrak{B}^k \mathfrak{D}g(\Sigma) = \mathbb{E} \Delta_1 \dots \Delta_k \varphi.$$

Since  $Dg$  is  $k$  times continuously differentiable and functions  $S(t_1, \dots, t_k)$ ,  $R(t_1, \dots, t_k)$  are polynomials with respect to  $t_1, \dots, t_k$ , the function  $\varphi$  is  $k$  times continuously differentiable in  $[0, 1]^k$ . Representation (17) follows from (20) and (19).  $\square$

Representation (17) implies a bound on  $\|\mathfrak{B}^k \mathfrak{D}g(\Sigma)\|$  of the order  $O\left(\left(\frac{d}{n}\right)^{k/2}\right)$ .

**Theorem 6.** Suppose  $k \leq d \leq n$  and let  $g \in L_\infty^O(\mathbb{C}_+(\mathbb{H}))$  be a  $k+1$  times continuously differentiable function with uniformly bounded derivatives  $D^j g$ ,  $j = 1, \dots, k+1$ . Then the following bound holds for some constant  $C > 0$ :

$$(21) \quad \|\mathfrak{B}^k \mathfrak{D}g(\Sigma)\| \leq C^{k^2} \max_{1 \leq j \leq k+1} \|D^j g\|_{L_\infty} (\|\Sigma\|^{k+1} \vee \|\Sigma\|) \left(\frac{d}{n}\right)^{k/2}, \quad \Sigma \in \mathbb{C}_+(\mathbb{H}).^1$$

<sup>1</sup>Note that  $j$ -th derivative  $D^j g(\Sigma)$  can be viewed as symmetric  $j$ -linear form  $D^j g(\Sigma)(H_1, \dots, H_j)$ ,  $H_1, \dots, H_j \in \mathfrak{B}_{sa}(\mathbb{H})$ . The space of such  $j$ -linear forms  $\mathcal{M}(H_1, \dots, H_j)$  is equipped with operator norm:  $\|\mathcal{M}\| := \sup_{\|H_1\|, \dots, \|H_j\| \leq 1} |\mathcal{M}(H_1, \dots, H_j)|$ . The  $L_\infty$ -norm  $\|D^j g\|_{L_\infty}$  is then defined as  $\|D^j g\|_{L_\infty} := \sup_{\Sigma \in \mathbb{C}_+(\mathbb{H})} \|D^j g(\Sigma)\|$ .

The proof is based on deriving the following bound on the partial derivative  $\frac{\partial^k \varphi(t_1, \dots, t_k)}{\partial t_1 \dots \partial t_k}$  in (17):

(22)

$$\left\| \frac{\partial^k \varphi(t_1, \dots, t_k)}{\partial t_1 \dots \partial t_k} \right\| \leq 3^k 2^{k(2k+1)} \max_{1 \leq j \leq k+1} \|D^j g\|_{L^\infty} (\|\Sigma\|^{k+1} \vee \|\Sigma\|) \prod_{i=1}^k (1 + \delta_i)^{2k+1} \delta_i,$$

where  $\delta_i := \|W_i - I\|$ . Substituting (22) in (17), using independence of r.v.  $\delta_i$  and bound (4), one can complete the proof.

Representation (17) can be also used to study differentiability of function  $\mathfrak{B}^k \mathfrak{D}g(\Sigma)$  and to obtain bounds on the remainder of its Taylor expansion. In view of representation (10) and properties of operators  $\mathcal{T}$ ,  $\mathfrak{B}$ ,  $\mathfrak{D}$  (see Proposition 2), this could be further used to prove concentration bounds for the remainder of first order Taylor expansion  $\langle S_{f_k}(\Sigma; \hat{\Sigma} - \Sigma), B \rangle$ , to prove normal approximation bounds for  $\langle f_k(\hat{\Sigma}) - \mathbb{E}_\Sigma f_k(\hat{\Sigma}), B \rangle$  and bounds on the bias  $\langle \mathbb{E}_\Sigma f_k(\hat{\Sigma}) - f(\Sigma), B \rangle$ , leading to the proof of Theorem 5 (see Koltchinskii [2017] for more details).

## 6 Open Problems

We discuss below several open problems related to estimation of smooth functionals of covariance.

1. It would be of interest to study asymptotically efficient estimation of functionals  $\langle f(\Sigma), B \rangle$  in a dimension-free framework with effective rank playing the role of complexity parameter and in the classes of covariance operators not necessarily of isotropic type. The question is whether a version of Theorem 5 holds for the class  $\mathfrak{G}_{f,B}(r_n; a; \sigma_0)$  (instead of  $\mathfrak{S}_{f,B}(d_n; a; \sigma_0)$ ) with  $r_n \leq n^\alpha, \alpha \in (0, 1)$ . The main difficulty is to understand how to control  $k$ -th order difference  $\sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(\hat{\Sigma}^{(j)})$  of smooth function  $f$  along the trajectory of Bootstrap Chain in this case (compare with the approach outlined in Section 5).

2. Another problem is to understand whether the smoothness threshold  $s > \frac{1}{1-\alpha}$  for asymptotically efficient estimation of functionals  $\langle f(\Sigma), B \rangle$  is sharp (a similar problem was solved in Ibragimov, Nemirovski, and Khasminskii [1986] and Nemirovski [2000] in the case of Gaussian shift model).

3. It would be also of interest to study minimax optimal convergence rates of estimation of functionals  $\langle f(\Sigma), B \rangle$  in the case when the nuclear norm of operator  $B$  is not bounded by a constant. This includes, for instance, functionals  $\text{tr}(f(\Sigma))$  (the case of  $B = I$ ). In such problems, the  $\sqrt{n}$ -convergence rate no longer holds (see, for instance, the example of estimation of log-determinant Cai, Liang, and Zhou [2015] for which the rate becomes of the order  $\asymp \sqrt{\frac{n}{d}}$ ). A more general problem is to study minimax optimal convergence

rate of estimation of smooth orthogonally invariant functionals of  $\Sigma$ . The Bootstrap Chain bias reduction could be still relevant in such problems.

4. One more problem is to study estimation of smooth functionals under further “complexity” constraints (such as smoothness or sparsity) on the set of possible covariance operators. For instance, if  $\mathcal{S} \subset \mathcal{C}_+(\mathbb{H})$  is a set of covariance operators and  $\mathfrak{M}$  is a family of finite-dimensional subspaces of  $\mathbb{H}$ , the complexity of  $\mathcal{S}$  could be characterized by quantities

$$d_m(\mathcal{S}; \mathfrak{M}) := \inf_{L \in \mathfrak{M}, \dim(L) \leq m} \sup_{\Sigma \in \mathcal{S}} \|\Sigma - P_L \Sigma P_L\|, m \geq 1.$$

Assuming that the dimension  $d = \dim(\mathbb{H})$  satisfies the condition  $d \leq n^\alpha$  for some  $\alpha > 0$  and  $d_m(\mathcal{S}; \mathfrak{M}) \lesssim m^{-\beta}$  for some  $\beta > 0$ , the question is to determine threshold  $s(\alpha, \beta)$  such that asymptotically efficient estimation is possible for functionals  $\langle f(\Sigma), B \rangle$  of smoothness  $s > s(\alpha, \beta)$  (and impossible for some functionals of smoothness  $s < s(\alpha, \beta)$ ).

5. Asymptotically efficient estimator  $\langle f_k(\hat{\Sigma}), B \rangle$  in [Theorem 5](#) is based on an approximate solution of Wishart integral equation  $\mathcal{T}g(\Sigma) = f(\Sigma)$ . The Wishart operator  $\mathcal{T}$  is well studied in the literature on Wishart distribution (see, e.g., [James \[1961\]](#), [Letac and Massam \[2004\]](#)). In particular, it is known that zonal polynomials [James \[1961\]](#) are its eigenfunctions. It would be of interest to study other approaches to regularized approximate solution of Wishart equation (and corresponding estimators of such functionals as  $\langle f(\Sigma), B \rangle$ ) that would use more directly the spectral properties of operator  $\mathcal{T}$  (and could require the tools from analysis on symmetric cones [Faraud and Korányi \[1994\]](#) and [Gross and Richards \[1987\]](#)).

## References

- A. B. Aleksandrov and V. V. Peller (2016). “Operator Lipschitz functions”. *Uspekhi Mat. Nauk* 71.4(430), pp. 3–106. arXiv: [1611.01593](#). MR: [3588921](#) (cit. on p. [2929](#), [2930](#)).
- T. W. Anderson (2003). *An introduction to multivariate statistical analysis*. Third. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, pp. xx+721. MR: [1990662](#) (cit. on p. [2930](#)).
- Z. D. Bai and Jack W. Silverstein (2004). “CLT for linear spectral statistics of large-dimensional sample covariance matrices”. *Ann. Probab.* 32.1A, pp. 553–605. MR: [2040792](#) (cit. on p. [2924](#)).
- Peter J. Bickel, Chris A. J. Klaassen, Ya’acov Ritov, and Jon A. Wellner (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, p. 560. MR: [1245941](#) (cit. on p. [2923](#)).

- T. Tony Cai, Tengyuan Liang, and Harrison H. Zhou (2015). “Law of log determinant of sample covariance matrix and optimal estimation of differential entropy for high-dimensional Gaussian distributions”. *J. Multivariate Anal.* 137, pp. 161–172. MR: [3332804](#) (cit. on pp. [2924](#), [2938](#)).
- T. Tony Cai and Mark G. Low (2005a). “Nonquadratic estimators of a quadratic functional”. *Ann. Statist.* 33.6, pp. 2930–2956. MR: [2253108](#) (cit. on p. [2923](#)).
- (2005b). “On adaptive estimation of linear functionals”. *Ann. Statist.* 33.5, pp. 2311–2343. MR: [2211088](#) (cit. on p. [2923](#)).
- Olivier Collier, Laëtitia Comminges, and Alexandre B. Tsybakov (2017). “Minimax estimation of linear and quadratic functionals on sparsity classes”. *Ann. Statist.* 45.3, pp. 923–958. MR: [3662444](#) (cit. on p. [2923](#)).
- Jianqing Fan, Philippe Rigollet, and Weichen Wang (2015). “Estimation of functionals of sparse covariance matrices”. *Ann. Statist.* 43.6, pp. 2706–2737. MR: [3405609](#) (cit. on p. [2924](#)).
- Jacques Faraut and Adam Korányi (1994). *Analysis on symmetric cones*. Oxford Mathematical Monographs. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, pp. xii+382. MR: [1446489](#) (cit. on p. [2939](#)).
- R.A. Fisher (1922). “On the mathematical foundation of theoretical statistics”. *Philosophical Transactions of the Royal Society of London, Series A* 222, pp. 309–368 (cit. on p. [2922](#)).
- (1925). “Theory of statistical estimation”. *Proceedings of the Cambridge Philosophical Society* 22, pp. 700–725 (cit. on p. [2922](#)).
- Chao Gao and Harrison H. Zhou (2016). “Bernstein–von Mises theorems for functionals of the covariance matrix”. *Electron. J. Stat.* 10.2, pp. 1751–1806. MR: [3522660](#) (cit. on p. [2924](#)).
- S. van de Geer, P. Bühlmann, Ya. Ritov, and R. Dezeure (2014). “On asymptotically optimal confidence regions and tests for high-dimensional models”. *Ann. Statist.* 42.3, pp. 1166–1202. MR: [3224285](#) (cit. on p. [2923](#)).
- R. D. Gill and B. Y. Levit (1995). “Applications of the Van Trees inequality: a Bayesian Cramér-Rao bound”. *Bernoulli* 1.1–2, pp. 59–79. MR: [1354456](#) (cit. on p. [2922](#)).
- Evarist Giné and Richard Nickl (2016). *Mathematical foundations of infinite-dimensional statistical models*. Cambridge Series in Statistical and Probabilistic Mathematics, [40]. Cambridge University Press, New York, pp. xiv+690. MR: [3588285](#) (cit. on p. [2923](#)).
- V. L. Girko (1987). “An introduction to general statistical analysis”. *Teor. Veroyatnost. i Primenen.* 32.2, pp. 252–265. MR: [902754](#) (cit. on p. [2924](#)).
- (1995). *Statistical analysis of observations of increasing dimension*. Vol. 28. Theory and Decision Library. Series B: Mathematical and Statistical Methods. Translated from the Russian. Kluwer Academic Publishers, Dordrecht, pp. xxii+287. MR: [1473719](#) (cit. on p. [2924](#)).

- Kenneth I. Gross and Donald St. P. Richards (1987). “Special functions of matrix argument. I. Algebraic induction, zonal polynomials, and hypergeometric functions”. *Trans. Amer. Math. Soc.* 301.2, pp. 781–811. MR: [882715](#) (cit. on p. [2939](#)).
- Jaroslav Hájek (1972). “Local asymptotic minimax and admissibility in estimation”. In: *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability, (Univ. California, Berkeley, Calif., 1970/1971), Vol. I: Theory of statistics*. Ed. by Lucien Le Cam, Jerzy Neyman, and Elizabeth L. Scott. Univ. California Press, Berkeley, Calif., pp. 175–194. MR: [0400513](#) (cit. on p. [2922](#)).
- I. A. Ibragimov and R. Z. Khasminskii (1981). *Statistical estimation*. Vol. 16. Applications of Mathematics. Asymptotic theory, Translated from the Russian by Samuel Kotz. Springer-Verlag, New York-Berlin, pp. vii+403. MR: [620321](#) (cit. on p. [2923](#)).
- I. A. Ibragimov, A. S. Nemirovski, and R. Z. Khasminskii (1986). “Some problems of nonparametric estimation in Gaussian white noise”. *Teor. Veroyatnost. i Primenen.* 31.3, pp. 451–466. MR: [866866](#) (cit. on pp. [2923](#), [2938](#)).
- Alan T. James (1961). “Zonal polynomials of the real positive definite symmetric matrices”. *Ann. of Math. (2)* 74, pp. 456–469. MR: [0140741](#) (cit. on pp. [2932](#), [2939](#)).
- J. Janková and S. van de Geer (2016). “Semi-parametric efficiency bounds for high-dimensional models”. To appear in *Annals of Statistics* (cit. on p. [2923](#)).
- Adel Javanmard and Andrea Montanari (2014). “Hypothesis testing in high-dimensional regression under the Gaussian random design model: asymptotic theory”. *IEEE Trans. Inform. Theory* 60.10, pp. 6522–6554. MR: [3265038](#) (cit. on p. [2923](#)).
- Jiantao Jiao, Yanjun Han, and Tsachy Weissman (Sept. 2017). “Bias Correction with Jackknife, Bootstrap, and Taylor Series”. arXiv: [1709.06183](#) (cit. on p. [2932](#)).
- Iain M. Johnstone (2001). “On the distribution of the largest eigenvalue in principal components analysis”. *Ann. Statist.* 29.2, pp. 295–327. MR: [1863961](#) (cit. on p. [2925](#)).
- Iain M. Johnstone and Arthur Yu Lu (2009). “On consistency and sparsity for principal components analysis in high dimensions”. *J. Amer. Statist. Assoc.* 104.486, pp. 682–693. MR: [2751448](#) (cit. on p. [2925](#)).
- Vladimir Koltchinskii (Oct. 2017). “Asymptotically Efficient Estimation of Smooth Functionals of Covariance Operators”. arXiv: [1710.09072](#) (cit. on pp. [2929](#), [2930](#), [2932](#), [2934](#), [2938](#)).
- Vladimir Koltchinskii, Matthias Löffler, and Richard Nickl (Aug. 2017). “Efficient Estimation of Linear Functionals of Principal Components”. arXiv: [1708.07642](#) (cit. on pp. [2926](#)–[2928](#)).
- Vladimir Koltchinskii and Karim Lounici (2016). “Asymptotics and concentration bounds for bilinear forms of spectral projectors of sample covariance”. *Ann. Inst. Henri Poincaré Probab. Stat.* 52.4, pp. 1976–2013. MR: [3573302](#) (cit. on pp. [2926](#)–[2928](#)).
- (2017a). “Concentration inequalities and moment bounds for sample covariance operators”. *Bernoulli* 23.1, pp. 110–133. MR: [3556768](#) (cit. on pp. [2925](#), [2926](#)).

- Vladimir Koltchinskii and Karim Lounici (2017b). “New asymptotic results in principal component analysis”. *Sankhya A* 79.2, pp. 254–297. MR: [3707422](#) (cit. on p. 2926).
- (2017c). “Normal approximation and concentration of spectral projectors of sample covariance”. *Ann. Statist.* 45.1, pp. 121–157. MR: [3611488](#) (cit. on p. 2926).
- Lucien LeCam (1953). “On some asymptotic properties of maximum likelihood estimates and related Bayes’ estimates”. *Univ. California Publ. Statist.* 1, pp. 277–329. MR: [0054913](#) (cit. on p. 2922).
- Gérard Letac and Hélène Massam (2004). “All invariant moments of the Wishart distribution”. *Scand. J. Statist.* 31.2, pp. 295–318. MR: [2066255](#) (cit. on pp. 2932, 2935, 2939).
- B. Y. Levit (1975). “The efficiency of a certain class of nonparametric estimators”. *Teor. Veroyatnost. i Primenen.* 20.4, pp. 738–754. MR: [0403052](#) (cit. on p. 2923).
- (1978). “Asymptotically efficient estimation of nonlinear functionals”. *Probl. Peredachi Inf. (Problems of Information Transmission)* 14.3, pp. 65–72. MR: [533450](#) (cit. on p. 2923).
- A. Lytova and L. Pastur (2009). “Central limit theorem for linear eigenvalue statistics of random matrices with independent entries”. *Ann. Probab.* 37.5, pp. 1778–1840. MR: [2561434](#) (cit. on p. 2924).
- A. S. Nemirovski (1990). “Necessary conditions for efficient estimation of functionals of a nonparametric signal observed in white noise”. *Teor. Veroyatnost. i Primenen.* 35.1, pp. 83–91. MR: [1050056](#) (cit. on p. 2923).
- (2000). “Topics in non-parametric statistics”. In: *Lectures on probability theory and statistics (Saint-Flour, 1998)*. Vol. 1738. Lecture Notes in Math. Springer, Berlin, pp. 85–277. MR: [1775640](#) (cit. on pp. 2923, 2938).
- Debashis Paul (2007). “Asymptotics of sample eigenstructure for a large dimensional spiked covariance model”. *Statist. Sinica* 17.4, pp. 1617–1642. MR: [2399865](#) (cit. on p. 2925).
- V. V. Peller (1985). “Hankel operators in the theory of perturbations of unitary and self-adjoint operators”. *Funktsional. Anal. i Prilozhen.* 19.2, pp. 37–51, 96. MR: [800919](#) (cit. on p. 2929).
- (2006). “Multiple operator integrals and higher operator derivatives”. *J. Funct. Anal.* 233.2, pp. 515–544. MR: [2214586](#) (cit. on p. 2929).
- Vilmos Totik (1994). “Approximation by Bernstein polynomials”. *Amer. J. Math.* 116.4, pp. 995–1018. MR: [1287945](#) (cit. on p. 2932).
- Hans Triebel (1983). *Theory of function spaces*. Vol. 78. Monographs in Mathematics. Birkhäuser Verlag, Basel, p. 284. MR: [781540](#) (cit. on p. 2929).
- Roman Vershynin (2012). “Introduction to the non-asymptotic analysis of random matrices”. In: *Compressed sensing*. Cambridge Univ. Press, Cambridge, pp. 210–268. MR: [2963170](#) (cit. on p. 2925).

Cun-Hui Zhang and Stephanie S. Zhang (2014). “[Confidence intervals for low dimensional parameters in high dimensional linear models](#)”. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 76.1, pp. 217–242. MR: [3153940](#) (cit. on p. [2923](#)).

Received 2017-11-29.

VLADIMIR KOLTCHINSKII  
SCHOOL OF MATHEMATICS  
GEORGIA INSTITUTE OF TECHNOLOGY  
ATLANTA, GA 30332-0160  
USA  
[vlad@math.gatech.edu](mailto:vlad@math.gatech.edu)





# CONCENTRATION OF RANDOM GRAPHS AND APPLICATION TO COMMUNITY DETECTION

CAN M. LE, ELIZAVETA LEVINA AND ROMAN VERSHYNIN

## Abstract

Random matrix theory has played an important role in recent work on statistical network analysis. In this paper, we review recent results on regimes of concentration of random graphs around their expectation, showing that dense graphs concentrate and sparse graphs concentrate after regularization. We also review relevant network models that may be of interest to probabilists considering directions for new random matrix theory developments, and random matrix theory tools that may be of interest to statisticians looking to prove properties of network algorithms. Applications of concentration results to the problem of community detection in networks are discussed in detail.

## 1 Introduction

A lot of recent interest in concentration of random graphs has been generated by problems in network analysis, a very active interdisciplinary research area with contributions from probability, statistics, physics, computer science, and the social sciences all playing a role. Networks represent relationships (edges) between objects (nodes), and a network between  $n$  nodes is typically represented by its  $n \times n$  adjacency matrix  $A$ . We will focus on the case of simple undirected networks, where  $A_{ij} = 1$  when nodes  $i$  and  $j$  are connected by an edge, and 0 otherwise, which makes  $A$  a symmetric matrix with binary entries. It is customary to assume the graph contains no self-loops, that is,  $A_{ii} = 0$  for all  $i$ , but this is not crucial. In general, networks may be directed ( $A$  is not symmetric), weighted (the entries of  $A$  have a numerical value representing the strength of connection), and/or signed (the entries of  $A$  have a sign representing whether the relationship is positive or negative in some sense).

Viewing networks as random realizations from an underlying network model enables analysis and inference, with the added difficulty that we often only observe a single realization of a given network. Quantities of interest to be inferred from this realization

may include various network properties such as the node degree distribution, the network radius, and community structure. Fundamental to these inferences is the question of how close a single realization of the matrix  $A$  is to the population mean, or the true model,  $\mathbb{E} A$ . If  $A$  is close to  $\mathbb{E} A$ , that is,  $A$  concentrates around its mean, then inferences drawn from  $A$  can be transferred to the population with high probability.

In this paper, we aim to answer the question “When does  $A$  concentrate around  $\mathbb{E} A$ ?” under a number of network models and asymptotic regimes. We also show that in some cases when the network does not concentrate, a simple regularization step can restore concentration. While the question of concentration is interesting in its own right, we especially focus on the implications for the problem of community detection, a problem that has attracted a lot of attention in the networks literature. When concentration holds, in many cases a simple spectral algorithm can recover communities, and thus concentration is of practical and not only theoretical interest.

## 2 Random network models

Our concentration results hold for quite general models, but, for the sake of clarity, we provide a brief review of network models, starting from the simplest model and building up in complexity.

**The Erdős–Rényi (ER) graph.** The simplest random network model is the Erdős–Rényi graph  $G(n, p)$  [Erdős and Rényi \[1959\]](#). Under this model, edges are independently drawn between all pairs of nodes according to a Bernoulli distribution with success probability  $p$ . Although the ER model provides an important building block in network modeling and is attractive to analyze, it almost never fits network data observed in practice.

**The stochastic block model (SBM).** The SBM is perhaps the simplest network model with community structure, first proposed by [Holland, Laskey, Leinhardt, and \[1983\]](#). Under this model, each node belongs to exactly one of  $K$  communities, and the node community membership  $c_i$  is drawn independently from a multinomial distribution on  $\{1, \dots, K\}$  with probabilities  $\pi_1, \dots, \pi_K$ . Conditional on the label vector  $c$ , edges are drawn independently between each pair of nodes  $i, j$ , with

$$\mathbb{P}(A_{ij} = 1) = B_{c_i c_j},$$

where  $B$  is a symmetric  $K \times K$  matrix controlling edge probabilities. Note that each community within SBM is an ER graph. The main question of interest in network analysis is estimating the label vector  $c$  from  $A$ , although model parameters  $\pi$  and  $P$  may also be of interest.

**The degree-corrected stochastic block model (DCSBM).** While the SBM does incorporate community structure, the assumption that each block is an ER graph is too restrictive for many real-world networks. In particular, ER graphs have a Poisson degree distribution, and real networks typically fit the power law or another heavy-tailed distribution better, since they often have “hubs”, influential nodes with many connections. An extension removing this limitation, the degree-corrected stochastic block model (DCSBM) was proposed by [Karrer and Newman \[2011\]](#). The DCSBM is like an SBM but with each node assigned an additional parameter  $\theta_i > 0$  that controls its expected degree, and edges drawn independently with

$$\mathbb{P}(A_{ij} = 1) = \theta_i \theta_j B_{c_i c_j}.$$

Additional constraints need to be imposed on  $\theta_i$  for model identifiability; see [Karrer and Newman \[2011\]](#) and [Zhao, Levina, and Zhu \[2012\]](#) for options.

**The latent space model (LSM).** Node labels under the SBM or the DCSBM can be thought of as latent (unobserved) node positions in a discrete space of  $K$  elements. More generally, latent positions can be modeled as coordinates in  $\mathbb{R}^d$ , or another set equipped with a distance measure. The LSM [Hoff, Raftery, and Handcock \[2002\]](#) assumes that each node  $i$  is associated with an unknown position  $x_i$  and edges are drawn independently between each pair of nodes  $i, j$  with probability inversely proportional to the distance between  $x_i$  and  $x_j$ . If latent positions  $x_i$  form clusters (for example, if they are drawn from a mixture of Gaussians), then a random network generated from this model exhibits community structure. Inferring the latent positions can in principle lead to insights into how the network was formed, beyond simple community assignments.

**Exchangeable random networks.** An analogue of de Finetti’s theorem for networks, due to Hoover and Aldous [Hoover \[1979\]](#) and [Aldous \[1981\]](#), shows that any network whose distribution is invariant under node permutations can be represented by

$$A_{ij} = g(\alpha, \xi_i, \xi_j, \lambda_{ij}),$$

where  $\alpha$ ,  $\xi_i$  and  $\xi_j$  are independent and uniformly distributed on  $[0, 1]$ , and  $g(u, v, w, z) = g(u, w, v, z)$  for all  $u, v, w, z$ . This model covers all the previously discussed models as special cases, and the function  $g$ , called the graphon, can be estimated up to a permutation under additional assumptions; see [Olhede and Wolfe \[2013\]](#), [Gao, Lu, and Zhou \[2015\]](#), and [Y. Zhang, Levina, and Zhu \[2017\]](#).

**Network models with overlapping communities.** In practice, it is often more reasonable to allow nodes to belong to more than one community. Multiple such models have

been proposed, including the Mixed Membership Stochastic Block Model (MMSBM) [Airoldi, Blei, Fienberg, and Xing \[2008\]](#), the Ball-Karrer-Newman Model (BKN) [Ball, Karrer, and Newman \[2011\]](#), and the OCCAM model [Y. Zhang, Levina, and Zhu \[2014\]](#). MMSBM allows different memberships depending on which node the given node interacts with; the BKN models edges as a sum of multiple edges corresponding to different communities; and OCCAM relaxes the membership vector  $c$  under the SBM to have entries between 0 and 1 instead of exactly one “1”. All of these models are also covered by the theory we present, because, conditional on node memberships, all these networks are distributed according to an inhomogeneous Erdős–Rényi model, the most general model we consider, described next.

**The inhomogeneous Erdős–Rényi model.** All models described above share an important property: conditioned on node latent positions, edges are formed independently. The most general form of such a model is the inhomogeneous Erdős–Rényi model (IERM) [Bollobás, Janson, and Riordan \[2007\]](#), where each edge is independently drawn, with  $\mathbb{P}(A_{ij} = 1) = P_{ij}$ , where  $P = (P_{ij}) = \mathbb{E} A$ . Evidently, additional assumptions have to be made if latent positions of nodes (however they are defined) are to be recovered from a single realization of  $A$ . We will state concentration results under the IERM as generally as possible, and then discuss additional assumptions under which latent positions can also be estimated reliably.

**Scaling.** We have so far defined all the models for a fixed number of nodes  $n$ , but in order to talk about concentration, we need to determine how the expectation  $P_n = \mathbb{E} A_n$  changes with  $n$ . Most of the literature defines

$$P_n = \rho_n P$$

where  $P$  is a matrix with constant non-negative entries, and  $\rho_n$  controls the average expected degree of the network,  $d = d_n = n\rho_n$ . Different asymptotic regimes have been studied, especially under the SBM; see [Abbe \[2017\]](#) for a review. Unless  $\rho_n \rightarrow 0$ , the average network degree  $d = \Omega(n)$ , and the network becomes dense as  $n$  grows. In the SBM literature, the regime  $d_n \gg \log n$  is sometimes referred to as semi-dense;  $d_n \rightarrow \infty$  but not faster than  $\log n$  is semi-sparse; and the constant degree regime  $d_n = O(1)$  is called sparse. We will elaborate on these regimes and their implications later on in the paper.

### 3 Useful random matrix results

We start from presenting a few powerful and general tools in random matrix theory which can help prove concentration bounds for random graphs.

**Theorem 3.1** (Bai-Yin law [Bai and Y. Q. Yin \[1988\]](#); see [Füredi and Komlós \[1981\]](#) an for earlier result). *Let  $M = (M_{ij})_{i,j=1}^\infty$  be an infinite, symmetric, and diagonal-free random matrix whose entries above the diagonal are i.i.d. random variables with zero mean and variance  $\sigma^2$ . Suppose further that  $\mathbb{E} M_{ij}^4 < \infty$ . Let  $M_n = (M_{ij})_{i,j=1}^n$  denote the principal minors of  $M$ . Then, as  $n \rightarrow \infty$ ,*

$$(3-1) \quad \frac{1}{\sqrt{n}} \|M_n\| \rightarrow 2 \quad \text{almost surely.}$$

**Theorem 3.2** (Matrix Bernstein's inequality). *Let  $X_1, \dots, X_N$  be independent, mean zero,  $n \times n$  symmetric random matrices, such that  $\|X_i\| \leq K$  almost surely for all  $i$ . Then, for every  $t \geq 0$  we have*

$$\mathbb{P} \left\{ \left\| \sum_{i=1}^N X_i \right\| \geq t \right\} \leq 2n \exp \left( - \frac{t^2/2}{\sigma^2 + Kt/3} \right).$$

Here  $\sigma^2 = \left\| \sum_{i=1}^N \mathbb{E} X_i^2 \right\|$  is the norm of the “matrix variance” of the sum.

**Corollary 3.3** (Expected norm of sum of random matrices). *We have*

$$\mathbb{E} \left\| \sum_{i=1}^N X_i \right\| \lesssim \sigma \sqrt{\log n} + K \log n.$$

The following result gives sharper bounds on random matrices than matrix Bernstein's inequality, but requires independence of entries.

**Theorem 3.4** (Bandeira-van Handel [Bandeira and van Handel \[2016\]](#) Corollary 3.6). *Let  $M$  be an  $n \times n$  symmetric random matrix with independent entries on and above the diagonal. Then*

$$\mathbb{E} \|M\| \lesssim \max_i \left( \sum_j \sigma_{ij}^2 \right)^{1/2} + \sqrt{\log n} \max_{i,j} K_{ij},$$

where  $\sigma_{ij}^2 = \mathbb{E} M_{ij}^2$  are the variances of entries and  $K_{ij} = \|M_{ij}\|_\infty$ .

**Theorem 3.5** (Seginer's theorem [Seginer \[2000\]](#)). *Let  $M$  be a  $n \times n$  symmetric random matrix with i.i.d. mean zero entries above the diagonal and arbitrary entries on the diagonal. Then*

$$\mathbb{E} \|M\| \asymp \mathbb{E} \max_i \|M_i\|_2$$

where  $M_i$  denote the columns of  $M$ .

The lower bound in Seginer’s theorem is trivial; it follows from the fact that the operator norm of a matrix is always bounded below by the Euclidean norm of any of its columns. The original paper of Seginer [Seginer \[2000\]](#) proved the upper bound for non-symmetric matrices with independent entries. The present statement of [Theorem 3.5](#) can be derived by a simple symmetrization argument, see [Hajek, Wu, and Xu \[2016, Section 4.1\]](#).

## 4 Dense networks concentrate

If  $A = A_n$  is the adjacency matrix of a  $G(n, p)$  random graph with a constant  $p$ , then the Bai-Yin law gives

$$\frac{1}{\sqrt{n}} \|A - \mathbb{E} A\| \rightarrow 2\sqrt{p(1-p)}.$$

In particular, we have

$$(4-1) \quad \|A - \mathbb{E} A\| \leq 2\sqrt{d}$$

with probability tending to one, where  $d = np$  is the expected node degree.

Can we expect a similar concentration for sparser Erdős–Rényi graphs, where  $p$  is allowed to decrease with  $n$ ? The method of [Friedman, Kahn, and Szemerédi \[1989\]](#) adapted by [Feige and Ofek \[2005\]](#) gives

$$(4-2) \quad \|A - \mathbb{E} A\| = O(\sqrt{d})$$

under the weaker condition  $d \gtrsim \log n$ , which is optimal, as we will see shortly. This argument actually yields (4-2) for inhomogeneous random graphs  $G(n, (p_{ij}))$  as well, and for  $d = \max_{ij} np_{ij}$ , see e.g. [Lei and Rinaldo \[2015\]](#) and [Chin, Rao, and V. Vu \[2015\]](#).

Under a weaker assumption  $d = np \gg \log^4 n$ , Vu [V. H. Vu \[2007\]](#) proved a sharper bound for  $G(n, p)$ , namely

$$(4-3) \quad \|A - \mathbb{E} A\| = (2 + o(1))\sqrt{d},$$

which essentially extends (4-1) to sparse random graphs. Very recently, [Benaych-Georges, Bordenave, and Knowles \[2017b\]](#) were able to derive (4-3) under the optimal condition  $d \gg \log n$ . More precisely, they showed that if  $4 \leq d \leq n^{2/13}$ , then

$$\mathbb{E} \|A - \mathbb{E} A\| \leq 2\sqrt{d} + C \sqrt{\frac{\log n}{1 + \log(\log(n)/d)}}.$$

The argument of [Benaych-Georges, Bordenave, and Knowles \[ibid.\]](#) applies more generally to inhomogeneous random graphs  $G(n, (p_{ij}))$  under a regularity condition on the

connection probabilities  $(p_{ij})$ . It even holds for more general random matrices that may not necessarily have binary entries.

To apply [Corollary 3.3](#) to the adjacency matrix  $A$  of an ER random graph  $G(n, p)$ , decompose  $A$  into a sum of independent random matrices  $A = \sum_{i \leq j} X_{ij}$ , where each matrix  $X_{ij}$  contains a pair of symmetric entries of  $A$ , i.e.  $X_{ij} = A_{ij}(e_i e_i^\top + e_j e_j^\top)$  where  $(e_i)$  denotes the canonical basis in  $\mathbb{R}^n$ . Then apply [Corollary 3.3](#) to the sum of mean zero matrices  $X_{ij} - p$ . It is quick to check that  $\sigma^2 \leq pn$  and obviously  $K \leq 2$ , and so we conclude that

$$(4-4) \quad \mathbb{E} \|A - \mathbb{E} A\| \lesssim \sqrt{d \log n} + \log n,$$

where  $d = np$  is the expected degree. The same argument applies more generally to inhomogeneous random graphs  $G(n, (p_{ij}))$ , and it still gives (4-4) when

$$d = \max_i \sum_j p_{ij}$$

is the maximal expected degree.

The logarithmic factors in bound (4-4) are not optimal, and can be improved by applying the result of Bandeira and van Handel ([Theorem 3.4](#)) to the centered adjacency matrix  $A - \mathbb{E} A$  of an inhomogeneous random graph  $G(n, (p_{ij}))$ . In this case,  $\sigma_{ij}^2 = p_{ij}$  and  $K_{ij} \leq 1$ , so we obtain the following sharpening of (4-4).

**Proposition 4.1** (Concentration of inhomogeneous random graphs). *Let  $A$  be the adjacency matrix of an inhomogeneous random graph  $G(n, (p_{ij}))$ . Then*

$$(4-5) \quad \mathbb{E} \|A - \mathbb{E} A\| \lesssim \sqrt{d} + \sqrt{\log n},$$

where  $d = \max_i \sum_j p_{ij}$  is the expected maximal degree.

In particular, if the graph is not too sparse, namely  $d \gtrsim \log n$ , then the optimal concentration (4-3) holds, i.e.

$$\mathbb{E} \|A - \mathbb{E} A\| \lesssim \sqrt{d}.$$

This recovers a result of [Feige and Ofek \[2005\]](#).

A similar bound can be alternatively proved using the general result of Seginer ([Theorem 3.5](#)). If  $A$  is the adjacency matrix of  $G(n, p)$ , it is easy to check that  $\mathbb{E} \max_i \|A_i\|_2 \lesssim \sqrt{d} + \sqrt{\log n}$ . Thus, Seginer's theorem implies the optimal concentration bound (4-5) as well. Using simple convexity arguments, one can extend this to inhomogeneous random graphs  $G(n, (p_{ij}))$ , and get the bound (4-5) for  $d = \max_{ij} np_{ij}$ , see [Hajek, Wu, and Xu \[2016, Section 4.1\]](#).



One may wonder if Seginer’s theorem holds for matrices with independent but not identically distributed entries. Unfortunately, this is not the case in general; a simple counterexample was found by Seginer [Seginer \[2000\]](#), see [Bandeira and van Handel \[2016, Remark 4.8\]](#). Nevertheless, it is an open conjecture of Latala that Seginer’s theorem does hold if  $M$  has independent *Gaussian* entries, see the papers [Riemer and Schütt \[2013\]](#) and [van Handel \[2017a\]](#) and the survey [van Handel \[2017b\]](#).

## 5 Sparse networks concentrate after regularization

**5.1 Sparse networks do not concentrate.** In the sparse regime  $d = np \ll \log n$ , the Bai-Yin’s law for  $G(n, p)$  fails. This is because in this case, degrees of some vertices are much higher than the expected degree  $d$ . This causes some rows of the adjacency matrix  $A$  to have Euclidean norms much larger than  $\sqrt{d}$ , which in turn gives

$$\|A - \mathbb{E} A\| \gg \sqrt{d}.$$

In other words, concentration fails for very sparse graphs; there exist outlying eigenvalues that escape the interval  $[-2, 2]$  where the spectrum of denser graphs lies according to (3-1). For precise description of this phenomenon, see the original paper [Krivelevich and Sudakov \[2003\]](#), a discussion in [Bandeira and van Handel \[2016, Section 4\]](#) and the very recent work [Benaych-Georges, Bordenave, and Knowles \[2017a\]](#).

**5.2 Sparse networks concentrate after regularization.** One way to regularize a random network in the sparse regime is to remove high degree vertices altogether from the network. Indeed, [Feige and Ofek \[2005\]](#) showed that for  $G(n, p)$ , if we drop all vertices with degrees, say, larger than  $2d$ , then the remaining part of the network satisfies  $\|A - \mathbb{E} A\| = O(\sqrt{d})$  with high probability. The argument in [Feige and Ofek \[ibid.\]](#) is based on the method developed by [Friedman, Kahn, and Szemerédi \[1989\]](#) and it is extended to the IERM in [Lei and Rinaldo \[2015\]](#) and [Chin, Rao, and V. Vu \[2015\]](#).

Although removal of high degree vertices restores concentration, in practice this is a bad idea, since the loss of edges associated with “hub” nodes in an already sparse network leads to a considerable loss of information, and in particular community detection tends to break down. A more gentle regularization proposed in [Le, Levina, and Vershynin \[2017\]](#) does not remove high degree vertices, but reduces the weights of their edges just enough to keep the degrees bounded by  $O(d)$ .

**Theorem 5.1** (Concentration of regularized adjacency matrices). *Consider a random graph from the inhomogeneous Erdős–Rényi model  $G(n, (p_{ij}))$ , and let  $d = \max_{ij} np_{ij}$ .*

*Consider any subset of at most  $10n/d$  vertices, and reduce the weights of the edges incident to those vertices in an arbitrary way, but so that all degrees of the new (weighted)*

network become bounded by  $2d$ . For any  $r \geq 1$ , with probability at least  $1 - n^{-r}$  the adjacency matrix  $A'$  of the new weighted graph satisfies

$$\|A' - \mathbb{E} A\| \leq Cr^{3/2}\sqrt{d}.$$

Proving concentration for this kind of general regularization requires different tools. One key result we state next is the Grothendieck-Pietsch factorization, a general and well-known result in functional analysis [Pietsch \[1980\]](#), [Pisier \[1986\]](#), [Tomczak-Jaegermann \[1989\]](#), and [Pisier \[2012\]](#) which has already been used in a similar probabilistic context [Ledoux and Talagrand \[1991, Proposition 15.11\]](#). It compares two matrix norms, the spectral norm  $\ell_2 \rightarrow \ell_2$  and the  $\ell_\infty \rightarrow \ell_2$  norm.

**Theorem 5.2** (Grothendieck-Pietsch factorization). *Let  $B$  be a  $k \times m$  real matrix. Then there exist positive weights  $\mu_j$  with  $\sum_{j=1}^m \mu_j = 1$  such that*

$$\|B\|_{\infty \rightarrow 2} \leq \|BD_\mu^{-1/2}\| \leq 2\|B\|_{\infty \rightarrow 2},$$

where  $D_\mu = \text{diag}(\mu_j)$  denotes the  $m \times m$  diagonal matrix with weights  $\mu_j$  on the diagonal.

**Idea of the proof of Theorem 5.1 by network decomposition.** The argument in [Feige and Ofek \[2005\]](#) becomes very complicated for handling the general regularization in [Theorem 5.1](#). A simpler alternative approach was developed by [Le, Levina, and Vershynin \[2017\]](#) for proving [Theorem 5.1](#). The main idea is to decompose the set of entries  $[n] \times [n]$  into different subsets with desirable properties. There exists a partition (see [Figure 1c](#) for illustration)

$$[n] \times [n] = \mathcal{N} \cup \mathcal{R} \cup \mathcal{C}$$

such that  $A$  concentrates on  $\mathcal{N}$  even without regularization, while restrictions of  $A$  onto  $\mathcal{R}$  and  $\mathcal{C}$  have small row and column sums, respectively. It is easy to see that the degree regularization does not destroy the properties of  $\mathcal{N}$ ,  $\mathcal{R}$  and  $\mathcal{C}$ . Moreover, it creates a new property, allowing for controlling the columns of  $\mathcal{R}$  and rows of  $\mathcal{C}$ . Together with the triangle inequality, this implies the concentration of the entire network.

The network decomposition is constructed by an iterative procedure. We first establish concentration of  $A$  in  $\ell_\infty \rightarrow \ell_2$  norm using standard probability techniques. Next, we upgrade this to concentration in the spectral norm  $\|(A - \mathbb{E} A)_{\mathcal{N}_0}\| = O(\sqrt{d})$  on an appropriate (large) subset  $\mathcal{N}_0 \subseteq [n] \times [n]$  using the Grothendieck-Pietsch factorization ([Theorem 5.2](#)). It remains to control  $A$  on the complement of  $\mathcal{N}_0$ . That set is small; it can be described as a union of a block  $\mathcal{C}_0$  with a small number of rows, a block  $\mathcal{R}_0$  with a small number of columns and an exceptional (small) block (see [Figure 1a](#)). Now we

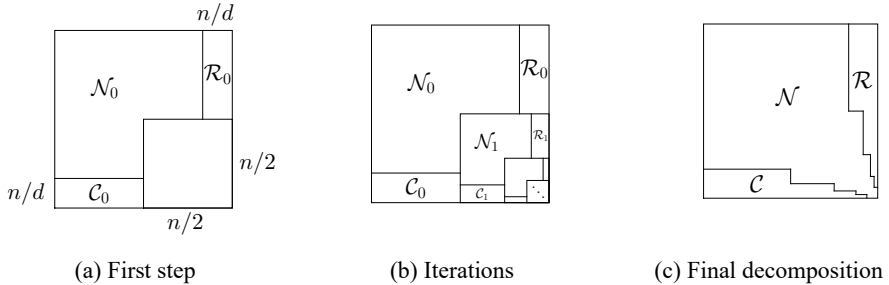


Figure 1: Constructing network decomposition iteratively.

repeat the process for the exceptional block, decomposing it into  $\mathcal{N}_1$ ,  $\mathcal{R}_1$ , and  $\mathcal{C}_1$ , and so on, as shown in Figure 1b. At the end, we set  $\mathcal{N} = \cup_i \mathcal{N}_i$ ,  $\mathcal{R} = \cup_i \mathcal{R}_i$  and  $\mathcal{C} = \cup_i \mathcal{C}_i$ . The cumulative error from this iterative procedure can be controlled appropriately; see [Le, Levina, and Vershynin \[2017\]](#) for details.

**5.3 Concentration of the graph Laplacian.** So far, we have looked at random graphs through the lens of their adjacency matrices. Another matrix that captures the structure of a random graph is the Laplacian. There are several ways to define the Laplacian; we focus on the symmetric, normalized Laplacian,

$$\mathcal{L}(A) = D^{-1/2} A D^{-1/2}.$$

Here  $D = \text{diag}(d_i)$  is the diagonal matrix with degrees  $d_i = \sum_{j=1}^n A_{ij}$  on the diagonal. The reader is referred to [F. R. K. Chung \[1997\]](#) for an introduction to graph Laplacians and their role in spectral graph theory. Here we mention just two basic facts: the spectrum of  $\mathcal{L}(A)$  is a subset of  $[-1, 1]$ , and the largest eigenvalue is always one.

In the networks literature in particular, community detection has been mainly done through spectral clustering on the Laplacian, not on the adjacency matrix. We will discuss this in more detail in [Section 6](#), but the primary reason for this is degree normalization: as discussed in [Section 2](#), real networks rarely have the Poisson or mixture of Poissons degree distribution that characterizes the stochastic block model; instead, “hubs”, or high degree vertices, are common, and they tend to break down spectral clustering on the adjacency matrix itself.

Concentration of Laplacians of random graphs has been studied by [S. Yin \[2008\]](#), [Chaudhuri, F. Chung, and Tsiatas \[2012\]](#), [Qin and Rohe \[2013\]](#), [Joseph and Yu \[2016\]](#), and [Gao, Ma, A. Y. Zhang, and Zhou \[2017\]](#). Just like the adjacency matrix, the Laplacian is known to concentrate in the dense regime  $d = \Omega(\log n)$ , and it fails to concentrate in the sparse regime. However, the reasons it fails to concentrate are different. For the

adjacency matrix, as we discussed, concentration fails in the sparse case because of high degree vertices. For the Laplacian, it is the low degree vertices that destroy concentration. In fact, it is easy to check that when  $d = o(\log n)$ , the probability of isolated vertices is non-vanishing; and each isolated vertex contributes an eigenvalue of 0 to the spectrum of  $\mathcal{L}(A)$ , which is easily seen to destroy concentration.

Multiple ways to regularize the Laplacian in order to deal with the low degree vertices have been proposed. Perhaps the two most common ones are adding a small constant to all the degrees on the diagonal of  $D$  [Chaudhuri, F. Chung, and Tsias \[2012\]](#), and adding a small constant to all the entries of  $A$  before computing the Laplacian. Here we focus on the latter regularization, proposed by [Amini, Chen, Bickel, and Levina \[2013\]](#) and analyzed by [Joseph and Yu \[2016\]](#) and [Gao, Ma, A. Y. Zhang, and Zhou \[2017\]](#). Choose  $\tau > 0$  and add the same number  $\tau/n$  to all entries of the adjacency matrix  $A$ , thereby replacing it with

$$(5-1) \quad A_\tau := A + \frac{\tau}{n} \mathbf{1}\mathbf{1}^\top$$

Then compute the Laplacian as usual using this new adjacency matrix. This regularization raises all degrees  $d_i$  to  $d_i + \tau$ , and eliminates isolated vertices, making the entire graph connected. The original paper [Amini, Chen, Bickel, and Levina \[2013\]](#) suggested the choice  $\tau = \rho \bar{d}$ , where  $\bar{d}$  is the average node degree and  $\rho \in (0, 1)$  is a constant. They showed the estimator is not particularly sensitive to  $\rho$  over a fairly wide range of values away from 0 (too little regularization) and 1 (too much noise). The choice of  $\rho = 0.25$  was recommended by [Amini, Chen, Bickel, and Levina \[ibid.\]](#) but this parameter can also be successfully chosen by cross-validation on the network [T. Li, Levina, and Zhu \[2016\]](#).

The following consequence of [Theorem 5.1](#) shows that regularization (5-1) indeed forces the Laplacian to concentrate.

**Theorem 5.3** (Concentration of the regularized Laplacian). *Consider a random graph drawn from the inhomogeneous Erdős–Rényi model  $G(n, (p_{ij}))$ , and let  $d = \max_{ij} np_{ij}$ . Choose a number  $\tau > 0$ . Then, for any  $r \geq 1$ , with probability at least  $1 - e^{-r}$  we have*

$$\|\mathcal{L}(A_\tau) - \mathcal{L}(\mathbb{E} A_\tau)\| \leq \frac{Cr^2}{\sqrt{\tau}} \left(1 + \frac{d}{\tau}\right)^{5/2}.$$

In the next section, we discuss why concentration of the adjacency matrix and/or its Laplacian is important in the context of community detection, the primary application of concentration in network analysis.

## 6 Application to community detection

Concentration of random graphs has been of such interest in networks analysis primarily because it relates to the problem of community detection; see [Fortunato \[2010\]](#), [Goldenberg, Zheng, Fienberg, Airoldi, et al. \[2010\]](#), and [Abbe \[2017\]](#) for reviews of community detection algorithms and results. We should specify that, perhaps in a slight misnomer, “community detection” refers to the task of assigning each node to a community (typically one and only one), not to the question of whether there are communities present, which might be a more natural use of the term “detection”.

Most of the theoretical work linking concentration of random graphs to community detection has focused on the stochastic block model (SBM), defined in [Section 2](#), which is one of the many special cases of the general IERM we consider. For the purpose of this paper, we focus on the simplest version of the SBM for which the largest number of results has been obtained so far, also known as the balanced planted partition model  $G(n, \frac{a}{n}, \frac{b}{n})$ . In this model, there are  $K = 2$  equal-sized communities with  $n/2$  nodes each. Edges between vertices within the same community are drawn independently with probability  $a/n$ , and edges between vertices in different communities with probability  $b/n$ . The task is to recover the community labels of vertices from a single realization of the adjacency matrix  $A$  drawn from this model. The large literature on both the recovery algorithms and the theory establishing when a recovery is possible is very nicely summarized in the recent excellent review [Abbe \[2017\]](#), where we refer the reader for details and analogues for a general  $K$  (now available for most results) and the asymmetric SBM (very few are available). In the following subsections we give a brief summary for the symmetric  $K = 2$  case which does not aim to be exhaustive.

**6.1 Community detection phase transition.** *Weak recovery*, sometimes also called detection, means performing better than randomly guessing the labels of vertices. The phase transition threshold for weak recovery was first conjectured in the physics literature by [Decelle, Krzakala, Moore, and Zdeborová \[2011\]](#), and proved rigorously by [Mossel, Neeman, and Sly \[2013, 2015, 2014\]](#), with follow-up and related work by [Abbe, Bandeira, and Hall \[2016\]](#), [Massoulié \[2014\]](#), and [Bordenave, Lelarge, and Massoulié \[2015\]](#). The phase transition result says that there exists a polynomial time algorithm which can classify more than 50% of the vertices correctly as  $n \rightarrow \infty$  with high probability if and only if

$$(a - b)^2 > 2(a + b).$$

Performing better than random guessing is the weakest possible guarantee of performance, which is of interest in the very sparse regime of  $d = (a + b)/2 = O(1)$ ; when the degree grows, weak recovery becomes trivial. This regime has been mostly studied by physicists and probabilists; in the statistics literature, *consistency* has been of more interest.

**6.2 Consistency of community detection.** Two types of consistency have been discussed in the literature. Strong consistency, also known as exact recovery, means labeling *all* vertices correctly with high probability, which is, as the name suggests, a very strong requirement. Weak consistency, or “almost exact” recovery, is the weaker and arguably more practically reasonable requirement that the fraction of misclassified vertices goes to 0 as  $n \rightarrow \infty$  with high probability.

*Strong consistency* was studied first, in a seminal paper [Bickel and Chen \[2009\]](#), as well as by [Mossel, Neeman, and Sly \[2014\]](#), [McSherry \[2001\]](#), [Hajek, Wu, and Xu \[2016\]](#), and [Cai and X. Li \[2015\]](#). Strong consistency is achievable, and achievable in polynomial time, if

$$\left| \sqrt{\frac{a}{\log n}} - \sqrt{\frac{b}{\log n}} \right| > \sqrt{2}$$

and not possible if  $\left| \sqrt{a/n} - \sqrt{b/n} \right| < \sqrt{2}$ . In particular, strong consistency is normally only considered in the semi-dense regime of  $d / \log n \rightarrow \infty$ .

*Weak consistency*, as one would expect, requires a stronger condition than weak recovery but a weaker one than strong consistency. Weak consistency is achievable if and only if

$$\frac{(a-b)^2}{a+b} = \omega(1)$$

see for example [Mossel, Neeman, and Sly \[2014\]](#). In particular, weak consistency is achievable in the semi-sparse regime of  $d \rightarrow \infty$ .

*Partial recovery*, finally, refers to the situation where the fraction of misclassified vertices does not go to 0, but remains bounded by a constant below 0.5. More specifically, partial recovery means that for a fixed  $\varepsilon > 0$  one can recover communities up to  $\varepsilon n$  mislabeled vertices. For the balanced symmetric case, this is true as long as

$$\frac{(a-b)^2}{a+b} = O(1)$$

which is primarily relevant when  $d = O(1)$ . Several types of algorithms are known to succeed at partial recovery in this very sparse regime, including non-backtracking walks [Mossel, Neeman, and Sly \[2013\]](#), [Massoulié \[2014\]](#), and [Bordenave, Lelarge, and Massoulié \[2015\]](#), spectral methods [Chin, Rao, and V. Vu \[2015\]](#) and methods based on semidefinite programming [Guédon and Vershynin \[2016\]](#) and [Montanari and Sen \[2016\]](#).

**6.3 Concentration implies recovery.** As an example application of the new concentration results, we demonstrate how to show that *regularized spectral clustering* [Amini, Chen, Bickel, and Levina \[2013\]](#) and [Joseph and Yu \[2016\]](#), one of the simplest and most popular

algorithms for community detection, can recover communities in the sparse regime of constant degrees. In general, spectral clustering works by computing the leading eigenvectors of either the adjacency matrix or the Laplacian, or their regularized versions, and running the  $k$ -means clustering algorithm on the rows of the  $n \times k$  matrix of leading eigenvectors to recover the node labels. In the simplest case of the balanced  $K = 2$  model  $G(n, \frac{a}{n}, \frac{b}{n})$ , one can simply assign nodes to two communities according to the sign of the entries of the eigenvector  $v_2(A')$  corresponding to the second smallest eigenvalue of the (regularized) adjacency matrix  $A'$ .

Let us briefly explain how concentration results validate recovery from the regularized adjacency matrix or regularized Laplacian. If concentration holds and the regularized matrix  $A'$  is shown to be close to  $\mathbb{E} A$ , then standard perturbation theory (i.e., the Davis-Kahan theorem, see e.g. [Bhatia \[1997\]](#)) implies that  $v_2(A')$  is close to  $v_2(\mathbb{E} A)$ , and in particular, the signs of these two eigenvectors must agree on most vertices. An easy calculation shows that the signs of  $v_2(\mathbb{E} A)$  recover the communities exactly: the eigenvector corresponding to the second smallest eigenvalue of  $\mathbb{E} A$  (or the second largest of  $\mathcal{L}(A)$ ) is a positive constant on one community and a negative constant on the other. Therefore, the signs of  $v_2(A')$  recover communities up to a small fraction of misclassified vertices and, as always, up to a permutation of community labels. This argument remains valid if we replace the regularized adjacency matrix  $A'$  with regularized Laplacian  $\mathcal{L}(A_\tau)$ .

**Corollary 6.1** (Partial recovery from a regularized adjacency matrix for sparse graphs). *Let  $\varepsilon > 0$  and  $r \geq 1$ . Let  $A$  be the adjacency matrix drawn from the stochastic block model  $G(n, \frac{a}{n}, \frac{b}{n})$ . Assume that*

$$(a - b)^2 > C(a + b)$$

*where  $C$  is a constant depending only on  $\varepsilon$  and  $r$ . For all nodes with degrees larger than  $2a$ , reduce the weights of the edges incident to them in an arbitrary way, but so that all degrees of the new (weighted) network become bounded by  $2a$ , resulting in a new matrix  $A'$ . Then with probability at least  $1 - e^{-r}$ , the signs of the entries of the eigenvector corresponding to the second smallest eigenvalue of  $A'$  correctly estimate the partition into two communities, up to at most  $\varepsilon n$  misclassified vertices.*

**Corollary 6.2** (Partial recovery from a regularized Laplacian for sparse graphs). *Let  $\varepsilon > 0$  and  $r \geq 1$ . Let  $A$  be the adjacency matrix drawn from the stochastic block model  $G(n, \frac{a}{n}, \frac{b}{n})$ . Assume that*

$$(6-1) \quad (a - b)^2 > C(a + b)$$

*where  $C$  is a constant depending only on  $\varepsilon$  and  $r$ . Choose  $\tau$  to be the average degree of the graph, i.e.  $\tau = (d_1 + \dots + d_n)/n$ . Then with probability at least  $1 - e^{-r}$ , the signs*

of the entries of the eigenvector corresponding to the second largest eigenvalue of  $\mathcal{L}(A_\tau)$  correctly estimate the partition into the two communities, up to at most  $\varepsilon n$  misclassified vertices.

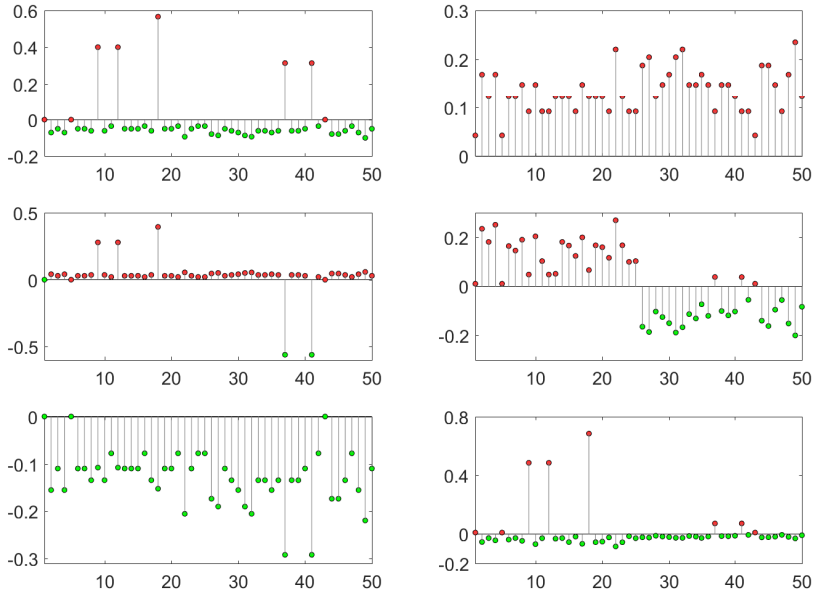


Figure 2: Three leading eigenvectors (from top to bottom) of the Laplacian (left) and the regularized Laplacian (right). The network is generated from  $G(n, \frac{a}{n}, \frac{b}{n})$  with  $n = 50$ ,  $a = 5$  and  $b = 0.1$ . Nodes are labeled so that the first 25 nodes belong to one community and the rest to the other community. Regularized Laplacian is computed from  $A + 0.1\bar{d}/n\mathbf{1}\mathbf{1}^\top$ .

As we have discussed, the Laplacian is typically preferred over the adjacency matrix in practice, because the variation in node degrees is reduced by the normalization factor  $D^{-1/2}$  [Sarkar and Bickel \[2015\]](#). [Figure 2](#) shows the effect of regularization for the Laplacian of a random network generated from  $G(n, \frac{a}{n}, \frac{b}{n})$  with  $n = 50$ ,  $a = 5$  and  $b = 0.1$ . For plotting purposes, we order the nodes so that the first  $n/2$  nodes belong to one community and the rest belong to the other community. Without regularization, the two leading eigenvectors of the Laplacian localize around a few low degree nodes, and therefore do not contain any information about the global community structure. In contrast, the second



leading eigenvector of the regularized Laplacian (with  $\tau = 0.1\bar{d}$ ) clearly reflects the communities, and the signs of this eigenvector alone recover community labels correctly for all but three nodes.

## 7 Discussion

Great progress has been made in recent years, and yet many problems remain open. Open questions on community detection under the SBM, in terms of exact and partial recovery and efficient (polynomial time) algorithms are discussed in Abbe [2017], and likely by the time this paper comes out in print, some of them will have been solved. Yet the focus on the SBM is unsatisfactory for many practitioners, since not many real networks fit this model well. Some of the more general models we discussed in Section 2 fix some of the problems of the SBM, allowing for heterogeneous degree distributions and overlapping communities, for instance. A bigger problem lies in the fixed  $K$  regime; it is not realistic to assume that as the size of the network grows, the number of communities remains fixed. A more realistic model is the “small world” scenario, where the size of communities remains bounded or grows very slowly with the number of nodes, the number of communities grows, and connections between many smaller communities happen primarily through hub nodes. Some consistency results have been obtained for a growing  $K$ , but we are not aware of any results in the sparse constant degree regime so far. An even bigger problem is presented by the so far nearly universal assumption of independent edges; this assumption violates commonly observed transitivity of friendships (if A is friends with B and B is friends with C, A is more likely to be friends with C). There are other types of network models that do not rely on this assumption, but hardly any random matrix results apply there. Ultimately, network analysis involves a lot more than community detection: link prediction, network denoising, predicting outcomes on networks, dynamic network modeling over time, and so on. We are a long way away from establishing rigorous theoretical guarantees for any of these problems to the extent that we have for community detection, but given how rapid progress in the latter area has been, we are hopeful that continued interest from the random matrix community will help shed light on other problems in network analysis.

## References

Emmanuel Abbe (Mar. 2017). “Community detection and stochastic block models: recent developments”. arXiv: 1703.10146 (cit. on pp. 2946, 2954, 2958).

- Emmanuel Abbe, Afonso S. Bandeira, and Georgina Hall (2016). “Exact recovery in the stochastic block model”. *IEEE Trans. Inform. Theory* 62.1, pp. 471–487. arXiv: [1405.3267](#). MR: [3447993](#) (cit. on p. [2954](#)).
- Edoardo M Airoldi, David M Blei, Stephen E Fienberg, and Eric P Xing (2008). “Mixed membership stochastic blockmodels”. *Journal of Machine Learning Research* 9.Sep, pp. 1981–2014 (cit. on p. [2946](#)).
- David J. Aldous (1981). “Representations for partially exchangeable arrays of random variables”. *J. Multivariate Anal.* 11.4, pp. 581–598. MR: [637937](#) (cit. on p. [2945](#)).
- Arash A. Amini, Aiyu Chen, Peter J. Bickel, and Elizaveta Levina (2013). “Pseudo-likelihood methods for community detection in large sparse networks”. *Ann. Statist.* 41.4, pp. 2097–2122. MR: [3127859](#) (cit. on pp. [2953](#), [2955](#)).
- Z. D. Bai and Y. Q. Yin (1988). “Necessary and sufficient conditions for almost sure convergence of the largest eigenvalue of a Wigner matrix”. *Ann. Probab.* 16.4, pp. 1729–1741. MR: [958213](#) (cit. on p. [2947](#)).
- B Ball, B. Karrer, and M. E. J. Newman (2011). “An efficient and principled method for detecting communities in networks”. *Physical Review E* 34, p. 036103 (cit. on p. [2946](#)).
- Afonso S. Bandeira and Ramon van Handel (2016). “Sharp nonasymptotic bounds on the norm of random matrices with independent entries”. *Ann. Probab.* 44.4, pp. 2479–2506. MR: [3531673](#) (cit. on pp. [2947](#), [2950](#)).
- Florent Benaych-Georges, Charles Bordenave, and Antti Knowles (Apr. 2017a). “Largest eigenvalues of sparse inhomogeneous Erdős–Rényi graphs”. arXiv: [1704.02953](#) (cit. on p. [2950](#)).
- (Apr. 2017b). “Spectral radii of sparse random matrices”. arXiv: [1704.02945](#) (cit. on p. [2948](#)).
- Rajendra Bhatia (1997). *Matrix analysis*. Vol. 169. Graduate Texts in Mathematics. Springer-Verlag, New York, pp. xii+347. MR: [1477662](#) (cit. on p. [2956](#)).
- Peter J Bickel and Aiyu Chen (2009). “A nonparametric view of network models and Newman–Girvan and other modularities”. *Proceedings of the National Academy of Sciences* 106.50, pp. 21068–21073 (cit. on p. [2955](#)).
- Béla Bollobás, Svante Janson, and Oliver Riordan (2007). “The phase transition in inhomogeneous random graphs”. *Random Structures Algorithms* 31.1, pp. 3–122. MR: [2337396](#) (cit. on p. [2946](#)).
- Charles Bordenave, Marc Lelarge, and Laurent Massoulié (2015). “Non-backtracking spectrum of random graphs: community detection and non-regular Ramanujan graphs”. In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS 2015*. IEEE Computer Soc., Los Alamitos, CA, pp. 1347–1357. arXiv: [1501.06087](#). MR: [3473374](#) (cit. on pp. [2954](#), [2955](#)).

- T. Tony Cai and Xiaodong Li (2015). “Robust and computationally feasible community detection in the presence of arbitrary outlier nodes”. *Ann. Statist.* 43.3, pp. 1027–1059. MR: [3346696](#) (cit. on p. [2955](#)).
- Kamalika Chaudhuri, Fan Chung, and Alexander Tsiatas (2012). “Spectral clustering of graphs with general degrees in the extended planted partition model”. In: *Proceedings of Machine Learning Research*, pp. 1–23 (cit. on pp. [2952](#), [2953](#)).
- P. Chin, A. Rao, and V. Vu (2015). “Stochastic block model and community detection in the sparse graphs : A spectral algorithm with optimal rate of recovery”. In: *Proceedings of Machine Learning Research*. Vol. 40, pp. 391–423 (cit. on pp. [2948](#), [2950](#), [2955](#)).
- Fan R. K. Chung (1997). *Spectral graph theory*. Vol. 92. CBMS Regional Conference Series in Mathematics. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI, pp. xii+207. MR: [1421568](#) (cit. on p. [2952](#)).
- Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová (2011). “Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications”. *Physical Review E* 84.6, p. 066106 (cit. on p. [2954](#)).
- P. Erdős and A. Rényi (1959). “On random graphs. I”. *Publ. Math. Debrecen* 6, pp. 290–297. MR: [0120167](#) (cit. on p. [2944](#)).
- Uriel Feige and Eran Ofek (2005). “Spectral techniques applied to sparse random graphs”. *Random Structures Algorithms* 27.2, pp. 251–275. MR: [2155709](#) (cit. on pp. [2948](#)–[2951](#)).
- Santo Fortunato (2010). “Community detection in graphs”. *Phys. Rep.* 486.3-5, pp. 75–174. MR: [2580414](#) (cit. on p. [2954](#)).
- Joel Friedman, Jeff Kahn, and Endre Szemerédi (1989). “On the second eigenvalue of random regular graphs”. In: *Proceedings of the twenty-first annual ACM symposium on Theory of computing*. ACM, pp. 587–598 (cit. on pp. [2948](#), [2950](#)).
- Z. Füredi and J. Komlós (1981). “The eigenvalues of random symmetric matrices”. *Combinatorica* 1.3, pp. 233–241. MR: [637828](#) (cit. on p. [2947](#)).
- Chao Gao, Yu Lu, and Harrison H. Zhou (2015). “Rate-optimal graphon estimation”. *Ann. Statist.* 43.6, pp. 2624–2652. MR: [3405606](#) (cit. on p. [2945](#)).
- Chao Gao, Zongming Ma, Anderson Y. Zhang, and Harrison H. Zhou (2017). “Achieving optimal misclassification proportion in stochastic block models”. *J. Mach. Learn. Res.* 18, Paper No. 60, 45. MR: [3687603](#) (cit. on pp. [2952](#), [2953](#)).
- Anna Goldenberg, Alice X Zheng, Stephen E Fienberg, Edoardo M Airolidi, et al. (2010). “A survey of statistical network models”. *Foundations and Trends in Machine Learning* 2.2, pp. 129–233 (cit. on p. [2954](#)).
- Olivier Guédon and Roman Vershynin (2016). “Community detection in sparse networks via Grothendieck’s inequality”. *Probab. Theory Related Fields* 165.3-4, pp. 1025–1049. MR: [3520025](#) (cit. on p. [2955](#)).

- Bruce Hajek, Yihong Wu, and Jiaming Xu (2016). “Achieving exact cluster recovery threshold via semidefinite programming”. *IEEE Trans. Inform. Theory* 62.5, pp. 2788–2797. MR: [3493879](#) (cit. on pp. [2948](#), [2949](#), [2955](#)).
- Ramon van Handel (2017a). “On the spectral norm of Gaussian random matrices”. *Trans. Amer. Math. Soc.* 369.11, pp. 8161–8178. MR: [3695857](#) (cit. on p. [2950](#)).
- (2017b). “Structured random matrices”. *Convexity and Concentration* 161, pp. 107–156 (cit. on p. [2950](#)).
- Peter D. Hoff, Adrian E. Raftery, and Mark S. Handcock (2002). “Latent space approaches to social network analysis”. *J. Amer. Statist. Assoc.* 97.460, pp. 1090–1098. MR: [1951262](#) (cit. on p. [2945](#)).
- Paul W. Holland, Kathryn Blackmond Laskey, Samuel Leinhardt, and and (1983). “Stochastic blockmodels: first steps”. *Social Networks* 5.2, pp. 109–137. MR: [718088](#) (cit. on p. [2944](#)).
- Douglas N Hoover (1979). “Relations on probability spaces and arrays of random variables”. *Technical report, Institute for Advanced Study, Princeton, NJ* 2 (cit. on p. [2945](#)).
- Antony Joseph and Bin Yu (2016). “Impact of regularization on spectral clustering”. *Ann. Statist.* 44.4, pp. 1765–1791. MR: [3519940](#) (cit. on pp. [2952](#), [2953](#), [2955](#)).
- Brian Karrer and M. E. J. Newman (2011). “Stochastic blockmodels and community structure in networks”. *Phys. Rev. E* (3) 83.1, pp. 016107, 10. MR: [2788206](#) (cit. on p. [2945](#)).
- Michael Krivelevich and Benny Sudakov (2003). “The largest eigenvalue of sparse random graphs”. *Combin. Probab. Comput.* 12.1, pp. 61–72. MR: [1967486](#) (cit. on p. [2950](#)).
- Can M. Le, Elizaveta Levina, and Roman Vershynin (2017). “Concentration and regularization of random graphs”. *Random Structures Algorithms* 51.3, pp. 538–561. MR: [3689343](#) (cit. on pp. [2950](#)–[2952](#)).
- Michel Ledoux and Michel Talagrand (1991). *Probability in Banach spaces*. Vol. 23. Ergebnisse der Mathematik und ihrer Grenzgebiete (3) [Results in Mathematics and Related Areas (3)]. Isoperimetry and processes. Springer-Verlag, Berlin, pp. xii+480. MR: [1102015](#) (cit. on p. [2951](#)).
- Jing Lei and Alessandro Rinaldo (2015). “Consistency of spectral clustering in stochastic block models”. *Ann. Statist.* 43.1, pp. 215–237. MR: [3285605](#) (cit. on pp. [2948](#), [2950](#)).
- Tianxi Li, Elizaveta Levina, and Ji Zhu (Dec. 2016). “Network cross-validation by edge sampling”. arXiv: [1612.04717](#) (cit. on p. [2953](#)).
- Laurent Massoulié (2014). “Community detection thresholds and the weak Ramanujan property”. In: *STOC’14—Proceedings of the 2014 ACM Symposium on Theory of Computing*. ACM, New York, pp. 694–703. MR: [3238997](#) (cit. on pp. [2954](#), [2955](#)).
- Frank McSherry (2001). “Spectral partitioning of random graphs”. In: *42nd IEEE Symposium on Foundations of Computer Science (Las Vegas, NV, 2001)*. IEEE Computer Soc., Los Alamitos, CA, pp. 529–537. MR: [1948742](#) (cit. on p. [2955](#)).

- Andrea Montanari and Subhabrata Sen (2016). “Semidefinite programs on sparse random graphs and their application to community detection”. In: *STOC’16—Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, pp. 814–827. MR: [3536616](#) (cit. on p. [2955](#)).
- Elchanan Mossel, Joe Neeman, and Allan Sly (Nov. 2013). “A Proof Of The Block Model Threshold Conjecture”. arXiv: [1311.4115](#) (cit. on pp. [2954](#), [2955](#)).
- (July 2014). “Consistency Thresholds for the Planted Bisection Model”. arXiv: [1407.1591](#) (cit. on pp. [2954](#), [2955](#)).
- (2015). “Reconstruction and estimation in the planted partition model”. *Probab. Theory Related Fields* 162.3–4, pp. 431–461. MR: [3383334](#) (cit. on p. [2954](#)).
- Sofia C Olhede and Patrick J Wolfe (2013). “Network histograms and universality of blockmodel approximation”. *Proceedings of the National Academy of Sciences* 111.41, pp. 14722–14727 (cit. on p. [2945](#)).
- Albrecht Pietsch (1980). *Operator ideals*. Vol. 20. North-Holland Mathematical Library. Translated from German by the author. North-Holland Publishing Co., Amsterdam–New York, p. 451. MR: [582655](#) (cit. on p. [2951](#)).
- Gilles Pisier (1986). *Factorization of linear operators and geometry of Banach spaces*. Vol. 60. CBMS Regional Conference Series in Mathematics. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI, pp. x+154. MR: [829919](#) (cit. on p. [2951](#)).
- (2012). “Grothendieck’s theorem, past and present”. *Bull. Amer. Math. Soc. (N.S.)* 49.2, pp. 237–323. MR: [2888168](#) (cit. on p. [2951](#)).
- Tai Qin and Karl Rohe (2013). “Regularized spectral clustering under the degree-corrected stochastic blockmodel”. In: *Advances in Neural Information Processing Systems*, pp. 3120–3128 (cit. on p. [2952](#)).
- Stiene Riemer and Carsten Schütt (2013). “On the expectation of the norm of random matrices with non-identically distributed entries”. *Electron. J. Probab.* 18, no. 29, 13. MR: [3035757](#) (cit. on p. [2950](#)).
- Purnamrita Sarkar and Peter J. Bickel (2015). “Role of normalization in spectral clustering for stochastic blockmodels”. *Ann. Statist.* 43.3, pp. 962–990. MR: [3346694](#) (cit. on p. [2957](#)).
- Yoav Seginer (2000). “The expected norm of random matrices”. *Combin. Probab. Comput.* 9.2, pp. 149–166. MR: [1762786](#) (cit. on pp. [2947](#), [2948](#), [2950](#)).
- Nicole Tomczak-Jaegermann (1989). *Banach-Mazur distances and finite-dimensional operator ideals*. Vol. 38. Pitman Monographs and Surveys in Pure and Applied Mathematics. Longman Scientific & Technical, Harlow; copublished in the United States with John Wiley & Sons, Inc., New York, pp. xii+395. MR: [993774](#) (cit. on p. [2951](#)).
- Van H. Vu (2007). “Spectral norm of random matrices”. *Combinatorica* 27.6, pp. 721–736. MR: [2384414](#) (cit. on p. [2948](#)).

- Shuhua Yin (2008). “Investigation on spectrum of the adjacency matrix and Laplacian matrix of graph  $\mathcal{G}_l$ ”. *WSEAS Trans. Syst.* 7.4, pp. 362–372. MR: [2447295](#) (cit. on p. [2952](#)).
- Yuan Zhang, Elizaveta Levina, and Ji Zhu (Dec. 2014). “[Detecting Overlapping Communities in Networks Using Spectral Methods](#)”. arXiv: [1412.3432](#) (cit. on p. [2946](#)).
- (2017). “[Estimating network edge probabilities by neighbourhood smoothing](#)”. *Biometrika* 104.4, pp. 771–783. MR: [3737303](#) (cit. on p. [2945](#)).
- Y. Zhao, E. Levina, and J. Zhu (2012). “Consistency of community detection in networks under degree-corrected stochastic block models”. *Annals of Statistics* 40.4, pp. 2266–2292 (cit. on p. [2945](#)).

Received 2017-12-22.

CAN M. LE

DEPARTMENT OF STATISTICS, UNIVERSITY OF CALIFORNIA, DAVIS, ONE SHIELDS AVE, DAVIS, CA 95616, U.S.A.  
[canle@ucdavis.edu](mailto:canle@ucdavis.edu)

ELIZAVETA LEVINA

DEPARTMENT OF STATISTICS, UNIVERSITY OF MICHIGAN, 1085 S. UNIVERSITY AVE, ANN ARBOR, MI 48109, U.S.A.  
[elevina@umich.edu](mailto:elevina@umich.edu)

ROMAN VERSHYNIN

UNIVERSITY OF CALIFORNIA IRVINE, 340 ROWLAND HALL, IRVINE, CA 92697, U.S.A.  
[rvershyn@uci.edu](mailto:rvershyn@uci.edu)



# LIOUVILLE QUANTUM GRAVITY AS A METRIC SPACE AND A SCALING LIMIT

JASON MILLER

## Abstract

Over the past few decades, two natural random surface models have emerged within physics and mathematics. The first is Liouville quantum gravity, which has its roots in string theory and conformal field theory from the 1980s and 1990s. The second is the Brownian map, which has its roots in planar map combinatorics from the 1960s together with recent scaling limit results. This article surveys a series of works with Sheffield in which it is shown that Liouville quantum gravity (LQG) with parameter  $\gamma = \sqrt{8/3}$  is equivalent to the Brownian map. We also briefly describe a series of works with Gwynne which use the  $\sqrt{8/3}$ -LQG metric to prove the convergence of self-avoiding walks and percolation on random planar maps towards  $\text{SLE}_{8/3}$  and  $\text{SLE}_6$ , respectively, on a Brownian surface.

## 1 Introduction

**1.1 The Gaussian free field.** Suppose that  $D \subseteq \mathbb{C}$  is a planar domain. Informally, the Gaussian free field (GFF)  $h$  on  $D$  is a Gaussian random variable with covariance function

$$\text{cov}(h(x), h(y)) = G(x, y) \quad \text{for } x, y \in D$$

where  $G$  denotes the Green's function for  $\Delta$  on  $D$ . Since  $G(x, y) \sim -\log|x - y|$  as  $x \rightarrow y$ , it turns out that it is not possible to make sense of the GFF as a function on  $D$  but rather it exists as a distribution on  $D$ . Perhaps the most natural construction of the GFF is using a series expansion. More precisely, one defines  $H_0^1(D)$  to be the Hilbert space closure of  $C_0^\infty(D)$  with respect to the *Dirichlet inner product*

$$(1-1) \quad (f, g)_\nabla = \frac{1}{2\pi} \int_D \nabla f(x) \cdot \nabla g(x) dx.$$



One then sets

$$(1-2) \quad h = \sum_n \alpha_n \phi_n$$

where  $(\alpha_n)$  is a sequence of i.i.d.  $N(0, 1)$  random variables and  $(\phi_n)$  is an orthonormal basis of  $H_0^1(D)$ . The convergence of the series (1-2) does not take place in  $H_0^1(D)$ , but rather in the space of distributions on  $D$ . (One similarly defines the GFF with free boundary conditions on a given boundary segment  $L$  using the same construction but including also the those functions can be non-zero on  $L$ .) Since the Dirichlet inner product is conformally invariant, so is the law of the GFF.

The GFF is a fundamental object in probability theory. Just like the Brownian motion arises as the scaling limit of many different types of random curves, the GFF describes the scaling limit of many different types of random surface models [Kenyon \[2000\]](#), [Ben Arous and Deuschel \[1996\]](#), [Giacomin, Olla, and Spohn \[2001\]](#), [Rider and Virág \[2007\]](#), and [Miller \[2011\]](#). It also has deep connections with many other important objects in probability theory, such as random walks and Brownian motion [Dynkin \[1984b,a\]](#) and [Le Jan \[2011\]](#) and the Schramm-Loewner evolution (SLE) [Schramm \[2000\]](#), [Schramm and Sheffield \[2009, 2013\]](#), [Sheffield \[2016a\]](#), [Dubédat \[2009\]](#), [Miller and Sheffield \[2016a,b,c, 2017\]](#), and [Duplantier, Miller, and Sheffield \[2014\]](#).

**1.2 Liouville quantum gravity.** Liouville quantum gravity (LQG) is one of several important random geometries that one can build from the GFF. It was introduced (non-rigorously) in the physics literature by Polyakov in the 1980s as a model for a string [A. M. Polyakov \[1981a,b\]](#) and [A. Polyakov \[1990\]](#). In its simplest form, it refers to the random Riemannian manifold with metric tensor given by

$$(1-3) \quad e^{\gamma h(z)}(dx^2 + dy^2)$$

where  $h$  is (some variant of) the GFF on  $D$ ,  $\gamma$  is a real parameter,  $z = x + iy \in D$ , and  $dx^2 + dy^2$  denotes the Euclidean metric on  $D$ . This expression does not make literal sense because  $h$  is only a distribution on  $D$  and not a function, hence does not take values at points.

There has been a considerable effort within the probability community over the course of the last decade or so to make rigorous sense of LQG. This is motivated by the goal of putting a heuristic from the physics literature, the so-called KPZ formula [Knizhnik, A. M. Polyakov, and Zamolodchikov \[1988\]](#) which has been used extensively to (non-rigorously) derive critical exponents for two-dimensional lattice models, onto firm mathematical ground. Another motivation comes from deep conjectures, which we will shortly describe in more detail, which state that LQG should describe the large-scale behavior of random planar maps.

The volume form

$$(1-4) \quad e^{\gamma h(z)} dx dy$$

associated with the metric (1-3) was constructed in Duplantier and Sheffield [2011] by regularizing  $h$  by considering its average  $h_\epsilon(z)$  on  $\partial B(z, \epsilon)$  and then setting

$$(1-5) \quad \mu_h^\gamma = \lim_{\epsilon \rightarrow 0} \epsilon^{\gamma^2/2} e^{\gamma h_\epsilon(z)} dx dy$$

where  $dx dy$  denotes Lebesgue measure on  $D$ . We note that there is no difficulty in making sense of the expression inside of the limit on the right hand side of (1-5) for each fixed  $\epsilon > 0$  since  $(z, \epsilon) \mapsto h_\epsilon(z)$  is a continuous function Duplantier and Sheffield [ibid.]. The factor  $\epsilon^{\gamma^2/2}$  appears because the leading order term in the variance of  $h_\epsilon(z)$  is  $\log \epsilon^{-1}$ . In the case that  $h$  is a GFF on a domain  $D$  with a linear boundary segment  $L$  and free boundary conditions along  $L$ , one can similarly construct a boundary length measure by setting

$$(1-6) \quad \nu_h^\gamma = \lim_{\epsilon \rightarrow 0} \epsilon^{\gamma^2/4} e^{\gamma h_\epsilon(x)/2} dx$$

where  $dx$  denotes Lebesgue measure on  $L$ . There is in fact a general theory of random measures with the same law as  $\mu_h, \nu_h$  which is referred to as *Gaussian multiplicative chaos* and was developed by Kahane [1985] in the 1980s; see Rhodes and Vargas [2014] for a more recent review. (Similar measures also appeared earlier in Höegh-Krohn [1971].)

The regularization procedure used to define the area and boundary measures leads to a natural change of coordinates formula for LQG. Namely, if  $h$  is a GFF on a domain  $D$ ,  $\varphi: \widetilde{D} \rightarrow D$  is a conformal transformation, and one takes

$$(1-7) \quad \widetilde{h} = h \circ \varphi + Q \log |\varphi'| \quad \text{where} \quad Q = \frac{2}{\gamma} + \frac{\gamma}{2}$$

then the area and boundary measures defined by  $\widetilde{h}$  are the same as the pushforward of the area and boundary measures defined by  $h$ . Therefore  $(D, h)$  and  $(\widetilde{D}, \widetilde{h})$  can be thought of as parameterizations of the same surface. Whenever two pairs  $(D, h)$  and  $(\widetilde{D}, \widetilde{h})$  are related as in (1-7), we say that they are equivalent as quantum surfaces and a *quantum surface* is an equivalence class under this relation. A particular choice of representative is referred to as an *embedding*. There are many different choices of embeddings of a given quantum surface which can be natural depending on the context, and this is a point we will come back to later. We note that when  $\gamma \rightarrow 0$  so that an LQG surface corresponds to a flat, Euclidean surface (i.e., the underlying planar domain), the change of coordinate formula (1-7) exactly corresponds to the usual change of coordinates formula.

**1.3 Random planar maps.** A *planar map* is a graph together with an embedding into the plane so that no two edges cross. Planar maps  $M_1, M_2$  are said to be equivalent if there exists an orientation preserving homeomorphism  $\varphi$  of  $\mathbb{R}^2$  which takes  $M_1$  to  $M_2$ . The *faces* of a planar map are the connected components of the complement of its edges. A map is called a triangulation (resp. quadrangulation) if it has the property that all faces have exactly three (resp. four) adjacent edges. The study of planar maps goes back to work of Tutte [1962], who worked on enumerating planar maps in the context of the four color theorem, and of Mullin [1967], who worked on enumerating planar maps decorated with a distinguished spanning tree. Random maps were intensively studied in physics by random matrix techniques, starting from Brézin, Itzykson, Parisi, and Zuber [1978] (see, e.g., the review Di Francesco, Ginsparg, and Zinn-Justin [1995]). This field has since been revitalized by the development of bijective techniques due to Cori and Vauquelin [1981] and Schaeffer [1998] and has remained a very active area in combinatorics and probability, especially in the last 20 or so years.

Since there are only a finite number of planar maps with  $n$  faces, one can pick one uniformly at random. We will now mention some of the main recent developments in the study of the metric properties of large uniformly random maps. First, it was shown by Chassaing and Schaeffer that the diameter of a uniformly random quadrangulation with  $n$  faces is typically of order  $n^{1/4}$  Chassaing and Schaeffer [2004]. It was then shown by Le Gall that if one rescales distances by the factor  $n^{-1/4}$  Le Gall [2007] then one obtains a tight sequence in the space of metric spaces (equipped with the Gromov-Hausdorff topology), which means that there exist subsequential limits in law. Le Gall also showed that the Hausdorff dimension of any subsequential limit is a.s. equal to 4 Le Gall [ibid.] and it was shown by Le Gall and Paulin [2008] (see also Miermont [2008]) that every subsequential limit is a.s. homeomorphic to the two-dimensional sphere  $\mathbb{S}^2$ . This line of work culminated with independent works of Le Gall [2013] and Miermont [2013], which both show that one has a true limit in distribution. The limiting random metric measure space is known as the *Brownian map*. (The term Brownian map first appeared in Marckert and Mokkadem [2006].) We will review the continuum construction of the Brownian map in Section 5.1. See Le Gall [2014] for a more in depth survey.

Building off the works Le Gall [2013] and Miermont [2013], this convergence has now been extended to a number of other topologies. In particular:

- Curien and Le Gall proved that the uniform quadrangulation of the whole plane (UIPQ), the local limit of a uniform quadrangulation with  $n$  faces, converges to the Brownian plane Curien and Le Gall [2014].

- Bettinelli and Miermont proved that quadrangulations of the disk with general boundary converge to the Brownian disk [Bettinelli and Miermont \[2017\]](#) (see also the extension [Gwynne and Miller \[2017c\]](#) to the case of quadrangulations of the disk with simple boundary).
- Building on [Bettinelli and Miermont \[2017\]](#), it was shown in [Gwynne and Miller \[2017d\]](#) and [Baur, Miermont, and Ray \[2016\]](#) that the uniform quadrangulation of the upper half-plane (UIHPQ), the local limit of a uniform quadrangulation of the disk near a boundary typical point, converges to the Brownian half-plane.

It has long been believed that LQG should describe the large scale behavior of random planar maps (see [Ambjørn, Durhuus, and Jonsson \[1997\]](#)), with the case of uniformly random planar maps corresponding to  $\gamma = \sqrt{8/3}$ . Precise conjectures have also been made more recently in the mathematics literature (see e.g. [Duplantier and Sheffield \[2011\]](#) and [David, Kupiainen, Rhodes, and Vargas \[2016\]](#)).

One approach is to view a quadrangulation as a surface by identifying each of the quadrilaterals with a copy of the Euclidean square  $[0, 1]^2$  which are identified according to boundary length using the adjacency structure of the map. One can then conformally map the resulting surface to  $\mathbb{S}^2$  and write the resulting metric in coordinates with respect to the Euclidean metric

$$(1-8) \quad e^{\lambda_n(z)}(dx^2 + dy^2).$$

The goal is then to show that  $\lambda_n$  converges in the limit as  $n \rightarrow \infty$  to  $\sqrt{8/3}$  times a form of the GFF. A variant of this conjecture is that the volume form associated with (1-8) converges in the limit to the  $\sqrt{8/3}$ -LQG measure associated with a form of the GFF. One can also consider other types of “discrete” conformal embeddings, such as circle packings, square tilings, or the Tutte (a.k.a. harmonic, or barycentric) embedding. (See [Gwynne, Miller, and Sheffield \[2017\]](#) for a convergence result of this type in the case of the so-called *mated-CRT map*.)

In this article, we will focus on the solution of a version of this conjecture carried out in [Miller and Sheffield \[2015b, 2016d,e, 2015a,c\]](#) in which it is shown that a  $\sqrt{8/3}$ -LQG surface determines a metric space structure and if one considers the correct law on  $\sqrt{8/3}$ -LQG surfaces then this metric space is an instance of the Brownian map.

**Acknowledgments.** We thank Bertrand Duplantier, Ewain Gwynne, and Scott Sheffield for helpful comments on an earlier version of this article.

## 2 Liouville quantum gravity surfaces

In order to make the connections between random planar maps and LQG precise, one needs to make precise the correct law on GFF-like distributions  $h$ . Since the definitions of these surfaces are quite important, we will now spend some time describing how to derive these laws (in the simply connected case), first in the setting of surfaces with boundary and then in the setting of surfaces without boundary. These constructions were first described in [Sheffield \[2016a\]](#) and carried out carefully in [Duplantier, Miller, and Sheffield \[2014\]](#).

**2.1 Surfaces with boundary.** The starting point for deriving the correct form of the distribution  $h$  is to understand the behavior of such a surface near a *boundary typical point*, that is, near a point  $x$  in the boundary chosen from  $\nu_h$ , for a general LQG surface with boundary. To make this more concrete, we consider the domain  $D = \mathbb{D} \cap \mathbb{H}$ , i.e., the upper semi-disk in  $\mathbb{H}$ . Let  $h$  be a GFF on  $D$  with free (resp. Dirichlet) boundary conditions on  $[-1, 1]$  (resp.  $\partial D \setminus [-1, 1]$ ). Following [Duplantier and Sheffield \[2011\]](#), we then consider the law whose Radon-Nikodym derivative with respect to  $h$  is given by a normalizing constant times  $\nu_h([-1, 1])$ . That is, if  $dh$  denotes the law of  $h$ , then the law we are considering is given by  $\nu_h([-1, 1])dh$  normalized to be a probability measure. From (1-6) this, in turn, is the same as the limit as  $\epsilon \rightarrow 0$  of the marginal of  $h$  under the law

$$(2-1) \quad \Theta_\epsilon(dx, dh) = \epsilon^{\gamma^2/4} e^{\gamma h_\epsilon(x)/2} dx dh$$

where  $dx$  denotes Lebesgue measure on  $[-1, 1]$ . We are going to in fact consider the limit  $\Theta$  as  $\epsilon \rightarrow 0$  of  $\Theta_\epsilon$  from (2-1). The reason for this is that in the limit as  $\epsilon \rightarrow 0$ , the conditional law of  $x$  given  $h$  converges to a point chosen from  $\nu_h$ . By performing an integration by parts, we can write  $h_\epsilon(x) = (h, \xi_\epsilon^x)_\nabla$  where  $\xi_\epsilon^x(y) = -\log \max(|x - y|, \epsilon) - \tilde{G}_\epsilon^x(z)$  and  $\tilde{G}_\epsilon^x$  is the function which is harmonic on  $D$  with Dirichlet boundary conditions given by  $y \mapsto -\log \max(|x - y|, \epsilon)$  on  $\partial D \setminus [-1, 1]$  and Neumann boundary conditions on  $[-1, 1]$ . In other words,  $\xi_\epsilon^x$  is a truncated form of the Green's function  $G$  for  $\Delta$  on  $D$  with Dirichlet boundary conditions on  $\partial D \setminus [-1, 1]$  and Neumann boundary conditions on  $[-1, 1]$ . Recall the following basic fact about the Gaussian distribution: if  $Z \sim N(0, 1)$  and we weight the law of  $Z$  by a normalizing constant times  $e^{\mu Z}$ , then the resulting distribution is that of a  $N(\mu, 1)$  random variable. By the infinite dimensional analog of this, we therefore have in the case of  $h$  that weighting its law by a constant times  $\exp(\gamma h_\epsilon(x)/2) = \exp(\gamma(h, \xi_\epsilon^x)_\nabla/2)$  is the same as shifting its mean by  $(\gamma/2)\xi_\epsilon^x$ . That is, under  $\Theta_\epsilon$ , we have that the conditional law of  $h$  given  $x$  is that of  $\tilde{h} + (\gamma/2)\xi_\epsilon^x$  where  $\tilde{h}$  is a GFF on  $D$  with free (resp. Dirichlet) boundary conditions on  $[-1, 1]$  (resp.  $\partial D \setminus [-1, 1]$ ). Taking a limit as  $\epsilon \rightarrow 0$ , we thus see that the conditional law of  $h$  given  $x$  under  $\Theta$  is given

by that of  $\tilde{h} + (\gamma/2)G(x, \cdot)$  where  $\tilde{h}$  is a GFF on  $D$  with free (resp. Dirichlet) boundary conditions on  $[-1, 1]$  (resp.  $\partial D \setminus [-1, 1]$ ).

This computation tells us that the local behavior of  $h$  near a point chosen from  $\nu_h$  is described by that of a GFF with free boundary conditions plus the singularity  $-\gamma \log |\cdot|$  (as the leading order behavior of  $G(x, y)$  is  $-2 \log |x - y|$  for  $x \in (-1, 1)$  and  $y$  close to  $x$ ). We now describe how to take an infinite volume limit near  $x$  in the aforementioned construction. Roughly speaking, we will “zoom in” by adding a large constant  $C$  to  $h$  (which has the effect of replacing  $\mu_h$  with  $e^{\gamma C} \mu_h$ ), centering so that  $x$  is at the origin, and then performing a rescaling so that  $\mathbb{D} \cap \mathbb{H}$  is assigned one unit of mass.

It is easiest to describe this procedure if we first apply a change of coordinates from  $D$  to the infinite half-strip  $\mathcal{S}_+ = \mathbb{R}_+ \times (0, \pi)$  using the unique conformal map  $\varphi: D \rightarrow \mathcal{S}_+$  which takes  $-1$  to  $0$ ,  $x$  to  $+\infty$ , and  $1$  to  $\pi i$ . Then the law of  $\tilde{h} = h \circ \varphi^{-1} + Q \log |(\varphi^{-1})'|$  is that of  $\hat{h} + (\gamma - Q)\text{Re}(\cdot)$  where  $\hat{h}$  is a GFF on  $\mathcal{S}_+$  with free (resp. Dirichlet) boundary conditions on  $\partial \mathcal{S}_+ \setminus [0, \pi i]$  (resp.  $[0, \pi i]$ ). For each  $u \geq 0$ , let  $\tilde{A}_u$  be the average of  $\tilde{h}$  on the vertical line  $[0, \pi i] + u$ . For such a GFF, it is possible to check that  $\tilde{A}_u = \tilde{B}_{2u} + (\gamma - Q)u$  where  $\tilde{B}$  is a standard Brownian motion. Suppose that  $C > 0$  is a large constant and  $\tilde{h}_C = \tilde{h}(\cdot + \tau_C) + C$  where  $\tau_C = \inf\{u \geq 0 : \tilde{A}_u + C = 0\}$ . Note that  $\tau_C$  is a.s. finite since  $(\gamma - Q) < 0$ . As  $C \rightarrow \infty$ , the law of  $\tilde{h}_C$  converges to that of a field on  $\mathcal{S}$  whose law can be sampled from using the following two step procedure:

- Take its average on vertical lines  $[0, \pi i] + u$  to be given by  $A_u$  where  $A_u$  for  $u > 0$  is given by  $B_{2u} + (\gamma - Q)u$  where  $B$  is a standard Brownian motion and for  $u < 0$  by  $\hat{B}_{-2u} + (\gamma - Q)u$  where  $\hat{B}$  is an independent standard Brownian motion conditioned so that  $\hat{B}_{2s} + (Q - \gamma)s \geq 0$  for all  $s \geq 0$ .
- Take its projection onto the  $(\cdot, \cdot)_{\nabla}$ -orthogonal complement of the subspace of functions which are constant on vertical lines of the form  $[0, \pi i] + u$  to be given by the corresponding projection of a GFF on  $\mathcal{S}$  with free boundary conditions on  $\mathcal{S}$  and which is independent of  $A$ .

The surface whose construction we have just described is called a  $\gamma$ -quantum wedge and, when parameterized by  $\mathbb{H}$ , can be thought of as a version of the GFF on  $\mathbb{H}$  with free boundary conditions and a  $-\gamma \log |\cdot|$  singularity with the additive constant fixed in a canonical manner. The construction generalizes to that of an  $\alpha$ -quantum wedge which is a similar type of quantum surface except with a  $-\alpha \log |\cdot|$  singularity. Quantum wedges are naturally marked by two points. When parameterized by  $\mathbb{H}$ , these correspond to  $0$  and  $\infty$  and when parameterized by  $\mathcal{S}$  correspond to  $-\infty$  and  $+\infty$ . This is emphasized with the notation  $(\mathbb{H}, h, 0, \infty)$  or  $(\mathcal{S}, h, -\infty, +\infty)$ .

A *quantum disk* is the finite volume analog of a  $\gamma$ -quantum wedge and, when parameterized by  $\mathcal{S}$ , the law of the associated field can be described in a manner which is analogous to that of a  $\gamma$ -quantum wedge. To make this more concrete, we recall that if  $B$  is a standard Brownian motion and  $a \in \mathbb{R}$ , then  $e^{B_t+at}$  reparameterized to have quadratic variation  $dt$  is a Bessel process of dimension  $\delta = 2 + 2a$ . Conversely, if  $Z$  is a Bessel process of dimension  $\delta$ , then  $\log Z$  reparameterized to have quadratic variation  $dt$  is a standard Brownian motion with drift  $at = (\delta - 2)t/2$ . The law of the process  $A$  described just above can therefore be sampled from by first sampling a Bessel process  $Z$  of dimension  $\delta = 8/\gamma^2$ , then reparameterizing  $(4/\gamma) \log Z$  to have quadratic variation  $2dt$ , and then reversing and centering time so that it first hits 0 at  $u = 0$ . The law of a quantum disk can be sampled from in the same manner except one replaces the Bessel process  $Z$  just above with a Bessel excursion sampled from the excursion measure of a Bessel process of dimension  $4 - 8/\gamma^2$ . A quantum disk parameterized by  $\mathcal{S}$  is marked by two points which correspond to  $-\infty, +\infty$  and is emphasized with the notation  $(\mathcal{S}, h, -\infty, +\infty)$ . It turns out that these two points have the law of independent samples from the boundary measure when one conditions on the quantum surface structure of a quantum disk.

**2.2 Surfaces without boundary.** The derivation of the surfaces without boundary proceeds along the same lines as the derivation of the surfaces with boundary except one analyzes the behavior of an LQG surface near an area typical point rather than a boundary typical point. The infinite volume surface is the  $\gamma$ -quantum cone and the finite volume surface is the *quantum sphere*. In this case, it is natural to parameterize such a surface by the infinite cylinder  $\mathcal{C} = \mathbb{R} \times [0, 2\pi]$  (with the top and bottom identified). The law of a  $\gamma$ -quantum cone parameterized by  $\mathcal{C}$  can be sampled from by:

- Taking its average on vertical lines  $[0, 2\pi i] + u$  to be given by  $A_u$  where  $A_u$  for  $u > 0$  is given by  $B_u + (\gamma - Q)u$  where  $B$  is a standard Brownian motion and for  $u < 0$  by  $\hat{B}_{-u} + (\gamma - Q)u$  where  $\hat{B}$  is an independent standard Brownian motion conditioned so that  $\hat{B}_s + (Q - \gamma)s \geq 0$  for all  $s \geq 0$ .
- Take its projection onto the  $(\cdot, \cdot)_{\nabla}$ -orthogonal complement of the subspace of functions which are constant on vertical lines of the form  $[0, 2\pi i] + u$  to be given by the corresponding projection of a GFF on  $\mathcal{C}$  and which is independent of  $A$ .

As in the case of a  $\gamma$ -quantum wedge, it is natural to describe  $A$  in terms of a Bessel process. In this case, one can sample from its law by first sampling a Bessel process  $Z$  of dimension  $\delta = 8/\gamma^2$ , then reparameterizing  $(2/\gamma) \log Z$  to have quadratic variation  $dt$ , and then reversing and centering time so that it first hits 0 at  $u = 0$ . A quantum sphere can be constructed in an analogous manner except one replaces the Bessel process  $Z$  just

above with a Bessel excursion sampled from the excursion measure of a Bessel process of dimension  $4 - 8/\gamma^2$ .

The  $\gamma$ -quantum cone parameterized by  $\mathbb{C}$  can be viewed as a whole-plane GFF plus  $-\gamma \log |\cdot|$  with the additive constant fixed in a canonical way. The  $\alpha$ -quantum cone is a generalization of this where the  $-\gamma \log |\cdot|$  singularity is replaced with a  $-\alpha \log |\cdot|$  singularity.

Quantum cones are naturally marked by two points. When parameterized by  $\mathbb{C}$ , these correspond to 0 and  $\infty$  and when parameterized by  $\mathcal{C}$  correspond to  $-\infty$  and  $+\infty$ . This is emphasized with the notation  $(\mathbb{C}, h, 0, \infty)$  or  $(\mathcal{C}, h, -\infty, +\infty)$ . It turns out that these two points have the law of independent samples from the area measure when one conditions on the quantum surface structure of a quantum sphere.

We note that a different perspective on quantum spheres was developed in [David, Kupiainen, Rhodes, and Vargas \[2016\]](#) which follows the construction of Polyakov. The equivalence of the construction in [David, Kupiainen, Rhodes, and Vargas \[ibid.\]](#) and the one described just above was established in [Aru, Huang, and Sun \[2017\]](#). (An approach similar to [Aru, Huang, and Sun \[ibid.\]](#) would likely yield the equivalence of the disk measures considered in [Huang, Rhodes, and Vargas \[2015\]](#) and the quantum disk defined earlier.)

Finally, we mention briefly that there are also some works which construct LQG on non-simply connected surfaces [David, Rhodes, and Vargas \[2016\]](#) and [Guillarmou, Rhodes, and Vargas \[2016\]](#).

### 3 SLE and Liouville quantum gravity

**3.1 The Schramm-Loewner evolution.** The Schramm-Loewner evolution (SLE) was introduced by [Schramm \[2000\]](#) to describe the scaling limits of the interfaces in discrete lattice models in two dimensions, the motivating examples being loop-erased random walk and critical percolation. We will discuss three variants of SLE: chordal, radial, and whole-plane.

Chordal SLE is a random fractal curve which connects two boundary points in a simply connected domain. It is most natural to define it first in  $\mathbb{H}$  and then for other domains by conformal mapping. Suppose that  $\eta$  is a curve in  $\overline{\mathbb{H}}$  from 0 to  $\infty$  which is non-self-crossing and non-self-tracing. For each  $t \geq 0$ , let  $\mathbb{H}_t$  be the unbounded component of  $\mathbb{H} \setminus \eta([0, t])$  and let  $g_t : \mathbb{H}_t \rightarrow \mathbb{H}$  be the unique conformal map with  $g_t(z) - z \rightarrow 0$  as  $z \rightarrow \infty$ . Then Loewner's theorem states that there exists a continuous function  $W : [0, \infty) \rightarrow \mathbb{R}$  such that the maps  $(g_t)$  satisfy the ODE

$$(3-1) \quad \partial_t g_t(z) = \frac{2}{g_t(z) - W_t}, \quad g_0(z) = z$$



(provided  $\eta$  is parameterized appropriately). The driving function  $W$  is explicitly given by  $W_t = g_t(\eta(t))$ . For  $\kappa > 0$ ,  $\text{SLE}_\kappa$  is the curve associated with the choice  $W = \sqrt{\kappa}B$  where  $B$  is a standard Brownian motion. This form of the driving function arises when one makes the assumption that the law of  $\eta$  is conformally invariant and satisfies the following Markov property: for each stopping time  $\tau$  for  $\eta$ , the conditional law of  $g_\tau(\eta|_{[\tau, \infty)}) - W_\tau$  is the same as the law of  $\eta$ . These properties are natural to assume for the scaling limits of two-dimensional discrete lattice models.

The behavior of  $\text{SLE}_\kappa$  strongly depends on the value of  $\kappa$ . When  $\kappa \in (0, 4]$ , it describes a simple curve, when  $\kappa \in (4, 8)$  it is a self-intersecting curve, and when  $\kappa \geq 8$  it is a space-filling curve [Rohde and Schramm \[2005\]](#). Special values of  $\kappa$  which have been proved or conjectured to correspond to discrete lattice models include:

- $\kappa = 1$ : Schnyder woods branches [Li, Sun, and Watson \[2017\]](#)
- $\kappa = 4/3$ : bipolar orientation branches [Kenyon, Miller, Sheffield, and Wilson \[2015, 2017\]](#)
- $\kappa = 2$ : loop-erased random walk [Lawler, Schramm, and Werner \[2004a\]](#)
- $\kappa = 8/3$ : self-avoiding walks [Lawler, Schramm, and Werner \[2004b\]](#)
- $\kappa = 3$ : critical Ising model [Smirnov \[2010\]](#)
- $\kappa = 4$ : level lines of the GFF [Schramm and Sheffield \[2009, 2013\]](#)
- $\kappa = 6$ : critical percolation [Smirnov \[2001\]](#)
- $\kappa = 16/3$ : FK-Ising model [Smirnov \[2010\]](#)
- $\kappa = 8$ : uniform spanning tree [Lawler, Schramm, and Werner \[2004a\]](#)
- $\kappa = 12$ : bipolar orientations [Kenyon, Miller, Sheffield, and Wilson \[2015, 2017\]](#)
- $\kappa = 16$ : Schnyder woods [Li, Sun, and Watson \[2017\]](#)

Radial SLE is a random fractal curve in a simply connected domain which connects a boundary point to an interior point. It is defined first in  $\mathbb{D}$  and then in other domains by conformal mapping. The definition is analogous to the case of chordal SLE, except one solves the radial Loewner ODE

$$(3-2) \quad \partial_t g_t(z) = -g_t(z) \frac{g_t(z) + e^{iW_t}}{g_t(z) - e^{iW_t}}$$

in place of (3-1). As in the case of (3-1), the radial Loewner ODE serves to encode a non-self-crossing and non-self-tracing curve  $\eta$  in terms of a continuous, real-valued function. For each  $t \geq 0$ ,  $g_t$  is the unique conformal map from the component of  $\mathbb{D} \setminus \eta([0, t])$

containing 0 to  $\mathbb{D}$  with  $g'_l(0) > 0$ . Radial  $\text{SLE}_\kappa$  corresponds to the case that  $W = \sqrt{\kappa}B$ . Whole-plane SLE is a random fractal curve which connects two points in the Riemann sphere. It is defined first for the points 0 and  $\infty$  and then for other pairs of points by applying a Möbius transformation. It can be constructed by starting with a radial  $\text{SLE}_\kappa$  in  $\mathbb{C} \setminus (\epsilon\mathbb{D})$  from  $\epsilon$  to  $\infty$  and then taking a limit as  $\epsilon \rightarrow 0$ .

**3.2 Exploring an LQG surface with an SLE.** There are two natural operations that one can perform in the context of planar maps (see the physics references in [Duplantier and Sheffield \[2011\]](#)). Namely:

- One can “glue” together two planar maps with boundary by identifying their edges along a marked boundary segment to produce a planar map decorated with a distinguished interface. If the two maps are chosen independently and uniformly at random, then this interface will in fact be a self-avoiding walk (SAW). (See [Section 6.1](#) for more details.)
- One can also decorate a planar map with an instance of a statistical physics model (e.g. a critical percolation configuration or a uniform spanning tree) and then explore the interfaces of the statistical physics model. (See [Section 6.2](#) for more details in the case of percolation.)

If one takes as an ansatz that LQG describes the large scale behavior of random planar maps, then it is natural to guess that one should be able to perform the same operations in the continuum on LQG. In order to make this mathematically precise, one needs to describe the precise form of the laws of:

- The field  $h$  which describes the underlying LQG surface and
- The law of the interfaces.

In view of the discussion in [Section 2](#), it is natural to expect that quantum wedges, disks, cones, and spheres will play the role of the former. In view of the conformal invariance ansatz for critical models in two-dimensional statistical mechanics, it is natural to expect that SLE-type curves should play the role of the latter.

Recall that LQG comes with the parameter  $\gamma$  and SLE comes with the parameter  $\kappa$ . As we will describe below in more detail, it is important that these parameters are correctly tuned. Namely, it will always be the case that

$$(3-3) \quad \gamma = \min \left( \sqrt{\kappa}, \frac{4}{\sqrt{\kappa}} \right).$$

We emphasize that (3-3) states that for each  $\gamma \in (0, 2)$ , there are precisely two compatible values of  $\kappa$ :  $\kappa = \gamma^2 \in (0, 4)$  and  $\kappa = 16/\gamma^2 > 4$ .

We will now describe some work of [Sheffield \[2016a\]](#), which is the first mathematical result relating SLE to LQG and should be interpreted as the continuous analog of the gluing operation for planar maps with boundary. It is motivated by earlier work of Duplantier from the physics literature [Duplantier \[1998, 1999a,b, 2000\]](#).

**Theorem 3.1.** *Fix  $\kappa \in (0, 4)$  and let  $\gamma = \sqrt{\kappa}$ . Suppose that  $\mathcal{W} = (\mathbb{H}, h, 0, \infty)$  is a  $(\gamma - 2/\gamma)$ -quantum wedge and that  $\eta$  is an independent  $\text{SLE}_\kappa$  process in  $\mathbb{H}$  from 0 to  $\infty$ . Let  $D_1$  (resp.  $D_2$ ) be the component of  $\mathbb{H} \setminus \eta$  which is to the left (resp. right) of  $\eta$ . Then the quantum surfaces  $\mathfrak{D}_1 = (D_1, h, 0, \infty)$  and  $\mathfrak{D}_2 = (D_2, h, 0, \infty)$  are independent  $\gamma$ -quantum wedges. Moreover,  $\mathcal{W}$  and  $\eta$  are a.s. determined by  $\mathfrak{D}_1, \mathfrak{D}_2$ .*

We first emphasize that the independence of  $\mathfrak{D}_1, \mathfrak{D}_2$  in [Theorem 3.1](#) is in terms of quantum surfaces, which are themselves defined modulo conformal transformation. The  $\gamma$ -quantum wedges  $\mathfrak{D}_1, \mathfrak{D}_2$  which are parameterized by the regions which are to the left and right of  $\eta$  each have their own boundary length measure. Therefore for any point  $z$  along  $\eta$ , one can measure the boundary length distance from  $z$  to 0 along the left or the right side of  $\eta$  (i.e., using  $\mathfrak{D}_1$  or  $\mathfrak{D}_2$ ). One of the other main results of [Sheffield \[2016a\]](#) is that these two quantities agree. An  $\text{SLE}_\kappa$  curve with  $\kappa \in (0, 4)$  therefore has a well-defined notion of quantum length. Moreover, this also allows one to think of [Sheffield \[ibid.\]](#) as a statement about welding quantum surfaces together.

The cutting/welding operation first established in [Sheffield \[ibid.\]](#) was substantially generalized in [Duplantier, Miller, and Sheffield \[2014\]](#), in which many other SLE explorations of LQG surfaces were studied. Let us mention one result in the context of an  $\text{SLE}_\kappa$  process for  $\kappa \in (4, 8)$ .

**Theorem 3.2.** *Fix  $\kappa \in (4, 8)$  and let  $\gamma = 4/\sqrt{\kappa}$ . Suppose that  $\mathcal{W} = (\mathbb{H}, h, 0, \infty)$  is a  $(4/\gamma - \gamma/2)$ -quantum wedge and that  $\eta$  is an independent  $\text{SLE}_\kappa$  process. Then the quantum surfaces parameterized by the components of  $\mathbb{H} \setminus \eta$  are conditionally independent quantum disks given their boundary lengths. The boundary lengths of these disks which are on the left (resp. right) side of  $\eta$  are in correspondence with the jumps of a  $\kappa/4$ -stable Lévy process  $L$  (resp.  $R$ ) with only downward jumps. Moreover,  $L$  and  $R$  are independent.*

The time parameterization of the Lévy processes  $L$  and  $R$  gives rise to an intrinsic notion of time for  $\eta$  which in [Duplantier, Miller, and Sheffield \[ibid.\]](#) is referred to as the *quantum natural time*.

There are also similar results for explorations of the finite volume surfaces by SLE processes. We will focus on one such result here (proved in [Miller and Sheffield \[2015c\]](#)) which is relevant for the construction of the metric on  $\sqrt{8/3}$ -LQG.

**Theorem 3.3.** *Suppose that  $\gamma = \sqrt{8/3}$  and that  $\mathfrak{S} = (\mathbb{C}, h, -\infty, +\infty)$  is a quantum sphere. Let  $\eta$  be a whole-plane  $\text{SLE}_6$  process in  $\mathbb{C}$  from  $-\infty$  to  $+\infty$  which is sampled*

independently of  $h$  and then reparameterized according to quantum natural time. Let  $X_t$  be the quantum boundary length of the component  $C_t$  of  $\mathcal{C} \setminus \eta([0, t])$  containing  $+\infty$ . Then  $X$  evolves as the time-reversal of a  $3/2$ -stable Lévy excursion with only upward jumps. For a given time  $t$ , the conditional law of the surface parameterized by  $C_t$  given  $X_t$  is that of a quantum disk with boundary length  $X_t$  weighted by its quantum area and the conditional law of the point  $\eta(t)$  is given by the quantum boundary length measure on  $\partial C_t$ . Moreover, the surfaces parameterized by the other components of  $\mathcal{C} \setminus \eta([0, t])$  are quantum disks which are conditionally independent given  $X|_{[0, t]}$ , each correspond to a downward jump of  $X|_{[0, t]}$ , and have quantum boundary length given by the corresponding jump.

## 4 Construction of the metric

In this section we will describe the construction of the metric on  $\sqrt{8/3}$ -LQG from [Miller and Sheffield \[2015b, 2016d,e\]](#). The construction is strongly motivated by discrete considerations, which we will review in [Section 4.1](#), before reviewing the continuum construction in [Section 4.2](#).

**4.1 The Eden growth model.** Suppose that  $G = (V, E)$  is a connected graph. In the Eden growth model [Eden \[1961\]](#) (or first-passage percolation [Hammersley and Welsh \[1965\]](#)), one associates with each  $e \in E$  an independent  $\exp(1)$  weight  $Z_e$ . The aim is then to understand the *random* metric  $d_{\text{FPP}}$  on  $G$  which assigns length  $Z_e$  to each  $e \in E$ . Suppose that  $x \in V$ . By the memoryless property of the exponential distribution, there is a simple Markovian way of growing the  $d_{\text{FPP}}$ -metric ball centered at  $x$ . Namely, one inductively defines an increasing sequence of clusters  $C_n$  as follows.

- Set  $C_0 = \{x\}$ .
- Given that  $C_n$  is defined, choose an edge  $e = \{y, z\}$  uniformly at random among those with  $y \in C_n$  and  $z \notin C_n$ , and then take  $C_{n+1} = C_n \cup \{z\}$ .

It is an interesting question to analyze the large-scale behavior of the clusters  $C_n$  on a given graph  $G$ . One of the most famous examples is the case  $G = \mathbb{Z}^2$ , i.e., the two-dimensional integer lattice (see the left side of [Figure 1](#)). It was shown by [Cox and Durrett \[1981\]](#) that the macroscopic shape of  $C_n$  is convex but computer simulations suggest that it is not a Euclidean ball. The reason for this is that  $\mathbb{Z}^2$  is not sufficiently isotropic. In [Vahidi-Asl and Wierman \[1990, 1992\]](#), Wiermann and Vahidi-Asl considered the Eden model on the Delaunay triangulation associated with the Voronoi tessellation of a Poisson point process with Lebesgue intensity on  $\mathbb{R}^2$  and showed that at large scales the Eden

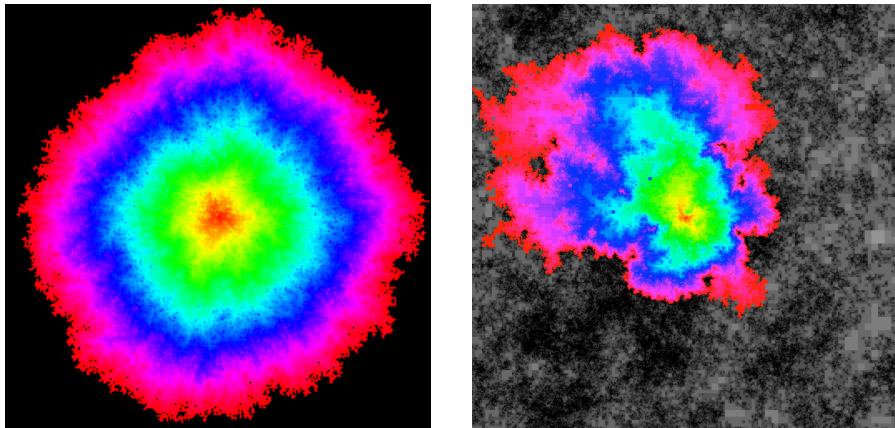


Figure 1: **Left:** Edén model on  $\mathbb{Z}^2$ . **Right:** Edén model on a graph approximation of  $\sqrt{8/3}$ -LQG. This serves as a discretization of  $\text{QLE}(8/3, 0)$ .

growth model is well-approximated by a Euclidean ball. The reason for the difference between the setting considered in [Vahidi-Asl and Wierman \[1990, 1992\]](#) and  $\mathbb{Z}^2$  is that such a Poisson point process does not have preferential directions due to the underlying randomness and the rotational invariance of Lebesgue measure.

Also being random, it is natural to expect that the Edén growth model on a random planar map should at large scales be approximated by a metric ball. This has now been proved by [Curien and Gall \[2015\]](#) in the case of the planar dual of a random triangulation. The Edén growth model on the planar dual of a random triangulation is natural to study because it can be described in terms of a so-called *peeling process*, in which one picks an edge uniformly on the boundary of the cluster so far and then reveals the opposing triangle. In particular, this exploration procedure respects the Markovian structure of a random triangulation in the sense that one can describe the law of the regions cut out by the growth process (uniform triangulations of the disk) as well as the evolution of the boundary length of the cluster. We will eventually want to make sense of the Edén model on a quantum sphere which satisfies the same properties. In order to motivate the construction, we will describe a variant of the Edén growth model on the planar dual of a random triangulation which involves two operations we know how to perform in the continuum (in view of the connection between SLE and LQG) and respects the Markovian structure of the quantum sphere in the same way as described above. Namely, we fix  $k \in \mathbb{N}$  and then define an increasing sequence of clusters  $C_n$  in the dual of the map as follows.

- Set  $C_0 = \{F\}$  where  $F$  is the root face.

- Given that  $C_n$  is defined, pick two edges  $e_1, e_2$  on the outer boundary of  $C_n$  uniformly at random and color the vertices on the clockwise (resp. counterclockwise) arc of the outer boundary of  $C_n$  from  $e_1$  to  $e_2$  blue (resp. yellow).
- Color the vertices in the remainder of the map blue (resp. yellow) independently with probability  $1/2$ . Take  $C_{n+1}$  to be the union of  $C_n$  and  $k$  edges which lie along the blue/yellow percolation interface starting from one of the marked edges described above.

This growth process can also be described in terms of a peeling process, hence it respects the Markovian structure of a random triangulation in the same way as the Eden growth model. It is also natural to expect there to be universality. Namely, the macroscopic behavior of the cluster should not depend on the specific geometry of the “chunks” that we are adding at each stage. That is, it should be the case that for each fixed  $k \in \mathbb{N}$ , the cluster  $C_n$  at large scales should be well-approximated by a metric ball.

**4.2 QLE(8/3, 0): The Eden model on the quantum sphere.** Suppose that  $(\mathcal{S}, x, y)$  is a quantum sphere marked by two independent samples  $x, y$  from the quantum measure. We will now describe a variant of the above construction on  $\mathcal{S}$  starting from  $x$  and targeted at  $y$ , with SLE<sub>6</sub> as the continuum analog of the percolation interface. Fix  $\delta > 0$  and let  $\eta_0$  be a whole-plane SLE<sub>6</sub> on  $\mathcal{S}$  from  $x$  to  $y$  with the quantum natural time parameterization as in [Theorem 3.2](#). Then [Theorem 3.2](#) implies that:

- The quantum surfaces parameterized by the components of  $\mathcal{S} \setminus \eta_0([0, \delta])$  which do not contain  $y$  are conditionally independent quantum disks.
- The component  $C_0$  of  $\mathcal{S} \setminus \eta_0([0, \delta])$  containing  $y$  has the law of a quantum disk weighted by its quantum area.
- The conditional law of  $\eta_0(\delta)$  is given by the quantum boundary measure on  $\partial C_0$ .

We now pick  $z_1$  from the quantum boundary measure on  $\partial C_0$  and then let  $\eta_1$  be a radial SLE<sub>6</sub> in  $C_0$  from  $z_1$  to  $y$ , parameterized by quantum natural time. Then it also holds that the quantum surfaces parameterized by the components of  $C_0 \setminus \eta_1([0, \delta])$  which do not contain  $y$  are conditionally independent quantum disks, the component  $C_1$  of  $C_0 \setminus \eta_1([0, \delta])$  containing  $y$  has the law of a quantum disk weighted by its quantum area, and  $\eta_1(\delta)$  is uniformly distributed according to the quantum boundary measure on  $\partial C_1$ .

The  $\delta$ -approximation  $\Gamma^{\delta, x \rightarrow y}$  to QLE(8/3, 0) (*quantum Loewner evolution* with parameters  $\gamma^2 = 8/3$  and  $\eta = 0$ ; see [Section 4.3](#) for more on the family of processes  $\text{QLE}(\gamma^2, \eta)$ ) from  $x$  to  $y$  is defined by iterating the above procedure until it eventually reaches  $y$ . One can view  $\Gamma^{\delta, x \rightarrow y}$  as arising by starting with a whole-plane SLE<sub>6</sub> process  $\eta$  on  $\mathcal{S}$  from  $x$  to

$y$  parameterized by quantum natural time and then resampling the location of the tip of  $\eta$  at each time of the form  $k\delta$  where  $k \in \mathbb{N}$ . Due to the way the process is constructed, we emphasize that the following hold:

- The surfaces parameterized by the components of  $\mathcal{S} \setminus \Gamma_t^{\delta, x \rightarrow y}$  which do not contain  $y$  are conditionally independent quantum disks.
- The surface parameterized by the component  $C_t$  of  $\mathcal{S} \setminus \Gamma_t^{\delta, x \rightarrow y}$  which contains  $y$  on its boundary is a quantum disk weighted by its area.
- The evolution of the quantum boundary length  $X_t$  of  $\partial C_t$  is the same as in the case of  $\eta$ , i.e., it is given by the time-reversal of a  $3/2$ -stable Lévy excursion.

That is,  $\Gamma^{\delta, x \rightarrow y}$  respects the Markovian structure of a quantum sphere. The growth process  $\text{QLE}(8/3, 0)$  is constructed by taking a limit as  $\delta \rightarrow 0$  of the  $\delta$ -approximation defined above. All of the above properties are preserved by the limit.

Parameterizing an  $\text{SLE}_6$  by quantum natural time on a quantum sphere is the continuum analog of parameterizing a percolation exploration on a random planar map according to the number of edges that the percolation has visited. This is not the correct notion of time if one wants to define a metric space structure since one should be adding particles to the growth process at a rate which is proportional to its boundary length. One is thus led to make the following change of time: set

$$D(t) = \int_0^t \frac{1}{X_u} du$$

and then let  $s(r) = \inf\{t \geq 0 : D(t) > r\}$ . Due to the above interpretation, the time-parameterization  $s(r)$  is called the *quantum distance time*. Let  $\Gamma^{x \rightarrow y}$  be a  $\text{QLE}(8/3, 0)$  parameterized by  $s(r)$  time. Then it will ultimately be the case that  $\Gamma_r^{x \rightarrow y}$  defines a metric ball of radius  $r$ , and we will sketch the proof of this fact in what follows.

Let  $(x_n)$  be a sequence of i.i.d. points chosen from the quantum measure on  $\mathcal{S}$ . For each  $i \neq j$ , we let  $\Gamma^{x_i \rightarrow x_j}$  be a conditionally independent (given  $\mathcal{S}$ )  $\text{QLE}(8/3, 0)$  from  $x_i$  to  $x_j$  with the quantum distance parameterization. We then set  $d_{\mathbb{Q}}(x_i, x_j)$  to be the amount of time it takes for  $\Gamma^{x_i \rightarrow x_j}$  to reach  $x_j$ . We want to show that  $d_{\mathbb{Q}}(x_i, x_j)$  defines a metric on the set  $(x_i)$  which is a.s. determined by  $\mathcal{S}$ . It is obvious from the construction that  $d_{\mathbb{Q}}(x_i, x_j) > 0$  for  $i \neq j$ , so to establish the metric property it suffices to prove that  $d_{\mathbb{Q}}$  is symmetric and satisfies the triangle inequality. This is proved by making use of a strategy developed by Sheffield, Watson, and Wu in the context of  $\text{CLE}_4$ .

Symmetry, the triangle inequality, and the fact that  $d_{\mathbb{Q}}$  is a.s. determined by  $\mathcal{S}$  all follow from the following stronger statement (taking without loss of generality  $x = x_1$  and  $y = x_2$ ). Let  $\Theta$  denote the law of  $(\mathcal{S}, x, y, \Gamma^{x \rightarrow y}, \Gamma^{y \rightarrow x}, U)$  where  $U$  is uniform in

$[0, 1]$  independently of everything else. Let  $\Theta^{x \rightarrow y}$  (resp.  $\Theta^{y \rightarrow x}$ ) be the law whose Radon-Nikodym derivative with respect to  $\Theta$  is given by  $d_{\mathbb{Q}}(x, y)$  (resp.  $d_{\mathbb{Q}}(y, x)$ ). That is,

$$\frac{d\Theta^{x \rightarrow y}}{d\Theta} = d_{\mathbb{Q}}(x, y) \quad \text{and} \quad \frac{d\Theta^{y \rightarrow x}}{d\Theta} = d_{\mathbb{Q}}(y, x).$$

We want to show that  $\Theta^{x \rightarrow y} = \Theta^{y \rightarrow x}$  because then the uniqueness of Radon-Nikodym derivatives implies that  $d_{\mathbb{Q}}(x, y) = d_{\mathbb{Q}}(y, x)$ . Since  $\Gamma^{x \rightarrow y}$  and  $\Gamma^{y \rightarrow x}$  were taken to be conditionally independent given  $\mathcal{S}$ , this also implies that the common value of  $d_{\mathbb{Q}}(x, y)$  and  $d_{\mathbb{Q}}(y, x)$  is a.s. determined by  $\mathcal{S}$ .

The main step in proving this is the following, which is a restatement of [Miller and Sheffield \[2015b\]](#), Lemma 1.2].

**Lemma 4.1.** *Let  $\tau = U d_{\mathbb{Q}}(x, y)$  so that  $\tau$  is uniform in  $[0, d_{\mathbb{Q}}(x, y)]$  and let  $\bar{\tau} = \inf\{t \geq 0 : \Gamma_{\tau}^{x \rightarrow y} \cap \Gamma_t^{y \rightarrow x} \neq \emptyset\}$ . We similarly let  $\bar{\sigma} = U d_{\mathbb{Q}}(y, x)$  and  $\sigma = \inf\{t \geq 0 : \Gamma_t^{x \rightarrow y} \cap \Gamma_{\bar{\sigma}}^{y \rightarrow x} \neq \emptyset\}$ . Then the  $\Theta^{x \rightarrow y}$  law of  $(\mathcal{S}, x, y, \Gamma^{x \rightarrow y}|_{[0, \tau]}, \Gamma^{y \rightarrow x}|_{[0, \bar{\tau}]})$  is the same as the  $\Theta^{y \rightarrow x}$  law of  $(\mathcal{S}, x, y, \Gamma^{x \rightarrow y}|_{[0, \sigma]}, \Gamma^{y \rightarrow x}|_{[0, \bar{\sigma}]})$ .*

Upon proving [Lemma 4.1](#), the proof is completed by showing that the  $\Theta^{x \rightarrow y}$  conditional law of  $\Gamma^{x \rightarrow y}, \Gamma^{y \rightarrow x}$  given  $(\mathcal{S}, x, y, \Gamma^{x \rightarrow y}|_{[0, \tau]}, \Gamma^{y \rightarrow x}|_{[0, \bar{\tau}]})$  is the same as the  $\Theta^{y \rightarrow x}$  conditional law of  $\Gamma^{x \rightarrow y}, \Gamma^{y \rightarrow x}$  given  $(\mathcal{S}, x, y, \Gamma^{x \rightarrow y}|_{[0, \sigma]}, \Gamma^{y \rightarrow x}|_{[0, \bar{\sigma}]})$ .

[Lemma 4.1](#) turns out to be a consequence of a corresponding symmetry statement for whole-plane SLE<sub>6</sub>, which in a certain sense reduces to the time-reversal symmetry of whole-plane SLE<sub>6</sub> [Miller and Sheffield \[2017\]](#), and then reshuffling the SLE<sub>6</sub> as described above to obtain a QLE(8/3, 0).

The rest of the program carried out in [Miller and Sheffield \[2016d,e\]](#) consists of showing that:

- The metric defined above extends uniquely in a continuous manner to the entire quantum sphere yielding a metric space which is homeomorphic to  $\mathbb{S}^2$ . Moreover, the resulting metric space is geodesic and isometric to the Brownian map using the characterization from [Miller and Sheffield \[2015a\]](#). We will describe this in further detail in the next section.
- The quantum sphere instance is a.s. determined by the metric measure space structure. This implies that the Brownian map possesses a canonical embedding into  $\mathbb{S}^2$  which takes it to a form of  $\sqrt{8/3}$ -LQG. This is based on an argument which is similar to that given in [Duplantier, Miller, and Sheffield \[2014\]](#), Section 10].

**4.3** QLE( $\gamma^2, \eta$ ). The process QLE(8/3, 0) described in [Section 4.2](#) is a part of a general family of growth processes which are the conjectural scaling limits of a family of growth models on random surfaces which we now describe.



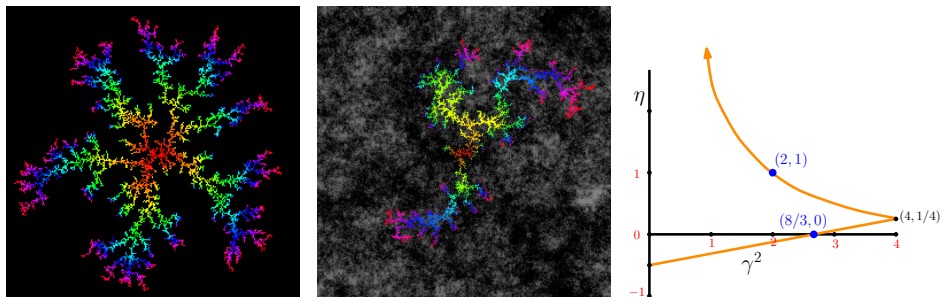


Figure 2: **Left:** DLA on  $\mathbb{Z}^2$ . **Middle:** DLA on a graph approximation of  $\sqrt{2}$ -LQG. This serves as a discretization of QLE(2, 1). **Right:** Plot of  $(\gamma^2, \eta)$  values for which QLE( $\gamma^2, \eta$ ) is constructed in Miller and Sheffield [2016f].

The dielectric breakdown model (DBM) Niemeyer, Pietronero, and Wiesmann [1984] with parameter  $\eta$  is a family of models which interpolate between the Eden model and diffusion limited aggregation (DLA). If  $\mu_{\text{HARM}}$  (resp.  $\mu_{\text{LEN}}$ ) denotes the natural harmonic (resp. length) measure on the underlying surface, then microscopic particles are added according to the measure

$$\left( \frac{d\mu_{\text{HARM}}}{d\mu_{\text{LEN}}} \right)^\eta d\mu_{\text{LEN}}.$$

In particular,  $\eta = 0$  corresponds to the Eden model and  $\eta = 1$  corresponds to DLA.

One can apply the tip-rerandomization procedure to SLE $_\kappa$  or SLE $_{\kappa'}$  coupled with  $\gamma$ -LQG for other values of  $\kappa, \kappa'$  and  $\gamma$  provided  $\kappa = \gamma^2$  and  $\kappa' = 16/\gamma^2$ . The resulting process, which is called QLE( $\gamma^2, \eta$ ) and is defined and analyzed in Miller and Sheffield [2016f], is the conjectural scaling limit of  $\eta$ -DBM on a  $\gamma$ -LQG surface where

$$\eta = \frac{3\gamma^2}{16} - \frac{1}{2} \quad \text{or} \quad \eta = \frac{3}{\gamma^2} - \frac{1}{2}.$$

(The parameter  $\gamma$  determines the type of LQG surface on which the process grows and the value of  $\eta$  determines the manner in which it grows.) Special parameter values include QLE(8/3, 0) and QLE(2, 1), the latter being the conjectural scaling limit of DLA on a tree-weighted random planar map. See Figure 2. It remains an open question to construct QLE( $\gamma^2, \eta$ ) for the full range of parameter values.

## 5 Equivalence with the Brownian map

In the previous section, we have described how to build a metric space structure on top of  $\sqrt{8/3}$ -LQG. We will describe here how it is checked that this metric space structure is equivalent with the Brownian map.

**5.1 Brownian map definition.** The Brownian map is constructed from a continuum version of the Cori-Vauquelin-Schaeffer bijection [Cori and Vauquelin \[1981\]](#) and [Schaeffer \[1998\]](#) using the *Brownian snake*. Suppose that  $Y : [0, T] \rightarrow \mathbb{R}_+$  is picked from the excursion measure for Brownian motion. Given  $Y$ , we let  $X$  be a centered Gaussian process with  $X_0 = 0$  and

$$\text{cov}(X_s, X_t) = \inf\{Y_r : r \in [s, t]\}.$$

For  $s < t$  and  $[t, s] = [0, T] \setminus (s, t)$ , we set

$$d^\circ(s, t) = X_s + X_t - 2 \max \left( \min_{r \in [s, t]} X_r, \min_{r \in [t, s]} X_r \right).$$

Let  $\mathcal{T}$  be the instance of the continuum random tree (CRT) [Aldous \[1991\]](#) encoded by  $Y$  and let  $\rho : [0, T] \rightarrow \mathcal{T}$  be the corresponding projection map. We then set

$$d_{\mathcal{T}}(a, b) = \min\{d^\circ(s, t) : \rho(s) = a, \quad \rho(t) = b\}.$$

Finally, for  $a, b \in \mathcal{T}$ , we set

$$d(a, b) = \inf \left\{ \sum_{j=1}^k d_{\mathcal{T}}^\circ(a_{j-1}, a_j) \right\}$$

where the infimum is over all  $k \in \mathbb{N}$  and  $a_0 = a, a_1, \dots, a_k = b$  in  $\mathcal{T}$ . Quotienting by the equivalence relation  $a \cong b$  if and only if  $d(a, b) = 0$  yields a metric space  $(S, d)$ . It is naturally equipped with a measure  $\nu$  by taking the projection of Lebesgue measure on  $[0, T]$ . Finally,  $(S, d, \nu)$  is naturally marked by the points  $x$  and  $y$  which are respectively given by the projections of  $t = 0$  and the value of  $t$  at which  $X$  attains its infimum. The space  $(S, d, \nu, x, y)$  is the (doubly marked) Brownian map and we denote its law by  $\mu_{\text{SPH}}^2$ . It is an infinite measure (since the Brownian excursion measure is an infinite measure). As mentioned earlier, it follows from [Le Gall and Paulin \[2008\]](#) (see also [Miermont \[2008\]](#)) that  $(S, d)$  a.s. has the topology of  $\mathbb{S}^2$ . Also, the law of  $(S, d, \nu, x, y)$  is invariant under the operation of resampling  $x, y$  independently from  $\nu$ . The standard unit area Brownian map arises by conditioning  $\mu_{\text{SPH}}^2$  to have total area equal to 1. Equivalently, one can take

the construction above and condition on the Brownian excursion  $Y$  to have length equal to 1.

Variants of the Brownian map with other topologies are defined in a similar manner. For example, the Brownian disk, half-plane, and plane are defined this way in [Bettinelli and Miermont \[2017\]](#), [Abraham and Gall \[2015\]](#), [Gall \[2017\]](#), [Baur, Miermont, and Ray \[2016\]](#), [Gwynne and Miller \[2017d\]](#), and [Curien and Le Gall \[2014\]](#).

**5.2 The  $\alpha$ -stable Lévy net.** Suppose that we have a doubly-marked metric space  $(S, d, x, y)$  which has the topology of  $\mathbb{S}^2$ . For each  $r \in [0, d(x, y)]$ , we define the *filled metric ball*  $B^\bullet(x, r)$  to be the closure of the complement of the  $y$ -containing component of  $S \setminus B(x, r)$ . The *metric net* of  $(S, d, x, y)$  is the closure of the union of  $\partial B^\bullet(x, r)$  over  $r \in [0, d(x, y)]$ . It turns out that it is possible to give an explicit description of the law of the metric net of the Brownian map and, as we will explain just below, this is one of the main ingredients which characterizes its law.

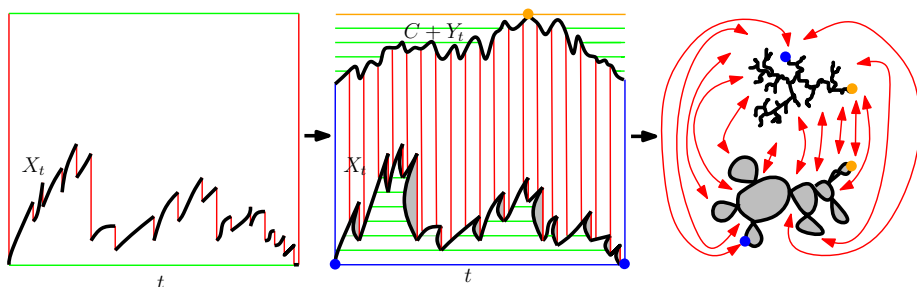


Figure 3: Illustration of the construction of the  $\alpha$ -stable Lévy net.

The  $\alpha$ -stable Lévy net is defined as follows. Fix  $\alpha \in (1, 2)$  and suppose that  $X$  is the time-reversal of an  $\alpha$ -stable Lévy excursion with only upward jumps (so that  $X$  has only downward jumps). We define the height process  $Y$  associated with  $X$  to be given at time  $t$  by the amount of local time that  $X|_{[t, T]}$  spends at its running infimum. Both  $X$  and  $Y$  encode trees. Namely, associated with  $X$  is a looptree (as first constructed and defined by Curien and Kortchemski [Curien and Kortchemski \[2014\]](#)) which is obtained by considering the graph of  $X$  and then replacing each of the downward jumps by a topological disk that does not otherwise cross the graph of  $X$  and with the disks pairwise disjoint. Points on the graph of  $X$  or the disk boundaries are considered to be equivalent if they can be connected by a horizontal chord which lies entirely below the graph of  $X$ . The tree associated with  $Y$  is defined by declaring two points on the graph of  $Y$  to be equivalent if they

can be connected by a horizontal chord which lies entirely above the graph of  $Y$ . We can then glue these two trees together as illustrated in [Figure 3](#) to obtain the  $\alpha$ -stable Lévy net. The tree associated with  $X$  (resp.  $Y$ ) is called the dual (resp. geodesic) tree of the Lévy net instance.

Due to the construction, points on the geodesic tree which are identified with each other all have the same distance to the root in the geodesic tree. Therefore every point in the Lévy net has a well-defined distance to the root, hence one can talk about metric balls which are centered at the root. It is not difficult to see that one can in fact associate with each such metric ball a boundary length and that this boundary length evolves as a so-called continuous state branching process (CSBP) as one reduces the radius of the ball. In fact, the boundary lengths between any finite collection of geodesics evolve as collection of independent CSBPs as the ball radius is reduced and this property essentially characterizes the  $\alpha$ -stable Lévy net (as it determines how the geodesics are glued together).

It is proved in [Miller and Sheffield \[2015a\]](#) that the law of the metric net of a sample from  $\mu_{\text{SPH}}^2$  is given by that of a  $3/2$ -stable Lévy net. The following is a restatement of [Miller and Sheffield \[ibid., Theorem 4.6\]](#).

**Theorem 5.1.** *The doubly marked Brownian map measure  $\mu_{\text{SPH}}^2$  is the unique infinite measure on doubly-marked metric measure spaces  $(S, d, \nu, x, y)$  with the topology of  $\mathbb{S}^2$  which satisfy the following properties:*

1. *The law of  $(S, d, \nu, x, y)$  is invariant under resampling  $x$  and  $y$  independently from  $\nu$ .*
2. *The law of the metric net from  $x$  to  $y$  agrees with that of the  $\alpha$ -stable Lévy net for some value of  $\alpha \in (1, 2)$ .*
3. *For each fixed  $r > 0$ , the metric measure spaces  $B^\bullet(x, r)$  and  $S \setminus B^\bullet(x, r)$  (equipped with the interior internal metric and the restriction of  $\nu$ ) are conditionally independent given the boundary length of  $\partial B^\bullet(x, r)$ .*

The proof of [Theorem 5.1](#) given in [Miller and Sheffield \[ibid.\]](#) shows that the assumptions necessarily imply that  $\alpha = 3/2$ , which is why we do not need to make this assumption explicitly in the statement of [Theorem 5.1](#).

### 5.3 QLE(8/3, 0) metric satisfies the axioms which characterize the Brownian map.

To check that the QLE(8/3, 0) metric defined on the  $\sqrt{8/3}$ -quantum sphere defines an instance of the Brownian map, it suffices to check that the axioms of [Theorem 5.1](#) are satisfied. The construction of the QLE(8/3, 0) metric in fact implies that the first and third axioms are satisfied, so the main challenge is to check the second axiom. This is one of the aims of [Miller and Sheffield \[2016d\]](#) and is closely related to the form of the

evolution of the boundary length when one performs an  $\text{SLE}_6$  exploration on a quantum sphere as described in [Theorem 3.3](#).

Upon proving the equivalence of the  $\text{QLE}(8/3, 0)$  metric on a quantum sphere with the Brownian map, it readily follows that several other types of quantum surfaces with  $\gamma = \sqrt{8/3}$  are equivalent to certain types of Brownian surfaces. Namely, the Brownian disk, half-plane, and plane are respectively equivalent to the quantum disk,  $\sqrt{8/3}$ -quantum wedge, and  $\sqrt{8/3}$ -quantum cone [Miller and Sheffield \[2016d\]](#), [Gwynne and Miller \[2017d\]](#), and [Gall \[2017\]](#).

## 6 Scaling limits

The equivalence of LQG and Brownian surfaces allows one to define SLE on a Brownian surface in a canonical way as the embedding of a Brownian surface is a.s. determined by the metric measure space structure [Miller and Sheffield \[2016e\]](#). This makes it possible to prove that certain statistical physics models on uniformly random planar maps converge to SLE. The natural topology of convergence is the so-called Gromov-Hausdorff-Prokhorov-uniform topology (GHPU) developed in [Gwynne and Miller \[2017d\]](#), which is an extension of the Gromov-Hausdorff topology to curve-decorated metric measure spaces. (We note that a number of other scaling limit results for random planar maps decorated with a statistical mechanics model toward SLE on LQG have been proved in e.g. [Sheffield \[2016b\]](#), [Kenyon, Miller, Sheffield, and Wilson \[2015\]](#), [Gwynne, Kassel, Miller, and Wilson \[2016\]](#), and [Li, Sun, and Watson \[2017\]](#) in the so-called peanosphere topology, which is developed in [Duplantier, Miller, and Sheffield \[2014\]](#).)

**6.1 Self-avoiding walks.** Recall that the self-avoiding walk (SAW) is the uniform measure on simple paths of a given length on a graph. The SAW on random planar maps was important historically because it was used by [Duplantier and I. Kostov \[1988\]](#) and [Duplantier and I. K. Kostov \[1990\]](#) as a test case of the KPZ formula [Knizhnik, A. M. Polyakov, and Zamolodchikov \[1988\]](#). It is a particularly natural model to consider as it admits a rather simple construction. Namely, one starts with two independent uniformly random quadrangulations of the disk with simple boundary, perimeter  $2\ell$ , and  $m$  faces as illustrated in the left side of [Figure 4](#). If one glues the two disks along a boundary segment of length  $2s < 2\ell$ , then one obtains a quadrangulation of the disk with perimeter  $2(\ell - s)$  and  $2m$  faces decorated by a distinguished path. Conditional on the map, it is not difficult to see that the path is uniformly random (i.e., a SAW) conditioned on having  $m$  faces to its left and right. In the limit as  $\ell, m \rightarrow \infty$ , one obtains a gluing of two UIHPQ's with simple boundary.

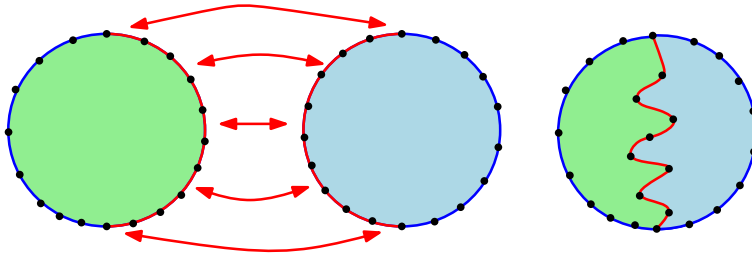


Figure 4: Two independent uniform quadrangulations of the disk with simple boundary, perimeter  $2\ell$ , and  $m$  faces are glued together along a marked boundary arc of length  $2s$  to produce a quadrangulation of the disk with a distinguished path of length  $2s$ . The conditional law of the path is uniform among all simple paths (i.e., a self-avoiding walk) conditioned on having  $m$  faces to its left and right.

The main result of [Gwynne and Miller \[2017d\]](#) implies that the UIHPQ's converge to independent Brownian half-planes, equivalently independent  $\sqrt{8/3}$ -quantum wedges together with their  $\text{QLE}(8/3, 0)$  metric. Proving the convergence of the SAW in this setting amounts to showing that the discrete graph gluing of the two UIHPQ's converges to the metric gluing of the limiting Brownian half-plane instances along their boundaries. This is accomplished in [Gwynne and Miller \[2016a\]](#). Each Brownian half-plane instance is equivalent to a  $\sqrt{8/3}$ -quantum wedge with its  $\text{QLE}(8/3, 0)$  metric. In order to identify the scaling limit of the SAW with  $\text{SLE}_{8/3}$ , it is necessary to show that the conformal welding of two such quantum wedges developed in [Sheffield \[2016a\]](#) is equivalent to the metric gluing of the two  $\sqrt{8/3}$ -quantum wedges with their  $\text{QLE}(8/3, 0)$  metric as it is shown in [Sheffield \[ibid.\]](#) that the interface is  $\text{SLE}_{8/3}$ . This is the main result of [Gwynne and Miller \[2016b\]](#). Combining everything gives the convergence of the SAW on random planar maps to  $\text{SLE}_{8/3}$  on  $\sqrt{8/3}$ -LQG.

**6.2 Percolation.** It is also natural to consider critical percolation on a uniformly random planar map. The critical percolation threshold has been computed for a number of different planar map types (see, e.g., [Angel \[2003\]](#) and [Angel and Curien \[2015\]](#)). The article [Gwynne and Miller \[2017b\]](#) establishes the convergence of the interfaces of critical face percolation on a uniformly random quadrangulation of the disk. This is the model in which the faces of a quadrangulation are declared to be either open or closed independently with probability  $3/4$  or  $1/4$  [Angel and Curien \[2015\]](#). (The reason that the critical threshold is  $3/4$  and not  $1/2$  is because open faces are adjacent if they share an edge and closed faces are adjacent if they share a vertex.) The underlying quadrangulation of the disk converges to the Brownian disk [Bettinelli and Miermont \[2017\]](#) and [Gwynne and Miller \[2017c\]](#),

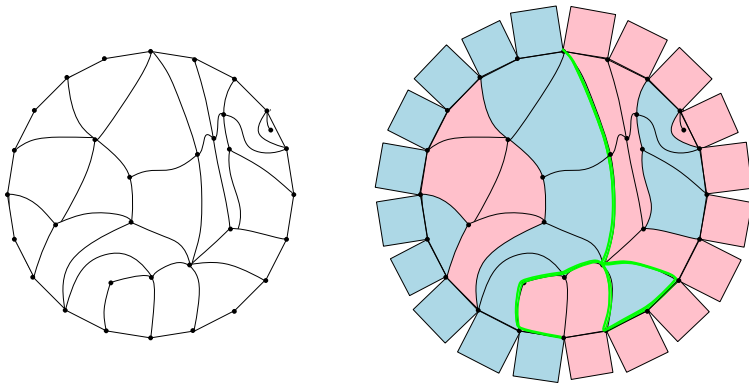


Figure 5: **Left:** A quadrangulation of the disk with simple boundary. **Right:** Blue (resp. red) quadrilaterals have been glued along two marked boundary arcs. Faces on the inside are colored blue (resp. red) independently with probability  $3/4$  (resp.  $1/4$ ). The green path is the interface between the cluster of blue (resp. red) quadrilaterals which are connected to the blue (resp. red) boundary arc.

equivalently a quantum disk. The main result of [Gwynne and Miller \[2017b\]](#) is that the percolation interface converges jointly with the underlying quadrangulation to  $\text{SLE}_6$  on the Brownian disk. The proof proceeds in a very different manner than the case of the SAW. Namely, the idea is to show that the percolation interface converges in the limit to a continuum path which is a Markovian exploration of a Brownian disk with the property that its complementary components are Brownian disks given their boundary lengths and it turns out that this property characterizes  $\text{SLE}_6$  on the Brownian disk [Gwynne and Miller \[2017a\]](#) (recall also [Theorem 3.2](#)).

## References

- Céline Abraham and Jean-François Le Gall (2015). “[Excursion theory for Brownian motion indexed by the Brownian tree](#)”. arXiv: [1509.06616](#) (cit. on p. [2982](#)).
- David Aldous (1991). “[The continuum random tree. I](#)”. *Ann. Probab.* 19.1, pp. 1–28. MR: [1085326](#) (cit. on p. [2981](#)).
- Jan Ambjørn, Bergfinnur Durhuus, and Thordur Jonsson (1997). *Quantum geometry*. Cambridge Monographs on Mathematical Physics. A statistical field theory approach. Cambridge University Press, Cambridge, pp. xiv+363. MR: [1465433](#) (cit. on p. [2967](#)).
- O. Angel (2003). “[Growth and percolation on the uniform infinite planar triangulation](#)”. *Geom. Funct. Anal.* 13.5, pp. 935–974. MR: [2024412](#) (cit. on p. [2985](#)).

- Omer Angel and Nicolas Curien (2015). “[Percolations on random maps I: Half-plane models](#)”. *Ann. Inst. Henri Poincaré Probab. Stat.* 51.2, pp. 405–431. MR: [3335009](#) (cit. on p. 2985).
- Juhan Aru, Yichao Huang, and Xin Sun (2017). “[Two perspectives of the 2D unit area quantum sphere and their equivalence](#)”. *Comm. Math. Phys.* 356.1, pp. 261–283. MR: [3694028](#) (cit. on p. 2971).
- Erich Baur, Grégory Miermont, and Gourab Ray (2016). “[Classification of scaling limits of uniform quadrangulations with a boundary](#)”. arXiv: [1608.01129](#) (cit. on pp. 2967, 2982).
- G. Ben Arous and J.-D. Deuschel (1996). “[The construction of the  \$\(d + 1\)\$ -dimensional Gaussian droplet](#)”. *Comm. Math. Phys.* 179.2, pp. 467–488. MR: [1400748](#) (cit. on p. 2964).
- Jérémie Bettinelli and Grégory Miermont (2017). “[Compact Brownian surfaces I: Brownian disks](#)”. *Probab. Theory Related Fields* 167.3-4, pp. 555–614. MR: [3627425](#) (cit. on pp. 2967, 2982, 2985).
- E. Brézin, C. Itzykson, G. Parisi, and J. B. Zuber (1978). “[Planar diagrams](#)”. *Comm. Math. Phys.* 59.1, pp. 35–51. MR: [0471676](#) (cit. on p. 2966).
- Philippe Chassaing and Gilles Schaeffer (2004). “[Random planar lattices and integrated superBrownian excursion](#)”. *Probab. Theory Related Fields* 128.2, pp. 161–212. MR: [2031225](#) (cit. on p. 2966).
- Robert Cori and Bernard Vauquelin (1981). “[Planar maps are well labeled trees](#)”. *Canad. J. Math.* 33.5, pp. 1023–1042. MR: [638363](#) (cit. on pp. 2966, 2981).
- J. Theodore Cox and Richard Durrett (1981). “[Some limit theorems for percolation processes with necessary and sufficient conditions](#)”. *Ann. Probab.* 9.4, pp. 583–603. MR: [624685](#) (cit. on p. 2975).
- Nicolas Curien and Jean-François Le Gall (2015). “[First-passage percolation and local modifications of distances in random triangulations](#)”. arXiv: [1511.04264](#) (cit. on p. 2976).
- Nicolas Curien and Igor Kortchemski (2014). “[Random stable looptrees](#)”. *Electron. J. Probab.* 19, no. 108, 35. MR: [3286462](#) (cit. on p. 2982).
- Nicolas Curien and Jean-François Le Gall (2014). “[The Brownian plane](#)”. *J. Theoret. Probab.* 27.4, pp. 1249–1291. MR: [3278940](#) (cit. on pp. 2966, 2982).
- François David, Antti Kupiainen, Rémi Rhodes, and Vincent Vargas (2016). “[Liouville quantum gravity on the Riemann sphere](#)”. *Comm. Math. Phys.* 342.3, pp. 869–907. MR: [3465434](#) (cit. on pp. 2967, 2971).
- François David, Rémi Rhodes, and Vincent Vargas (2016). “[Liouville quantum gravity on complex tori](#)”. *J. Math. Phys.* 57.2, pp. 022302, 25. MR: [3450564](#) (cit. on p. 2971).
- P. Di Francesco, P. Ginsparg, and J. Zinn-Justin (1995). “[2D gravity and random matrices](#)”. *Phys. Rep.* 254.1-2, p. 133. MR: [1320471](#) (cit. on p. 2966).



- Julien Dubédat (2009). “[SLE and the free field: partition functions and couplings](#)”. *J. Amer. Math. Soc.* 22.4, pp. 995–1054. MR: [2525778](#) (cit. on p. 2964).
- Bertrand Duplantier (1998). “[Random walks and quantum gravity in two dimensions](#)”. *Phys. Rev. Lett.* 81.25, pp. 5489–5492. MR: [1666816](#) (cit. on p. 2974).
- (1999a). “[Harmonic measure exponents for two-dimensional percolation](#)”. *Phys. Rev. Lett.* 82.20, pp. 3940–3943. MR: [1688869](#) (cit. on p. 2974).
  - (1999b). “Two-dimensional copolymers and exact conformal multifractality”. *Physical review letters* 82.5, pp. 880–883 (cit. on p. 2974).
  - (2000). “[Conformally invariant fractals and potential theory](#)”. *Phys. Rev. Lett.* 84.7, pp. 1363–1367. MR: [1740371](#) (cit. on p. 2974).
- Bertrand Duplantier and Ivan Kostov (1988). “[Conformal spectra of polymers on a random surface](#)”. *Phys. Rev. Lett.* 61.13, pp. 1433–1437. MR: [960093](#) (cit. on p. 2984).
- Bertrand Duplantier and Ivan K. Kostov (1990). “[Geometrical critical phenomena on a random surface of arbitrary genus](#)”. *Nuclear Phys. B* 340.2-3, pp. 491–541. MR: [1068092](#) (cit. on p. 2984).
- Bertrand Duplantier, J. Miller, and S. Sheffield (2014). “Liouville quantum gravity, as a mating of trees” (cit. on pp. [2964](#), [2968](#), [2974](#), [2979](#), [2984](#)).
- Bertrand Duplantier and Scott Sheffield (2011). “[Liouville quantum gravity and KPZ](#)”. *Invent. Math.* 185.2, pp. 333–393. MR: [2819163](#) (cit. on pp. [2965](#), [2967](#), [2968](#), [2973](#)).
- E. B. Dynkin (1984a). “[Gaussian and non-Gaussian random fields associated with Markov processes](#)”. *J. Funct. Anal.* 55.3, pp. 344–376. MR: [734803](#) (cit. on p. 2964).
- (1984b). “Local times and quantum fields”. In: *Seminar on stochastic processes, 1983 (Gainesville, Fla., 1983)*. Vol. 7. Progr. Probab. Statist. Birkhäuser Boston, Boston, MA, pp. 69–83. MR: [902412](#) (cit. on p. 2964).
- Murray Eden (1961). “A two-dimensional growth process”. In: *Proc. 4th Berkeley Sympos. Math. Statist. and Prob., Vol. IV*. Univ. California Press, Berkeley, Calif., pp. 223–239. MR: [0136460](#) (cit. on p. 2975).
- J.-F. Le Gall (2017). “Brownian disks and the Brownian snake” (cit. on pp. [2982](#), [2984](#)).
- Giambattista Giacomin, Stefano Olla, and Herbert Spohn (2001). “[Equilibrium fluctuations for  \$\nabla\phi\$  interface model](#)”. *Ann. Probab.* 29.3, pp. 1138–1172. MR: [1872740](#) (cit. on p. 2964).
- Colin Guillarmou, Rémi Rhodes, and Vincent Vargas (2016). “[Polyakov’s formulation of 2d bosonic string theory](#)”. arXiv: [1607.08467](#) (cit. on p. 2971).
- Ewain Gwynne, Adrien Kassel, Jason Miller, and David B Wilson (2016). “[Active spanning trees with bending energy on planar maps and SLE-decorated Liouville quantum gravity for  \$\kappa \geq 8\$](#) ”. arXiv: [1603.09722](#) (cit. on p. 2984).
- Ewain Gwynne and Jason Miller (2016a). “[Convergence of the self-avoiding walk on random quadrangulations to  \$\text{SLE}\_{8/3}\$  on  \$\sqrt{8/3}\$ -Liouville quantum gravity](#)”. arXiv: [1608.00956](#) (cit. on p. 2985).

- (2016b). “Metric gluing of Brownian and  $\sqrt{8/3}$ -Liouville quantum gravity surfaces”. arXiv: 1608.00955 (cit. on p. 2985).
  - (2017a). “Characterizations of  $\text{SLE}_\kappa$  for  $\kappa \in (4, 8)$  on Liouville quantum gravity”. arXiv: 1701.05174 (cit. on p. 2986).
  - (2017b). “Convergence of percolation on uniform quadrangulations with boundary to  $\text{SLE}_6$  on  $\sqrt{8/3}$ -Liouville quantum gravity”. arXiv: 1701.05175 (cit. on pp. 2985, 2986).
  - (2017c). “Convergence of the free Boltzmann quadrangulation with simple boundary to the Brownian disk”. arXiv: 1701.05173 (cit. on pp. 2967, 2985).
  - (2017d). “Scaling limit of the uniform infinite half-plane quadrangulation in the Gromov-Hausdorff-Prokhorov-uniform topology”. *Electron. J. Probab.* 22, Paper No. 84, 47. MR: 3718712 (cit. on pp. 2967, 2982, 2984, 2985).
- Ewain Gwynne, Jason Miller, and Scott Sheffield (2017). “The Tutte embedding of the mated-CRT map converges to Liouville quantum gravity”. arXiv: 1705.11161 (cit. on p. 2967).
- J. M. Hammersley and D. J. A. Welsh (1965). “First-passage percolation, subadditive processes, stochastic networks, and generalized renewal theory”. In: *Proc. Internat. Res. Semin., Statist. Lab., Univ. California, Berkeley, Calif.* Springer-Verlag, New York, pp. 61–110. MR: 0198576 (cit. on p. 2975).
- Raphael Höegh-Krohn (1971). “A general class of quantum fields without cut-offs in two space-time dimensions”. *Comm. Math. Phys.* 21, pp. 244–255. MR: 0292433 (cit. on p. 2965).
- Yichao Huang, Rémi Rhodes, and Vincent Vargas (2015). “Liouville Quantum Gravity on the unit disk”. arXiv: 1502.04343 (cit. on p. 2971).
- Jean-Pierre Kahane (1985). “Sur le chaos multiplicatif”. *Ann. Sci. Math. Québec* 9.2, pp. 105–150. MR: 829798 (cit. on p. 2965).
- Richard Kenyon (2000). “Conformal invariance of domino tiling”. *Ann. Probab.* 28.2, pp. 759–795. MR: 1782431 (cit. on p. 2964).
- Richard Kenyon, Jason Miller, Scott Sheffield, and David B Wilson (2015). “Bipolar orientations on planar maps and  $\text{SLE}_{12}$ ”. arXiv: 1511.04068 (cit. on pp. 2972, 2984).
- (2017). “Six-vertex model and Schramm-Loewner evolution”. *Physical Review E* 95.5, p. 052146 (cit. on p. 2972).
- V. G. Knizhnik, A. M. Polyakov, and A. B. Zamolodchikov (1988). “Fractal structure of 2D-quantum gravity”. *Modern Phys. Lett. A* 3.8, pp. 819–826. MR: 947880 (cit. on pp. 2964, 2984).
- Gregory F. Lawler, Oded Schramm, and Wendelin Werner (2004a). “Conformal invariance of planar loop-erased random walks and uniform spanning trees”. *Ann. Probab.* 32.1B, pp. 939–995. MR: 2044671 (cit. on p. 2972).

- Gregory F. Lawler, Oded Schramm, and Wendelin Werner (2004b). “On the scaling limit of planar self-avoiding walk”. In: *Fractal geometry and applications: a jubilee of Benoît Mandelbrot, Part 2*. Vol. 72. Proc. Sympos. Pure Math. Amer. Math. Soc., Providence, RI, pp. 339–364. MR: [2112127](#) (cit. on p. [2972](#)).
- Jean-François Le Gall (2007). “The topological structure of scaling limits of large planar maps”. *Invent. Math.* 169.3, pp. 621–670. MR: [2336042](#) (cit. on p. [2966](#)).
- (2013). “Uniqueness and universality of the Brownian map”. *Ann. Probab.* 41.4, pp. 2880–2960. MR: [3112934](#) (cit. on p. [2966](#)).
  - (2014). “Random geometry on the sphere”. In: *Proceedings of the International Congress of Mathematicians—Seoul 2014. Vol. 1*. Kyung Moon Sa, Seoul, pp. 421–442. MR: [3728478](#) (cit. on p. [2966](#)).
- Jean-François Le Gall and Frédéric Paulin (2008). “Scaling limits of bipartite planar maps are homeomorphic to the 2-sphere”. *Geom. Funct. Anal.* 18.3, pp. 893–918. MR: [2438999](#) (cit. on pp. [2966](#), [2981](#)).
- Yves Le Jan (2011). *Markov paths, loops and fields*. Vol. 2026. Lecture Notes in Mathematics. Lectures from the 38th Probability Summer School held in Saint-Flour, 2008, École d’Été de Probabilités de Saint-Flour. [Saint-Flour Probability Summer School]. Springer, Heidelberg, pp. viii+124. MR: [2815763](#) (cit. on p. [2964](#)).
- Yiting Li, Xin Sun, and Samuel S Watson (2017). “Schnyder woods, SLE(16), and Liouville quantum gravity”. arXiv: [1705.03573](#) (cit. on pp. [2972](#), [2984](#)).
- Jean-François Marckert and Abdelkader Mokkadem (2006). “Limit of normalized quadrangulations: the Brownian map”. *Ann. Probab.* 34.6, pp. 2144–2202. MR: [2294979](#) (cit. on p. [2966](#)).
- Grégory Miermont (2008). “On the sphericity of scaling limits of random planar quadrangulations”. *Electron. Commun. Probab.* 13, pp. 248–257. MR: [2399286](#) (cit. on pp. [2966](#), [2981](#)).
- (2013). “The Brownian map is the scaling limit of uniform random plane quadrangulations”. *Acta Math.* 210.2, pp. 319–401. MR: [3070569](#) (cit. on p. [2966](#)).
- Jason Miller (2011). “Fluctuations for the Ginzburg-Landau  $\nabla\phi$  interface model on a bounded domain”. *Comm. Math. Phys.* 308.3, pp. 591–639. MR: [2855536](#) (cit. on p. [2964](#)).
- Jason Miller and Scott Sheffield (2015a). “An axiomatic characterization of the Brownian map”. arXiv: [1506.03806](#) (cit. on pp. [2967](#), [2979](#), [2983](#)).
- (2015b). “Liouville quantum gravity and the Brownian map I: The QLE (8/3, 0) metric”. arXiv: [1507.00719](#) (cit. on pp. [2967](#), [2975](#), [2979](#)).
  - (2015c). “Liouville quantum gravity spheres as matings of finite-diameter trees”. arXiv: [1506.03804](#) (cit. on pp. [2967](#), [2974](#)).
  - (2016a). “Imaginary geometry I: interacting SLEs”. *Probab. Theory Related Fields* 164.3–4, pp. 553–705. MR: [3477777](#) (cit. on p. [2964](#)).

- (2016b). “Imaginary geometry II: reversibility of  $SLE_\kappa(\rho_1; \rho_2)$  for  $\kappa \in (0, 4)$ ”. *Ann. Probab.* 44.3, pp. 1647–1722. MR: [3502592](#) (cit. on p. 2964).
- (2016c). “Imaginary geometry III: reversibility of  $SLE_\kappa$  for  $\kappa \in (4, 8)$ ”. *Ann. of Math.* (2) 184.2, pp. 455–486. MR: [3548530](#) (cit. on p. 2964).
- (2016d). “Liouville quantum gravity and the Brownian map II: geodesics and continuity of the embedding”. arXiv: [1605.03563](#) (cit. on pp. 2967, 2975, 2979, 2983, 2984).
- (2016e). “Liouville quantum gravity and the Brownian map III: the conformal structure is determined”. arXiv: [1608.05391](#) (cit. on pp. 2967, 2975, 2979, 2984).
- (2016f). “Quantum Loewner evolution”. *Duke Math. J.* 165.17, pp. 3241–3378. MR: [3572845](#) (cit. on p. 2980).
- (2017). “Imaginary geometry IV: interior rays, whole-plane reversibility, and space-filling trees”. *Probab. Theory Related Fields* 169.3–4, pp. 729–869. MR: [3719057](#) (cit. on pp. 2964, 2979).
- R. C. Mullin (1967). “On the enumeration of tree-rooted maps”. *Canad. J. Math.* 19, pp. 174–183. MR: [0205882](#) (cit. on p. 2966).
- L. Niemeyer, L. Pietronero, and H. J. Wiesmann (1984). “Fractal dimension of dielectric breakdown”. *Phys. Rev. Lett.* 52.12, pp. 1033–1036. MR: [736820](#) (cit. on p. 2980).
- A. Polyakov (1990). “Two-dimensional quantum gravity. Superconductivity at high  $T_C$ ”. In: *Champs, cordes et phénomènes critiques (Les Houches, 1988)*. North-Holland, Amsterdam, pp. 305–368. MR: [1052937](#) (cit. on p. 2964).
- A. M. Polyakov (1981a). “Quantum geometry of bosonic strings”. *Phys. Lett. B* 103.3, pp. 207–210. MR: [623209](#) (cit. on p. 2964).
- (1981b). “Quantum geometry of fermionic strings”. *Phys. Lett. B* 103.3, pp. 211–213. MR: [623210](#) (cit. on p. 2964).
- Rémi Rhodes and Vincent Vargas (2014). “Gaussian multiplicative chaos and applications: a review”. *Probab. Surv.* 11, pp. 315–392. MR: [3274356](#) (cit. on p. 2965).
- Brian Rider and Bálint Virág (2007). “The noise in the circular law and the Gaussian free field”. *Int. Math. Res. Not. IMRN* 2, Art. ID rnm006, 33. MR: [2361453](#) (cit. on p. 2964).
- Steffen Rohde and Oded Schramm (2005). “Basic properties of SLE”. *Ann. of Math.* (2) 161.2, pp. 883–924. MR: [2153402](#) (cit. on p. 2972).
- Gilles Schaeffer (1998). “Conjugaison d’arbres et cartes combinatoires aléatoires”. PhD thesis. Bordeaux 1 (cit. on pp. 2966, 2981).
- Oded Schramm (2000). “Scaling limits of loop-erased random walks and uniform spanning trees”. *Israel J. Math.* 118, pp. 221–288. MR: [1776084](#) (cit. on pp. 2964, 2971).
- Oded Schramm and Scott Sheffield (2009). “Contour lines of the two-dimensional discrete Gaussian free field”. *Acta Math.* 202.1, pp. 21–137. MR: [2486487](#) (cit. on pp. 2964, 2972).
- (2013). “A contour line of the continuum Gaussian free field”. *Probab. Theory Related Fields* 157.1–2, pp. 47–80. MR: [3101840](#) (cit. on pp. 2964, 2972).

- Scott Sheffield (2016a). “Conformal weldings of random surfaces: SLE and the quantum gravity zipper”. *Ann. Probab.* 44.5, pp. 3474–3545. MR: [3551203](#) (cit. on pp. [2964](#), [2968](#), [2974](#), [2985](#)).
- (2016b). “Quantum gravity and inventory accumulation”. *Ann. Probab.* 44.6, pp. 3804–3848. MR: [3572324](#) (cit. on p. [2984](#)).
- Stanislav Smirnov (2001). “Critical percolation in the plane: conformal invariance, Cardy’s formula, scaling limits”. *C. R. Acad. Sci. Paris Sér. I Math.* 333.3, pp. 239–244. MR: [1851632](#) (cit. on p. [2972](#)).
- (2010). “Conformal invariance in random cluster models. I. Holomorphic fermions in the Ising model”. *Ann. of Math. (2)* 172.2, pp. 1435–1467. MR: [2680496](#) (cit. on p. [2972](#)).
- W. T. Tutte (1962). “A census of planar triangulations”. *Canad. J. Math.* 14, pp. 21–38. MR: [0130841](#) (cit. on p. [2966](#)).
- Mohammad Q. Vahidi-Asl and John C. Wierman (1990). “First-passage percolation on the Voronoï tessellation and Delaunay triangulation”. In: *Random graphs ’87 (Poznań, 1987)*. Wiley, Chichester, pp. 341–359. MR: [1094141](#) (cit. on pp. [2975](#), [2976](#)).
- (1992). “A shape result for first-passage percolation on the Voronoï tessellation and Delaunay triangulation”. In: *Random graphs, Vol. 2 (Poznań, 1989)*. Wiley-Intersci. Publ. Wiley, New York, pp. 247–262. MR: [1166620](#) (cit. on pp. [2975](#), [2976](#)).

Received 2017-12-05.

JASON MILLER

[jpmiller@statslab.cam.ac.uk](mailto:jpmiller@statslab.cam.ac.uk)

# MEAN FIELD ASYMPTOTICS IN HIGH-DIMENSIONAL STATISTICS: FROM EXACT RESULTS TO EFFICIENT ALGORITHMS

ANDREA MONTANARI

## Abstract

Modern data analysis challenges require building complex statistical models with massive numbers of parameters. It is nowadays commonplace to learn models with millions of parameters by using iterative optimization algorithms. What are typical properties of the estimated models? In some cases, the high-dimensional limit of a statistical estimator is analogous to the thermodynamic limit of a certain (disordered) statistical mechanics system. Building on mathematical ideas from the mean-field theory of disordered systems, exact asymptotics can be computed for high-dimensional statistical learning problems.

This theory suggests new practical algorithms and new procedures for statistical inference. Also, it leads to intriguing conjectures about the fundamental computational limits for statistical estimation.

## 1 Introduction

Natural and social sciences as well as engineering disciplines are nowadays blessed with abundant data which are used to construct ever more complex statistical models. This scenario requires new methodologies and new mathematical techniques to analyze these methods. In this article I will briefly overview some recent progress on two prototypical problems in this research area: high-dimensional regression and principal component analysis. This overview will be far from exhaustive, and will follow a viewpoint that builds on connections with mean field theory in mathematical physics and probability theory (see [Section 5](#) for further context).

**High-dimensional regression.** We are given data points  $(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)$  that are independent draws from a common (unknown) distribution. Here  $\mathbf{x}_i \in \mathbb{R}^d$  is a feature

vector (or vector of covariates), and  $y_i \in \mathbb{R}$  is a label or response variable. We would like to model the dependency of the response variable upon the feature vector as

$$(1-1) \quad y_i = \langle \boldsymbol{\theta}_0, \mathbf{x}_i \rangle + w_i,$$

where  $\boldsymbol{\theta}_0 \in \mathbb{R}^d$  is a vector of parameters (coefficients), and  $w_i$  captures non-linear dependence as well as random effects. This simple linear model (and its variants) has an impressive number of applications ranging from genomics [Shevade and Keerthi \[2003\]](#), to online commerce [McMahan et al. \[2013\]](#), to signal processing [D. L. Donoho \[2006\]](#) and [Candès, Romberg, and Tao \[2006\]](#).

**Principal component analysis.** We are given unlabeled data  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$ , that are i.i.d. with zero mean and common covariance  $\Sigma \equiv \mathbb{E}\{\mathbf{x}_1 \mathbf{x}_1^T\}$ . We would like to estimate the directions of maximal variability of these data. Namely, denoting by  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$  the ordered eigenvalues of  $\Sigma$  and by  $\mathbf{v}_1(\Sigma), \dots, \mathbf{v}_n(\Sigma)$  the correspondent eigenvectors, we would like to estimate  $\mathbf{v}_1(\Sigma), \dots, \mathbf{v}_k(\Sigma)$  for  $k \ll d$  a fixed number. This task is a fundamental component of dimensionality reduction and clustering [Kannan, S. Vempala, and Vetta \[2004\]](#), and is often used in neuroscience [Rossant et al. \[2016\]](#) and genomics [Abraham and Inouye \[2014\]](#).

## 2 High-dimensional regression

Since [Gauss \[2011\]](#), least squares has been the method of choice for estimating the parameter vector  $\boldsymbol{\theta}_0$  in the linear model (1-1). Least squares does not make assumptions on the coefficients  $\boldsymbol{\theta}_0$ , but implicitly assumes the errors  $w_i$  to be unbiased and all of roughly the same magnitude. In this is the case (for ‘non-degenerate’ features  $\mathbf{x}_i$ ), consistent estimation is possible if and only if  $n/p \gg 1$ .

In contrast, many modern applications are characterized by a large amount of data, together with extremely complex models. In other words, both  $n$  and  $p$  are large and often comparable. The prototypical approach to this regime is provided by the following  $\ell_1$  regularized least squares problem, known as the Lasso [Tibshirani \[1996\]](#) or basis pursuit denoising [Chen and D. L. Donoho \[1995\]](#):

$$(2-1) \quad \hat{\boldsymbol{\theta}}(\lambda; \mathbf{y}, \mathbf{X}) \equiv \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2 + \lambda \|\boldsymbol{\theta}\|_1 \right\}.$$

Here  $\mathbf{y} = (y_1, \dots, y_n) \in \mathbb{R}^n$  is the vector of response variables, and  $\mathbf{X} \in \mathbb{R}^{n \times p}$  is the matrix whose  $i$ -th row is the  $i$ -th feature vector  $\mathbf{x}_i$ . Since the problem (2-1) is convex (and of a particularly simple form) it can be solved efficiently. In the following, we will drop the dependence of  $\hat{\boldsymbol{\theta}}$  upon  $\mathbf{y}, \mathbf{X}$  unless needed for clarity.

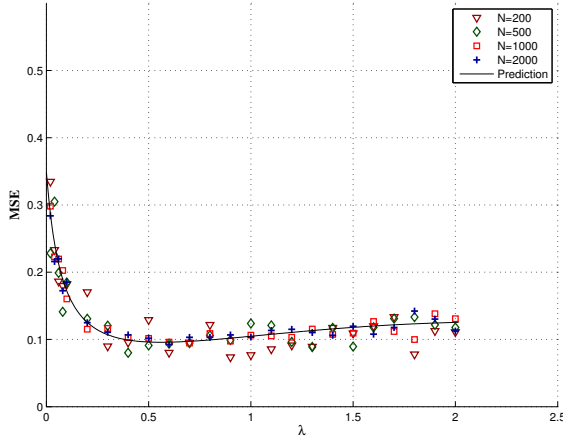


Figure 1: Mean square estimation error of the Lasso per dimension  $\|\hat{\theta}(\lambda) - \theta_0\|_2^2/d$ , as a function of the regularization parameter  $\lambda$ . Each point corresponds to a different instance of the problem (2-1) with symbols representing the dimension  $d = N \in \{100, 500, 1000, 2000\}$ . The number of samples is  $n = d\delta$ , with  $\delta = 0.64$ , and the noise level  $\sigma^2 = 0.2 \cdot n$ . The ‘true’ coefficients were generated with i.i.d. coordinates  $\theta_{0,i} \in \{0, +1, -1\}$  and  $\mathbb{P}(\theta_{0,i} = +1) = \mathbb{P}(\theta_{0,i} = -1) = 0.064$ .

Over the last ten years, a sequence of beautiful works [Candes and Tao \[2007\]](#) and [Bickel, Ritov, and Tsybakov \[2009\]](#) has developed order-optimal bounds on the performance of the Lasso estimator. Analysis typically assumes that the data are generated according to model (1-1), with some vector  $\theta_0$ , and i.i.d. noise  $(w_i)_{i \leq n}$ : to be concrete we will assume here  $w_i \sim \mathcal{N}(0, \sigma^2)$ . For instance, if  $\theta_0$  has at most  $s_0$  non-zero elements, and under suitable conditions on the matrix  $X$ , it is known that (with high probability)

$$(2-2) \quad \|\hat{\theta}(\lambda) - \theta_0\|_2^2 \leq \frac{C s_0 \sigma^2}{n} \log d ,$$

where  $C$  is a numerical constant.

This type of results give confidence in the use of the Lasso, and explain the origins of its effectiveness. However, they are not precise enough to compare different estimators with the same error rate or –say– different ways of selecting the regularization parameter  $\lambda$ . Also, they provide limited insight on the distribution of  $\hat{\theta}(\lambda)$ , an issue that is crucial for statistical inference.



**2.1 Exact asymptotics for the Lasso.** In order to address these questions, a different type of analysis makes probabilistic assumptions about the feature vectors  $\mathbf{x}_i$ , and derives an asymptotically exact characterization of the high-dimensional estimator. In order to state a result of this type for the case of the Lasso, it is useful to introduce the proximal operator of the  $\ell_1$  norm (in one dimension):

$$(2-3) \quad \eta(y; \alpha) \equiv \arg \min_{x \in \mathbb{R}} \left\{ \frac{1}{2} (y - x)^2 + \alpha |x| \right\}.$$

Explicitly, we have  $\eta(y; \alpha) = (|y| - \alpha) \text{sign}(y)$ . We also note that the following simple consequence of the first-order stationarity conditions for problem (2-1) holds for any  $\alpha > 0$ :

$$(2-4) \quad \hat{\boldsymbol{\theta}}(\lambda) = \eta(\hat{\boldsymbol{\theta}}^d; \alpha), \quad \hat{\boldsymbol{\theta}}^d(\alpha, \lambda) \equiv \hat{\boldsymbol{\theta}}(\lambda) + \frac{\alpha}{n\lambda} \mathbf{X}^\top (\mathbf{y} - \mathbf{X} \hat{\boldsymbol{\theta}}(\lambda)).$$

We say that a function  $\psi : \mathbb{R}^d \rightarrow \mathbb{R}$  is pseudo-Lipschitz function of order  $k$  (and write  $\psi \in \text{PL}(k)$ ) if  $|\psi(\mathbf{x}) - \psi(\mathbf{y})| \leq L(1 + (\|\mathbf{x}\|_2/\sqrt{d})^{k-1} + (\|\mathbf{y}\|_2/\sqrt{d})^{k-1})\|\mathbf{x} - \mathbf{y}\|_2$  for any  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ . Also recall that a sequence of probability distributions  $\nu_n$  on  $\mathbb{R}^d$  converges in Wasserstein- $k$  distance to  $\nu$  if and only if  $\int \psi(\mathbf{x}) \nu_n(d\mathbf{x}) \rightarrow \int \psi(\mathbf{x}) \nu(d\mathbf{x})$  for each  $\psi \in \text{PL}(k)$ .

**Theorem 1.** *Consider a sequence of linear models (1-1) indexed by  $n$ , with  $d = d(n)$  such that  $\lim_{n \rightarrow \infty} n/d(n) = \delta \in (0, \infty)$ , and let  $\sigma = \sigma(n)$  be such that  $\lim_{n \rightarrow \infty} \sigma(n)/\sqrt{n} = \sigma_0$ . Assume  $\mathbf{x}_i \sim \mathcal{N}(0, \mathbf{I}_d)$  and  $w_i \sim \mathcal{N}(0, \sigma^2)$  independent and that the empirical distribution  $d^{-1} \sum_{i=1}^d \delta_{\theta_{0,i}}$  converges in  $W_k$  to the law  $p_\Theta$  of a random variable  $\Theta$ .*

*Let  $\alpha_*, \tau_*^2 \in \mathbb{R}_{>0}$  be the unique solution of the pair of equations*

$$(2-5) \quad \lambda = \alpha \left\{ 1 - \frac{1}{\delta} \mathbb{P}(|\Theta_* + \tau_* Z| \geq \alpha_*) \right\},$$

$$(2-6) \quad \tau_*^2 = \sigma_0^2 + \frac{1}{\delta} \mathbb{E} \{ [\eta(\Theta + \tau_* Z; \alpha_*) - \Theta]^2 \},$$

where expectation is with respect to  $\Theta$  and  $Z \sim \mathcal{N}(0, 1)$  independent. Then, taking  $\hat{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}^d(\alpha, \lambda)$ , for any  $\psi : \mathbb{R}^2 \rightarrow \mathbb{R}$ ,  $\psi \in \text{PL}(k)$ , we have, almost surely,

$$(2-7) \quad \lim_{n, d \rightarrow \infty} \frac{1}{d} \sum_{i=1}^d \psi(\theta_{0,i}, \hat{\theta}_i^d) = \mathbb{E} \{ \psi(\Theta, \Theta + \tau_* Z) \}.$$

The proof of this result [Bayati and Montanari \[2012\]](#) consists in introducing an iterative algorithm that converges rapidly to  $\hat{\boldsymbol{\theta}}(\lambda)$  and can be analyzed exactly. Of course, the

existence of such an algorithm is of independent interest, cf. [Section 4](#). Alternative proof techniques have been developed as well, and are briefly mentioned in the next section. All of these proofs take advantage in a crucial way of the fact that the optimization problem (2-1) is convex, which in turn is a choice dictated by computational tractability. However, for  $\delta < 1$ , the cost function is not strongly convex (since the kernel of  $\mathbf{X}$  has dimension  $n(1 - \delta)$ , with high probability), which poses interesting challenges.

**Remark 2.1.** A first obvious use of [Theorem 1](#) is to derive asymptotic expressions for the risk of the Lasso. Using the stationarity condition (2-4) and choosing  $\psi(x, y) = [x - \eta(y, \alpha_*)]^2$ , we obtain

$$(2-8) \quad \lim_{n, p \rightarrow \infty} \frac{1}{d} \|\hat{\boldsymbol{\theta}}(\lambda) - \boldsymbol{\theta}_0\|_2^2 = \mathbb{E} \{ [\eta(\Theta + \tau_* Z; \alpha_*) - \Theta]^2 \}.$$

For applications, this prediction has the disadvantage of depending on the asymptotic empirical distribution of the entries of  $\boldsymbol{\theta}_0$ , which is not known. One possible way to overcome this problem is to consider the worst case distribution [D. L. Donoho, Maleki, and Montanari \[2011\]](#). Assuming  $\boldsymbol{\theta}_0$  has at most  $s_0 = p\varepsilon$  non-zero entries (and under the same assumptions of the last theorem), this results in the bound

$$(2-9) \quad \lim_{n, p \rightarrow \infty} \frac{1}{d} \|\hat{\boldsymbol{\theta}}(\lambda) - \boldsymbol{\theta}_0\|_2^2 \leq \frac{M(\varepsilon)}{1 - M(\varepsilon)/\delta} \sigma_0^2.$$

Where  $M(\varepsilon)$  is explicitly given in [D. L. Donoho, Maleki, and Montanari \[2011\]](#) and [Montanari \[2012\]](#), and behaves as  $M(\varepsilon) = 2\varepsilon \log(1/\varepsilon) + O(\varepsilon)$  for small  $\varepsilon$ . This bound is tight in the sense that there exists sequences of vectors  $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_0(n)$  for which the bound holds with equality.

**Remark 2.2.** Interestingly, [Theorem 1](#) also characterizes the joint distribution of  $\hat{\boldsymbol{\theta}}^d$  and the true parameter vector. Namely  $\hat{\theta}_i^d$  is asymptotically Gaussian, with mean equal to the true parameter  $\theta_{0,i}$  and variance  $\tau_*^2$ . This is somewhat surprising, given that the Lasso estimator  $\hat{\boldsymbol{\theta}} = \eta(\hat{\boldsymbol{\theta}}^d; \alpha_*)$  is highly non-Gaussian (in particular is  $\hat{\theta}_i = 0$  for a positive fraction of the entries).

This Gaussian limit suggests a possible approach to statistical inference. In particular, a confidence interval for  $\theta_{0,i}$  can be constructed by letting  $J_i(c) = [\hat{\theta}_i^d - c\tau_*, \hat{\theta}_i^d + c\tau_*]$ . The above theorem implies the following coverage guarantee

$$(2-10) \quad \lim_{n, d \rightarrow \infty} \frac{1}{d} \sum_{i=1}^n \mathbb{P}(\theta_{0,i} \in J_i(c)) = 1 - 2\Phi(-c),$$

where  $\Phi(x) \equiv \int e^{-t^2/2} dt / \sqrt{2\pi}$  is the Gaussian distribution function. In other words, the confidence interval is valid on average [Javanmard and Montanari \[2014b\]](#).

Ideally, one would like stronger guarantees than in (2-10), for instance ensuring coverage for each coordinate, rather than on average over coordinates. Results of this type were proven in [C.-H. Zhang and S. S. Zhang \[2014\]](#), [van de Geer, Bühlmann, Ritov, and Dezeure \[2014\]](#), and [Javanmard and Montanari \[2014a, 2015\]](#) (these papers however do not address the regime  $n/d \rightarrow \delta \in (0, \infty)$ ).

**Remark 2.3.** [Theorem 1](#) assumes the entries of the design matrix  $\mathbf{X}$  to be i.i.d. standard Gaussian. It is expected this result to enjoy some degree of universality, for instance with respect to matrices with i.i.d. entries with the same first two moments and sufficiently light tails. Universality results were proven in [Korada and Montanari \[2011\]](#), [Bayati, Lelarge, and Montanari \[2015\]](#), and [Oymak and Tropp \[2015\]](#), mainly focusing on the noiseless case  $\sigma = 0$  which is addressed by solving the problem (2-1) in the limit  $\lambda \rightarrow 0$  (equivalently, finding the solution of  $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}$  that minimizes  $\|\boldsymbol{\theta}\|_1$ ). Classical tools of probability theory, in particular the moment method and Lindeberg swapping trick are successfully applied in this case.

Beyond matrices with i.i.d. entries, there is empirical evidence [D. Donoho and Tanner \[2009\]](#) and heuristic results [Tulino, Caire, Verdu, and Shamai \[2013\]](#) and [Javanmard and Montanari \[2014b\]](#) suggesting universality or (in some cases) generalizations of the prediction of [Theorem 1](#).

**2.2 Generalizations and comparisons.** When the data  $(y_i, \mathbf{x}_i)$ ,  $1 \leq i \leq n$  contain outliers, the sum of square residuals  $\|\mathbf{y} - \mathbf{X}\boldsymbol{\theta}\|_2^2$  in [Equation \(2-1\)](#) is overly influenced by such outliers resulting in poor estimates. Robust regression [Huber and Ronchetti \[2009\]](#) suggests to use the following estimator instead (focusing for simplicity on the un-regularized case):

$$(2-11) \quad \hat{\boldsymbol{\theta}} = \arg \min_{\boldsymbol{\theta} \in \mathbb{R}^p} \sum_{i=1}^n \rho(y_i - \langle \boldsymbol{\theta}, \mathbf{x}_i \rangle),$$

where  $\rho : \mathbb{R} \rightarrow \mathbb{R}$  is often chosen to be convex in order to ensure computational tractability. For instance, [Huber \[1964, 1973\]](#) advocated the use of  $\rho(x) = \rho_{\text{Huber}}(x; c)$  defined by  $\rho_{\text{Huber}}(x; c) = x^2/2$  for  $|x| \leq c$  and  $\rho_{\text{Huber}}(x; c) = c|x| - c^2/2$  otherwise. Results analogous to [Theorem 1](#) were proven for robust estimators of the form (2-11) in [Karoui \[2013\]](#) and [D. Donoho and Montanari \[2016\]](#), following earlier conjectures in [El Karoui, Bean, Bickel, Lim, and Yu \[2013\]](#).

A second possibility for generalizing [Theorem 1](#) is to modify the penalty function  $\lambda\|\boldsymbol{\theta}\|_1$ , and replacing it by  $f(\boldsymbol{\theta})$  for  $f : \mathbb{R}^d \rightarrow \mathbb{R}$  a convex function. General results

in this setting were proven in [Chandrasekaran, Recht, Parrilo, and Willsky \[2012\]](#) and [Thrampoulidis, Oymak, and Hassibi \[2015\]](#) via a different approach that builds on Gordon's minimax theorem [Gordon \[1988\]](#).

Finally, let us emphasize that sparsity of  $\theta_0$ —while motivating the Lasso estimator (2-1)—does not play any role in [Theorem 1](#), which in fact holds for non-sparse  $\theta_0$  as well. Given this, it is natural to ask what is the best estimate for any given  $\theta_0$ . Under the assumption of [Theorem 1](#), it is natural to treat the  $(\theta_i)_{i \leq d}$  as i.i.d. draws with common distribution  $p_\Theta$ . If this is the case, we can consider the posterior expectation estimator

$$(2-12) \quad \hat{\theta}^{\text{Bayes}}(\mathbf{y}, \mathbf{X}) \equiv \mathbb{E}_{p_\Theta}(\theta | \mathbf{y}, \mathbf{X}).$$

The analysis of this estimator requires introducing two functions associated with the scalar problem of estimating  $\Theta \sim p_\Theta$  from observations  $Y = \sqrt{s} \Theta + Z$ ,  $Z \sim \mathcal{N}(0, 1)$ :

$$(2-13) \quad I(s) \equiv I(\Theta; \sqrt{s}\Theta + Z), \quad \text{mmse}(s) = \mathbb{E}\{[\Theta - \mathbb{E}(\Theta | \sqrt{s}\Theta + Z)]^2\},$$

These two quantities are intimately related since  $\frac{dI}{ds}(s) = \frac{1}{2} \text{mmse}(s)$  [Stam \[1959\]](#) and [Guo, Shamai, and Verdú \[2005\]](#). The following is a restatement of a theorem proved in [Reeves and Pfister \[2016\]](#),

**Theorem 2.** *Under the assumptions of [Theorem 1](#), define the function  $\tau^2 \mapsto \Psi(\tau^2)$  by*

$$(2-14) \quad \Psi(\tau^2) = I(\tau^{-2}) + \frac{\delta}{2} \left( \log(\delta \tau^2) - \frac{\delta}{2} + \frac{\delta \sigma_0^2}{2\tau^2} \right).$$

*If, for  $\sigma_0^2 > 0$ ,  $\tau^2 \mapsto \Psi(\tau^2)$  has at most three critical point and  $\tau_{\text{Bayes}}^2 \equiv \arg \min_{\tau^2 > 0} \Psi(\tau^2)$  is unique, then*

$$(2-15) \quad \lim_{n, d \rightarrow \infty} \frac{1}{d} \mathbb{E}\{\|\hat{\theta}(\mathbf{y}, \mathbf{X}) - \theta\|_2\} = \text{mmse}(\tau_{\text{Bayes}}^{-2}).$$

A substantial generalization of this theorem was proved recently in [Barbier, Macris, Dia, and Krzakala \[2017\]](#), encompassing in particular a class of generalized linear models.

Notice that  $\tau_{\text{Bayes}}$  must satisfy the following first-order stationarity condition (which is obtained by differentiating  $\Psi(\cdot)$ ):

$$(2-16) \quad \tau_{\text{Bayes}}^2 = \sigma^2 + \frac{1}{\delta} \text{mmse}(\tau_{\text{Bayes}}^{-2}).$$

The form of this equation is tantalizingly similar to the one for the Lasso mean square error, cf. [Equation \(2-6\)](#). In both case the right-hand side is given in terms of the error in

estimating the scalar  $\Theta \sim p_\Theta$  from noisy observations  $Y = \Theta + \tau Z$ . While Equation (2-6) corresponds to the error of proximal denoising using  $\ell_1$  norm, the Bayes estimation error appears in Equation (2-16).

**2.3 Decoupling.** A key property is shared by the Lasso and other convex estimators, as well as the Bayes-optimal estimators of Section 2.2. It will also hold for the message passing algorithms of Section 4 and it is sometimes referred to as ‘decoupling’. Notice that Equation (2-7) of Theorem 1 can be interpreted as follows. By Equation (2-4), we can use the estimate  $\hat{\theta}$  to construct new ‘pseudo-data’  $\hat{\theta}^d$  with the following remarkable property. Each coordinate of the pseudo-data  $\hat{\theta}_i^d$  is approximately distributed as a Gaussian noisy observation of the true parameter  $\theta_{0,i}$ .

This naturally raises the question of the joint distribution of  $k$  coordinates  $\hat{\theta}_{i(1)}^d, \dots, \hat{\theta}_{i(k)}^d$ . Decoupling occurs when these are approximately distributed as observations of  $\theta_{i(1)}, \dots, \theta_{i(k)}$  with independent noise. For instance, in the case of Theorem 1, this can be formalized as

$$(2-17) \quad \lim_{n,d \rightarrow \infty} \frac{1}{d^k} \sum_{i(1), \dots, i(k)=1}^d \psi(\theta_{0,i(1)}, \dots, \theta_{0,i(k)}; \hat{\theta}_{i(1)}^d, \dots, \hat{\theta}_{i(k)}^d) = \mathbb{E} \{ \psi(\Theta_1, \dots, \Theta_k; \Theta_1 + \tau_* Z_1, \dots, \Theta_k + \tau_* Z_k) \},$$

where  $\psi$  is a bounded continuous function and  $(\Theta_\ell)_{\ell \leq k} \sim_{iid} p_\Theta$  independent of  $(Z_\ell)_{\ell \leq k} \sim \mathcal{N}(0, 1)$ . In this form, decoupling is in fact an immediate consequence of Equation (2-7), but other forms of decoupling are proved in the literature. (And sometimes the model of interest has to be perturbed in order to obtain decoupling.)

### 3 Principal component analysis

A standard model for principal component analysis assumes that the vectors  $\mathbf{x}_1, \dots, \mathbf{x}_n \in \mathbb{R}^d$  are centered Gaussian, with covariance  $\Sigma = \theta_0 \theta_0^\top + \mathbf{I}_n$ , for  $\theta_0$  a fixed unknown vector. Equivalently, if we let  $\mathbf{X} \in \mathbb{R}^{n \times d}$  be the matrix whose  $i$ -th row is the vector  $\mathbf{x}_i$ , we have  $\mathbf{X} = \mathbf{u} \theta_0^\top + \mathbf{W}$ , where  $\mathbf{u} = (u_i)_{i \leq n}$  is a vector with i.i.d. entries  $u_i \sim \mathcal{N}(0, 1)$ , and  $(W_{ij})_{i \leq n, j \leq d} \sim \mathcal{N}(0, 1)$ .

For the sake of simplicity, we shall consider here the symmetric version of this model. The data consists in a symmetric matrix  $\mathbf{X} \in \mathbb{R}^{n \times n}$ , where

$$(3-1) \quad \mathbf{X} = \frac{\lambda}{n} \theta_0 \theta_0^\top + \mathbf{W},$$

where  $\mathbf{W}$  is a noise matrix from the  $\text{GOE}(n)$  ensemble, namely  $(W_{ij})_{i < j \leq n} \sim \mathcal{N}(0, 1/n)$  are independent of  $(W_{ii})_{i \leq n} \sim \mathcal{N}(0, 2/n)$ , and  $\mathbf{W} = \mathbf{W}^\top$ . We further assume  $\lambda \geq 0$  and  $\|\boldsymbol{\theta}_0\|_2^n/n \rightarrow 1$  as  $n \rightarrow \infty$ . This normalization is chosen to make the problem nontrivial when  $\lambda = \Theta(1)$ .

We are asked to estimate  $\boldsymbol{\theta}_0 \in \mathbb{R}^n$  from a single observation of the matrix  $\mathbf{X}$ . Spectral methods are –by far– the best studied approach to this problem, and the asymptotic spectral properties of  $\mathbf{X}$  have been studied in exquisite detail across probability theory and statistics [Baik, Ben Arous, and P     \[2005\]](#), [Baik and Silverstein \[2006\]](#), [F     and P     \[2007\]](#), [Johnstone \[2001\]](#), [Paul \[2007\]](#), [Capitaine, Donati-Martin, and F     \[2009\]](#), [Benaych-Georges and Nadakuditi \[2011, 2012\]](#), and [Knowles and Yin \[2013\]](#). In particular, letting  $\hat{\boldsymbol{\theta}}^{\text{PCA}}(\mathbf{X})$  denote the principal eigenvector of  $\mathbf{X}$ , we have

$$(3-2) \quad \lim_{n \rightarrow \infty} \frac{|\langle \hat{\boldsymbol{\theta}}^{\text{PCA}}(\mathbf{X}), \boldsymbol{\theta}_0 \rangle|}{\|\hat{\boldsymbol{\theta}}^{\text{PCA}}(\mathbf{X})\|_2 \|\boldsymbol{\theta}_0\|} = \begin{cases} 0 & \text{if } \lambda \leq 1, \\ \sqrt{1 - \lambda^{-2}} & \text{if } \lambda > 1. \end{cases}$$

In other words, the spectral estimator achieves a positive correlation with the unknown vector  $\boldsymbol{\theta}_0$  provided  $\lambda > 1$ : this phenomenon is known as the BBAP phase transition [Baik, Ben Arous, and P     \[2005\]](#).

From a statistical perspective, the principal eigenvector is known to be an asymptotically optimal estimator if no additional information is available about  $\boldsymbol{\theta}_0$ . In particular, it is asymptotically equivalent to the Bayes-optimal estimator when the prior of  $\boldsymbol{\theta}_0$  is uniformly distributed on a sphere of radius  $\sqrt{n}$ . However, in many problems of interest, additional information is available on  $\boldsymbol{\theta}_0$ : exploiting this information optimally requires to move away from spectral methods and from the familiar grounds of random matrix theory.

**3.1  $\mathbb{Z}_2$ -synchronization.** In some cases, all the entries of  $\boldsymbol{\theta}_0$  are known to have equal magnitude. For instance, in the community detection problem we might be required to partition the vertices of a graph in two communities such that vertices are better connected within each part than across the partition. Under the so-called stochastic block model [Decelle, Krzakala, Moore, and Zdeborov   \[2011\]](#) and [Abbe \[2017\]](#), the adjacency matrix of the graph is of the form (3-1) (albet with Bernoulli rather than Gaussian noise) whereby  $\theta_{0,i} \in \{+1, -1\}$  is the label of vertex  $i \in [n]$ . Another motivation comes from group synchronization [Wang and Singer \[2013\]](#), which is a relative of model (3-1) whereby the unknowns  $\theta_{0,i}$  are elements of a compact matrix group  $\mathcal{G}$ . In the special case  $\mathcal{G} = \mathbb{Z}_2 = (\{+1, -1\}, \cdot)$ , the resulting model is a special case of (3-1).

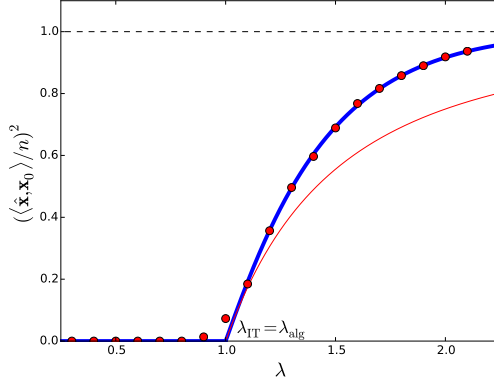


Figure 2: Estimation accuracy  $(\langle \hat{\boldsymbol{\theta}}^{\text{Bayes}}, \boldsymbol{\theta}_0 \rangle / n)^2$  within the  $\mathbb{Z}_2$ -synchronization problem. Red circles: numerical simulations with the AMP algorithm (form matrices of dimension  $n = 2000$  and  $t = 200$  iterations). Continuous thick blue line: Bayes optimal estimation accuracy, cf. [Theorem 3](#). Dashed blue line: other fixed points of state evolution. Red line: Accuracy achieved by principal component analysis.

The following theorem follows from [Deshpande, Abbe, and Montanari \[2017\]](#) and [Montanari and Venkataramanan \[2017\]](#) and provides an asymptotically exact characterization of optimal estimation in the  $\mathbb{Z}_2$ -synchronization problem, with respect to the metric in [Equation \(3-2\)](#).

**Theorem 3.** Consider the model [\(3-1\)](#) and let  $\gamma_* \in [0, \infty)$  denote the largest solution of

$$(3-3) \quad \gamma = \lambda(1 - \text{mmse}(\gamma)),$$

where  $\text{mmse}(\cdot)$  is defined as in [Equation \(2-13\)](#), with  $p_\Theta = (1/2)\delta_{+1} + (1/2)\delta_{-1}$ .

Then, there exists an estimator  $\hat{\boldsymbol{\theta}}^{\text{Bayes}} : \mathbf{X} \mapsto \hat{\boldsymbol{\theta}}^{\text{Bayes}}(\mathbf{X})$  such that, almost surely,

$$(3-4) \quad \lim_{n \rightarrow \infty} \frac{|\langle \hat{\boldsymbol{\theta}}^{\text{Bayes}}(\mathbf{X}), \boldsymbol{\theta}_0 \rangle|}{\|\hat{\boldsymbol{\theta}}^{\text{Bayes}}(\mathbf{X})\|_2 \|\boldsymbol{\theta}_0\|_2} = \sqrt{1 - \text{mmse}(\gamma_*)}.$$

Further, this accuracy can be approximated within arbitrarily small (constant) additive error  $\varepsilon$  by a polynomial-time message passing algorithm, cf. [Section 4](#). Finally, no estimator can achieve a better correlation than in [Equation \(3-4\)](#).

This prediction is illustrated in [Figure 2](#). Notice that it undergoes a phase transition at the spectral threshold  $\lambda = 1$ . For  $\lambda < 1$  no estimator can achieve a correlation that is bounded away from zero.

**Remark 3.1.** Substantial generalizations of the last theorem were proved in several papers [Barbier, Dia, Macris, Krzakala, Lesieur, and Zdeborová \[2016\]](#), [Lelarge and Miolane \[2016\]](#), and [Miolane \[2017\]](#). These generalization use new proof techniques inspired by mathematical spin glass theory and cover the case of vectors  $\theta$  whose entries have general distributions  $p_\Theta$ , as well as the rectangular and higher rank cases.

In particularly, [Theorem 3](#) holds almost verbatimly if  $\theta_0$  has i.i.d. entries with known distribution  $p_\Theta$  such that  $\int \theta^2 p_\Theta(d\theta) = 1$  and  $\int \theta^4 p_\Theta(d\theta) < \infty$ . One important difference is that in this more general setting, [Equation \(3-3\)](#) can have multiple solutions, and [Barbier, Dia, Macris, Krzakala, Lesieur, and Zdeborová \[2016\]](#), [Lelarge and Miolane \[2016\]](#), and [Miolane \[2017\]](#) provide a way to select the ‘correct’ solution that is analogous to the one in [Theorem 2](#).

**Remark 3.2.** As in the linear regression problem, the fixed point [Equation \(3-3\)](#) points at a connection between the high-dimensional estimation problem of [Equation \(3-1\)](#), where we are required to estimate  $n$  bits of information  $\theta_{0,i} \in \{+1, -1\}$ , to a much simpler scalar problem. The underlying mechanism is again the decoupling phenomenon of [Section 2.3](#). An alternative viewpoint on the same phenomenon is provided by the analysis of message passing algorithms outlined in [Section 4](#).

**Remark 3.3.** Replacing the minimum mean-square estimator  $\mathbb{E}\{\Theta|Y\}$  with the optimal linear estimator  $\hat{\Theta}(Y) = aY$  in the definition of [Equation \(2-13\)](#) yields the general upper bound (for  $\mathbb{E}\{\Theta^2\} = 1$ )  $\text{mmse}(s) \leq 1/(1+s)$ . Substituting in [Equations \(3-3\)](#) and [\(3-4\)](#) this yields in turn

$$(3-5) \quad \lim_{n \rightarrow \infty} \frac{|\langle \hat{\theta}^{\text{Bayes}}(X), \theta_0 \rangle|}{\|\hat{\theta}^{\text{Bayes}}(X)\|_2 \|\theta_0\|_2} \geq \sqrt{\left(1 - \frac{1}{\lambda^2}\right)_+}.$$

We thus recover the predicted accuracy of spectral methods, cf. [3-2](#). It is not hard to show that this inequality is strict unless the coordinates of  $\theta_0$  are asymptotically Gaussian. [Figure 2](#) compares the Bayes optimal accuracy of [Theorem 3](#) with this spectral lower bound.

While [Theorem 3](#) states that there exists a message passing algorithm that essentially achieves Bayes-optimal performances, this type of algorithms can be sensitive to model misspecification. It is therefore interesting to consider other algorithmic approaches. One standard starting point is to consider the maximum likelihood estimator that is obtained



by solving the following optimization problem:

$$(3-6) \quad \begin{aligned} & \text{maximize} && \langle X, \theta \theta^\top \rangle, \\ & \text{subject to} && \theta \in \{+1, -1\}^n. \end{aligned}$$

Semidefinite programing (SDP) relaxations provide a canonical path to obtain a tractable algorithm for such combinatorial problems. A very popular relaxation for the present case [Goemans and Williamson \[1995\]](#) and [Nesterov \[1998\]](#) is the following program in the decision variable  $Q \in \mathbb{R}^{n \times n}$ :

$$(3-7) \quad \begin{aligned} & \text{maximize} && \langle X, Q \rangle, \\ & \text{subject to} && Q \succeq 0, \\ & && Q_{ii} = 1 \text{ for all } i \in \{1, \dots, n\}. \end{aligned}$$

The matrix  $Q$  can be interpreted as a covariance matrix for a certain distribution on the vector  $\theta$ . Once a solution  $Q_*$  of this SDP is computed, we can use it to produce an estimate  $\hat{\theta}^{\text{SDP}} \in \{+1, -1\}^n$  in many ways (this step is called ‘rounding’ in theoretical computer science). For instance, we can take the sign of its principal eigenvector:  $\hat{\theta} = \text{sign}(v_1(Q_*))$ . There are many open questions concerning the SDP ([Equation \(3-7\)](#)). In particular [Javanmard, Montanari, and Ricci-Tersenghi \[2016\]](#) uses statistical physics methods to obtain close form expression for its asymptotic accuracy, that are still unproven. On the positive side, [Montanari and Sen \[2016\]](#) establishes the following positive result.

**Theorem 4.** *Let  $X$  be generated according to the model (3-1) with  $\theta_0 \in \{+1, -1\}^n$ , and denote by  $Q_*$  the solution of the SDP ([Equation \(3-7\)](#)). Then there exists a rounding procedure that produces  $\hat{\theta}^{\text{SDP}} = \hat{\theta}^{\text{SDP}}(Q_*) \in \{+1, -1\}^n$  such that for any  $\lambda > 1$  there exists  $\varepsilon > 0$  such that, with high probability*

$$(3-8) \quad \frac{|\langle \hat{\theta}^{\text{SDP}}, \theta_0 \rangle|}{\|\hat{\theta}^{\text{SDP}}\|_2 \|\theta_0\|_2} \geq \varepsilon.$$

In other words, semidefinite programming matches the optimal threshold.

**3.2 The computation/information gap and the hidden clique problem.** It is worth emphasizing one specific aspect of [Theorem 3](#). Within the spiked matrix model (3-1), there exists a polynomial-time computable estimator that nearly achieves Bayes-optimal performances, despite the underlying estimation problem is combinatorial in nature:  $\theta_0 \in \{+1, -1\}^n$ .

It is important to stress that the existence of a polynomial-time estimator for the problem (3-1) is far from being the norm, when changing the distribution  $p_\Theta$ , and the signal-to-noise ratio  $\lambda$ . In certain cases, simple algorithms achieve nearly optimal performances. In others, even highly sophisticated approaches (for instance SDP relaxations from the sum-of-squares hierarchy Barak and Steurer [2014]) fail.

Developing a theory of which statistical estimation problems are solvable by polynomial-time algorithms is a central open problem in this area, and a very difficult one. For certain classes of problems, a bold conjecture was put forward on the basis of statistical physics insights.

In order to formulate this conjecture in the context of model (3-1), it is useful to state the following theorem from Montanari and Venkataramanan [2017] that concerns the case of a general distribution  $p_\Theta$  of the entries of  $\theta_0$ .

**Theorem 5.** *Consider –to be specific– model (3-1), with  $\theta_0$  having i.i.d. entries with known distribution  $p_\Theta$ . Assume  $p_\Theta$  and  $\lambda$  to be independent of  $n$  and known, with  $\int \theta^2 p_\Theta(d\theta) = 1$ . If  $\int \theta p_\Theta(d\theta) = 0$ , further assume  $\lambda > 1$ . Then there exists a polynomial time (message passing) algorithm that outputs an estimator  $\hat{\theta}^{\text{AMP}} = \hat{\theta}^{\text{AMP}}(X)$  such that*

$$(3-9) \quad \lim_{n \rightarrow \infty} \frac{|\langle \hat{\theta}^{\text{AMP}}(X), \theta_0 \rangle|}{\|\hat{\theta}^{\text{AMP}}(X)\|_2 \|\theta_0\|_2} = \sqrt{1 - \text{mmse}(\gamma_{\text{AMP}})}.$$

where  $\text{mmse}(\cdot)$  is defined as in Equation (2-13) and  $\gamma_{\text{AMP}}$  is the smallest non-zero fixed point of Equation (3-3).

Within the setting of this theorem, it is conjectured that Equation (3-9) is the optimal accuracy achieved by polynomial time estimators Barbier, Dia, Macris, Krzakala, Lesieur, and Zdeborová [2016], Lelarge and Miolane [2016], Lesieur, Krzakala, and Zdeborová [2017], and Montanari and Venkataramanan [2017]. Together with Remark 3.1, this provides a precise –albeit conjectural– picture of the gap between fundamental statistical limits (the Bayes optimal accuracy) and computationally efficient methods. This is sometimes referred to as the information-computation gap. The same phenomenon was pointed out earlier in other statistical estimation problems, e.g. in the context of error correcting codes Mézard and Montanari [2009].

The hidden clique problem is the prototypical example of a statistical estimation problem in which a large information-computation gap is present, and it is the problem for which this phenomenon is best studied. Nature generates a graph over  $n$  vertices as follows: a subset  $S \subseteq [n]$  of size  $|S| = k$  is chosen uniformly at random. Conditional on  $S$ ,

for any pair of vertices  $\{i, j\}$ , an edge is added independently with probability

$$(3-10) \quad \mathbb{P}(\{i, j\} \in E | S) = \begin{cases} 1 & \text{if } \{i, j\} \subseteq S, \\ 1/2 & \text{otherwise.} \end{cases}$$

We are given one realization  $G$  such a graph, and are requested to identify the set  $S$ . In order to clarify the connection with the rank-one plus noise model (3-1), denote by  $A$  the  $+/ -$  adjacency matrix of  $G$ . This is the  $n \times n$  matrix whose entry  $i, j$  is  $A_{ij} = +1$  if  $(i, j) \in E$  and  $-1$  otherwise (in what follows, all matrices have diagonal entries equal to  $+1$ ). Then it is easy to see that

$$(3-11) \quad \frac{1}{\sqrt{n}} A = \lambda \theta_0 \theta_0^\top + W - W_{S,S},$$

$$(3-12) \quad \theta_0 = \frac{1}{\sqrt{k}} \mathbf{1}_S, \quad \lambda = \frac{k}{\sqrt{n}},$$

where  $W_{S,S}$  is the restriction of matrix  $W = W^\top$  to rows/columns with index in  $S$  and  $(W_{ij})_{i < j} \sim_{iid} \text{Unif}(+1/\sqrt{n}, -1/\sqrt{n})$ . This model has a few differences with respect to the one in Equation (3-1): (i) The noise is Radamacher instead of Gaussian; (ii) The term  $W_{S,S}$  of noise is subtracted; (iii) The distribution of the entries of  $\theta_0$  is  $p_\Theta = (k/n)\delta_{1/\sqrt{k}} + (1 - (k/n))\delta_0$ ; hence, for  $\lambda = k/\sqrt{n}$  fixed,  $p_\Theta$  depends on  $n$ . Of these differences, only the last one is really important for our purposes, and changes some qualitative features of the problem.

From a purely statistical point of view, the set  $S$  can be reconstructed with high probability provided that  $k \geq 2(1 + \varepsilon) \log_2(n)$ , by searching over all subsets of  $k$  vertices. On the other hand, a variety of polynomial-time algorithms have been analyzed, including Monte Carlo Markov Chain [Jerrum \[1992\]](#), spectral algorithms [Alon, Krivelevich, and Sudakov \[1998\]](#), message passing algorithms [Deshpande and Montanari \[2015\]](#), semidefinite programming relaxations in the [Feige and Krauthgamer \[2003\]](#) and sum-of-squares [Barak, Hopkins, Kelner, Kothari, Moitra, and Potechin \[2016\]](#) hierarchies, statistical query models [Feldman, Grigorescu, Reyzin, S. S. Vempala, and Xiao \[2013\]](#). Despite all of these efforts, no polynomial-time algorithms is known to be effective with high probability for  $k \leq n^{1/2-\varepsilon}$ , suggesting the possibility of a large information/computation gap for the hidden clique problem. As shown in [Deshpande and Montanari \[2015\]](#), this is consistent with the general picture emerging from statistical physics (although the hidden clique problem does not fit in the setting of the conjecture mentioned above).

## 4 Message passing algorithms

Message passing algorithms were already mentioned a few times in the previous pages and provide one natural class of algorithms to deal with random structures. Also, they are intimately connected to mean field approximations in statistical physics. Given an undirected graph  $G = (V, E)$ , we introduce the set of directed edges  $\vec{E} = \{(i \rightarrow j) : (i, j) \in E\}$  (namely, for each edge  $(i, j) \in E$ , we introduce the two directed edges  $(i \rightarrow j)$  and  $(j \rightarrow i)$ ). A message passing algorithm operates on messages  $(v_{i \rightarrow j}^t)_{(i \rightarrow j) \in \vec{E}} \in \mathcal{M}^{\vec{E}}$  taking values in a set  $\mathcal{M}$ , with  $t$  a time index. Messages are updated according to local rules:

$$(4-1) \quad v_{i \rightarrow j}^{t+1} = \Psi_{i \rightarrow j}^{(t)}(v_{k \rightarrow i}^t : k \in \partial i \setminus j),$$

In other words, a message outgoing vertex  $i$  at time  $t + 1$  is a function of messages ingoing the same vertex at time  $t$ , with the exception of the message along the same edge. Here all edges are updated synchronously: asynchronous schemes are of interest as well.

Notice that rather than an algorithm, (4-1) describes a general class of dynamical systems: we did not specify what the updating function  $\Psi_{i \rightarrow j}^{(t)}$  are, what is the space  $\mathcal{M}$  in which messages live, and not even what is the problem that we are trying to solve. We only insisted on locality and the ‘non-backtracking information’ condition: these turn out to be sufficient to lead to some interesting properties of the dynamical system (4-1) when the underlying graph is a tree or locally tree-like [Richardson and Urbanke \[2008\]](#).

Special forms of the dynamics (4-1) are used for Bayesian inference [Koller and Friedman \[2009\]](#), decoding in digital communications [Richardson and Urbanke \[2008\]](#), and combinatorial optimization [Mézard and Montanari \[2009\]](#). To the best of my knowledge, the first appearance an algorithm of the form (4-1) (and its analysis) dates back to Gallager Ph.D. thesis on low-density parity check codes in the early sixties [Gallager \[1962\]](#). As an analytical tool, recursions of this type have been in use in physics at least since Bethe’s work in the thirties [Bethe \[1935\]](#).

At first sight, message passing algorithms might seem immaterial to the problems discussed in the rest of this paper: typically these are not associated to a locally tree-like graph (possibly with the exception of some sparse-graph versions of the hidden-clique problem [Deshpande and Montanari \[2015\]](#)). Somewhat surprisingly, there exists a natural class of algorithms whose datum is not a locally tree-like graph but a (dense) random matrix, and which can be considered a close relative of message passing algorithms. In fact, they can be thought as the limit of message passing algorithm when the average degree of the underlying graph diverges (see, for instance, [Bayati and Montanari \[2011\]](#)). These algorithms

are known as *approximate message passing*: for the sake of simplicity we will briefly discuss them in the case in which the data consists of a matrix  $\mathbf{A} \sim \text{GOE}(n)$ . The algorithm operates on variables  $\hat{\boldsymbol{\theta}}^t \in \mathbb{R}^{n \times k}$  where  $k$  is considered as fixed as  $n \rightarrow \infty$ . This state is updated according to

$$(4-2) \quad \begin{aligned} \hat{\boldsymbol{\theta}}^{t+1} &= \mathbf{A} f_t(\hat{\boldsymbol{\theta}}^t) - f_{t-1}(\hat{\boldsymbol{\theta}}^{t-1}) \mathbf{B}_t^\top, \\ \mathbf{B}_t &= \frac{1}{n} \sum_{i=1}^m \frac{\partial f_t}{\partial \hat{\boldsymbol{\theta}}_i^t}(\hat{\boldsymbol{\theta}}_i^t). \end{aligned}$$

Here  $f_t : \mathbb{R}^k \rightarrow \mathbb{R}^k$  is a Lipschitz continuous function and we denote by  $f_t(\hat{\boldsymbol{\theta}}^t) \in \mathbb{R}^{n \times k}$  the matrix that is obtained by applying  $f_t$  row-by-row to  $\hat{\boldsymbol{\theta}}^t$ . The  $i$ -th row of  $\hat{\boldsymbol{\theta}}^t$  is denoted by  $\hat{\boldsymbol{\theta}}_i^t$  and, by convention,  $\mathbf{B}_0 = 0$ . Once again, Equation (4-2) does not specify the update functions  $f_t$ , nor the problem we are trying to solve: rather it define a class of dynamical systems. However, special cases can be developed for Bayesian inference, statistical estimation, optimization, and so on.

In the Bayesian case, the functions  $f_t(\cdot)$  take the form of conditional expectations with respect to certain distributions, and the fixed point version of the iteration (4-2) dates back to the work of Thouless, Anderson, Palmer (TAP) on mean field spin glasses [Thouless, Anderson, and Palmer \[1977\]](#). Iterative solutions of the TAP equations were studied among others in [Bolthausen \[2014\]](#). The general (non-Bayesian) formulation was developed and analyzed in [D. L. Donoho, Maleki, and Montanari \[2009\]](#) and [Bayati and Montanari \[2011\]](#), with the original motivation being its application to compressed sensing.

Crucially, the recursion (4-2) admits an asymptotically exact characterization in the limit  $n \rightarrow \infty$  with  $t$  fixed. This type of analysis is known as *state evolution*.

**Theorem 6.** *Consider the AMP iteration (4-2) with  $f_t$  Lipschitz continuous,  $\mathbf{A} \sim \text{GOE}(n)$ , and deterministic initialization  $\hat{\boldsymbol{\theta}}^0$  such that  $\lim_{n \rightarrow \infty} f_0(\hat{\boldsymbol{\theta}}^0)^\top f_0(\hat{\boldsymbol{\theta}}_0)/n = \Sigma_0 \in \mathbb{R}^{k \times k}$ . Define the sequence  $\Sigma_t \in \mathbb{R}^{k \times k}$  via the recursion:*

$$(4-3) \quad \Sigma_{t+1} = \mathbb{E} \{ f_t(\Sigma_t^{1/2} \mathbf{g}) f_t(\Sigma_t^{1/2} \mathbf{g})^\top \},$$

where expectation is with respect to  $\mathbf{g} \sim \mathcal{N}(0, \mathbf{I}_k)$ . Then, for any  $t$  and any test function  $\psi : \mathbb{R}^k \rightarrow \mathbb{R}$  that is continuous and with at most quadratic growth at infinity, the following holds almost surely

$$(4-4) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n \psi(\hat{\boldsymbol{\theta}}_i^t) = \mathbb{E} \{ \psi(\Sigma_t^{1/2} \mathbf{g}) \}.$$

A theorem of this type was first proved in the case of the TAP equations for the Sherrington-Kirkpatrick model in [Bolthausen \[2014\]](#) and then in general in [Bayati and Montanari \[2011\]](#). Generalizations have also been proved for matrices with non-i.i.d. entries [Javanmard and Montanari \[2013\]](#), non-Gaussian random matrices [Bayati, Lelarge, and Montanari \[2015\]](#), non-separable functions  $f_t$  [Berthier, Montanari, and Nguyen \[2017\]](#), invariant matrix ensembles [Schniter, Rangan, and Fletcher \[2016\]](#), non-asymptotic settings [Rush and Venkataramanan \[2016\]](#), non-deterministic initializations [Montanari and Venkataramanan \[2017\]](#).

This type of analysis is used to prove the algorithmic part of [Theorem 3](#), as well as algorithmic versions of the other theorems in this paper.

## 5 Context and conclusion

For the greatest part of the last century, mean field theory has been an important tool used by physicists to understand the behavior of systems with a large number of degrees of freedom [Landau \[1937\]](#). Classical mean field theory describes homogeneous states, e.g. the state of a fluid in which each molecule interacts with the average environment created by all the other molecules. Starting in the late seventies, a new class mean-field ideas was developed to deal with heterogeneous states, where all particles look statistically the same, but typical configurations are highly heterogeneous, as is the case with disordered solids and spin glasses [Kirkpatrick and Sherrington \[1978\]](#) and [Parisi \[1979\]](#). This opened the way to applying the same tools to a variety of probabilistic models without apparent connection to physics, including combinatorial optimization and neural networks (see [Mézard, Parisi, and Virasoro \[1987\]](#) for seminal papers in this direction).

Over the last few years, this circle of ideas has gone through a spectacular renaissance for at least three reasons: (i) Mathematical methods have been developed to prove (part of) physicists' predictions [Talagrand \[2007\]](#), [Panchenko \[2013\]](#), and [Ding, Sly, and Sun \[2015\]](#); (ii) Structural insights from physics have unveiled new computational phenomena; (iii) New applications of these techniques have emerged within high-dimensional statistics and machine learning, generating interest across several communities.

This brief overview focused on the last two points, and hopefully will provide the reader with an entrypoint in this rapidly evolving literature.

## References

Emmanuel Abbe (2017). “Community detection and stochastic block models: recent developments”. arXiv: 1703.10146 (cit. on p. 2999).

- Gad Abraham and Michael Inouye (2014). “Fast principal component analysis of large-scale genome-wide data”. *PLoS one* 9.4, e93766 (cit. on p. [2992](#)).
- Noga Alon, Michael Krivelevich, and Benny Sudakov (1998). “Finding a large hidden clique in a random graph”. In: *Proceedings of the Ninth Annual ACM-SIAM Symposium on Discrete Algorithms (San Francisco, CA, 1998)*. ACM, New York, pp. 594–598. MR: [1642973](#) (cit. on p. [3004](#)).
- Jinho Baik, Gérard Ben Arous, and Sandrine Péché (2005). “Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices”. *Ann. Probab.* 33.5, pp. 1643–1697. MR: [2165575](#) (cit. on p. [2999](#)).
- Jinho Baik and Jack W. Silverstein (2006). “Eigenvalues of large sample covariance matrices of spiked population models”. *J. Multivariate Anal.* 97.6, pp. 1382–1408. MR: [2279680](#) (cit. on p. [2999](#)).
- Boaz Barak, Samuel B. Hopkins, Jonathan Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin (2016). “A nearly tight sum-of-squares lower bound for the planted clique problem”. In: *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*. IEEE Computer Soc., Los Alamitos, CA, pp. 428–437. MR: [3631005](#) (cit. on p. [3004](#)).
- Boaz Barak and David Steurer (2014). “Sum-of-squares proofs and the quest toward optimal algorithms”. arXiv: [1404.5236](#) (cit. on p. [3003](#)).
- Jean Barbier, Mohamad Dia, Nicolas Macris, Florent Krzakala, Thibault Lesieur, and Lenka Zdeborová (2016). “Mutual information for symmetric rank-one matrix estimation: A proof of the replica formula”. In: *Advances in Neural Information Processing Systems*, pp. 424–432 (cit. on pp. [3001](#), [3003](#)).
- Jean Barbier, Nicolas Macris, Mohamad Dia, and Florent Krzakala (2017). “Mutual Information and Optimality of Approximate Message-Passing in Random Linear Estimation”. arXiv: [1701.05823](#) (cit. on p. [2997](#)).
- Mohsen Bayati, Marc Lelarge, and Andrea Montanari (2015). “Universality in polytope phase transitions and message passing algorithms”. *Ann. Appl. Probab.* 25.2, pp. 753–822. MR: [3313755](#) (cit. on pp. [2996](#), [3007](#)).
- Mohsen Bayati and Andrea Montanari (2011). “The dynamics of message passing on dense graphs, with applications to compressed sensing”. *IEEE Trans. Inform. Theory* 57.2, pp. 764–785. MR: [2810285](#) (cit. on pp. [3005](#)–[3007](#)).
- (2012). “The LASSO risk for Gaussian matrices”. *IEEE Trans. Inform. Theory* 58.4, pp. 1997–2017. MR: [2951312](#) (cit. on p. [2994](#)).
- Florent Benaych-Georges and Raj Rao Nadakuditi (2011). “The eigenvalues and eigenvectors of finite, low rank perturbations of large random matrices”. *Adv. Math.* 227.1, pp. 494–521. MR: [2782201](#) (cit. on p. [2999](#)).

- (2012). “The singular values and vectors of low rank perturbations of large rectangular random matrices”. *J. Multivariate Anal.* 111, pp. 120–135. MR: [2944410](#) (cit. on p. 2999).
- Raphael Berthier, Andrea Montanari, and Phan-Minh Nguyen (2017). “State Evolution for Approximate Message Passing with Non-Separable Functions”. arXiv: [1708.03950](#) (cit. on p. 3007).
- Hans A Bethe (1935). “Statistical theory of superlattices”. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences* 150.871, pp. 552–575 (cit. on p. 3005).
- Peter J. Bickel, Ya’acov Ritov, and Alexandre B. Tsybakov (2009). “Simultaneous analysis of lasso and Dantzig selector”. *Ann. Statist.* 37.4, pp. 1705–1732. MR: [2533469](#) (cit. on p. 2993).
- Erwin Bolthausen (2014). “An iterative construction of solutions of the TAP equations for the Sherrington-Kirkpatrick model”. *Comm. Math. Phys.* 325.1, pp. 333–366. MR: [3147441](#) (cit. on pp. 3006, 3007).
- Emmanuel J. Candès, Justin Romberg, and Terence Tao (2006). “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information”. *IEEE Trans. Inform. Theory* 52.2, pp. 489–509. MR: [2236170](#) (cit. on p. 2992).
- Emmanuel Candes and Terence Tao (2007). “The Dantzig selector: statistical estimation when  $p$  is much larger than  $n$ ”. *Ann. Statist.* 35.6, pp. 2313–2351. MR: [2382644](#) (cit. on p. 2993).
- Mireille Capitaine, Catherine Donati-Martin, and Delphine Féral (2009). “The largest eigenvalues of finite rank deformation of large Wigner matrices: convergence and nonuniversality of the fluctuations”. *Ann. Probab.* 37.1, pp. 1–47. MR: [2489158](#) (cit. on p. 2999).
- Venkat Chandrasekaran, Benjamin Recht, Pablo A. Parrilo, and Alan S. Willsky (2012). “The convex geometry of linear inverse problems”. *Found. Comput. Math.* 12.6, pp. 805–849. MR: [2989474](#) (cit. on p. 2997).
- Scott Chen and David L Donoho (1995). “Examples of basis pursuit”. In: *Wavelet Applications in Signal and Image Processing III*. Vol. 2569. International Society for Optics and Photonics, pp. 564–575 (cit. on p. 2992).
- Aurelien Decelle, Florent Krzakala, Cristopher Moore, and Lenka Zdeborová (2011). “Asymptotic analysis of the stochastic block model for modular networks and its algorithmic applications”. *Physical Review E* 84.6, p. 066106 (cit. on p. 2999).
- Yash Deshpande, Emmanuel Abbe, and Andrea Montanari (2017). “Asymptotic mutual information for the balanced binary stochastic block model”. *Inf. Inference* 6.2, pp. 125–170. MR: [3671474](#) (cit. on p. 3000).



- Yash Deshpande and Andrea Montanari (2015). “Finding hidden cliques of size  $\sqrt{N/e}$  in nearly linear time”. *Found. Comput. Math.* 15.4, pp. 1069–1128. MR: [3371378](#) (cit. on pp. [3004](#), [3005](#)).
- Jian Ding, Allan Sly, and Nike Sun (2015). “Proof of the satisfiability conjecture for large  $k$  [extended abstract]”. In: *STOC’15—Proceedings of the 2015 ACM Symposium on Theory of Computing*. ACM, New York, pp. 59–68. MR: [3388183](#) (cit. on p. [3007](#)).
- David L. Donoho (2006). “Compressed sensing”. *IEEE Trans. Inform. Theory* 52.4, pp. 1289–1306. MR: [2241189](#) (cit. on p. [2992](#)).
- David L. Donoho, Arian Maleki, and Andrea Montanari (2009). “Message Passing Algorithms for Compressed Sensing”. *Proceedings of the National Academy of Sciences* 106, pp. 18914–18919 (cit. on p. [3006](#)).
- (2011). “The noise-sensitivity phase transition in compressed sensing”. *IEEE Trans. Inform. Theory* 57.10, pp. 6920–6941. MR: [2882271](#) (cit. on p. [2995](#)).
- David Donoho and Andrea Montanari (2016). “High dimensional robust M-estimation: asymptotic variance via approximate message passing”. *Probab. Theory Related Fields* 166.3-4, pp. 935–969. MR: [3568043](#) (cit. on p. [2996](#)).
- David Donoho and Jared Tanner (2009). “Observed universality of phase transitions in high-dimensional geometry, with implications for modern data analysis and signal processing”. *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.* 367.1906. With electronic supplementary materials available online, pp. 4273–4293. MR: [2546388](#) (cit. on p. [2996](#)).
- Noureddine El Karoui, Derek Bean, Peter J Bickel, Chinghay Lim, and Bin Yu (2013). “On robust regression with high-dimensional predictors”. *Proceedings of the National Academy of Sciences* 110.36, pp. 14557–14562 (cit. on p. [2996](#)).
- Uriel Feige and Robert Krauthgamer (2003). “The probable value of the Lovász-Schrijver relaxations for maximum independent set”. *SIAM J. Comput.* 32.2, pp. 345–370. MR: [1969394](#) (cit. on p. [3004](#)).
- Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh S. Vempala, and Ying Xiao (2013). “Statistical algorithms and a lower bound for detecting planted cliques”. In: *STOC’13—Proceedings of the 2013 ACM Symposium on Theory of Computing*. ACM, New York, pp. 655–664. MR: [3210827](#) (cit. on p. [3004](#)).
- Delphine Féral and Sandrine Péché (2007). “The largest eigenvalue of rank one deformation of large Wigner matrices”. *Comm. Math. Phys.* 272.1, pp. 185–228. MR: [2291807](#) (cit. on p. [2999](#)).
- R. G. Gallager (1962). “Low-density parity-check codes”. *IRE Trans. IT*-8, pp. 21–28. MR: [0136009](#) (cit. on p. [3005](#)).
- Carl Friedrich Gauss (2011). *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*. Cambridge Library Collection. Reprint of the 1809 original. Cambridge University Press, Cambridge, pp. xii+228+21. MR: [2858122](#) (cit. on p. [2992](#)).

- Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure (2014). “On asymptotically optimal confidence regions and tests for high-dimensional models”. *Ann. Statist.* 42.3, pp. 1166–1202. MR: [3224285](#) (cit. on p. 2996).
- Michel X. Goemans and David P. Williamson (1995). “Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming”. *J. Assoc. Comput. Mach.* 42.6, pp. 1115–1145. MR: [1412228](#) (cit. on p. 3002).
- Y. Gordon (1988). “On Milman’s inequality and random subspaces which escape through a mesh in  $\mathbf{R}^n$ ”. In: *Geometric aspects of functional analysis (1986/87)*. Vol. 1317. Lecture Notes in Math. Springer, Berlin, pp. 84–106. MR: [950977](#) (cit. on p. 2997).
- Dongning Guo, Shlomo Shamai, and Sergio Verdú (2005). “Mutual information and minimum mean-square error in Gaussian channels”. *IEEE Trans. Inform. Theory* 51.4, pp. 1261–1282. MR: [2241490](#) (cit. on p. 2997).
- Peter J. Huber (1964). “Robust estimation of a location parameter”. *Ann. Math. Statist.* 35, pp. 73–101. MR: [0161415](#) (cit. on p. 2996).
- (1973). “Robust regression: asymptotics, conjectures and Monte Carlo”. *Ann. Statist.* 1, pp. 799–821. MR: [0356373](#) (cit. on p. 2996).
- Peter J. Huber and Elvezio M. Ronchetti (2009). *Robust statistics*. Second. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc., Hoboken, NJ, xvi+354 pp. + loose erratum. MR: [2488795](#) (cit. on p. 2996).
- Adel Javanmard and Andrea Montanari (2013). “State evolution for general approximate message passing algorithms, with applications to spatial coupling”. *Inf. Inference* 2.2, pp. 115–144. MR: [3311445](#) (cit. on p. 3007).
- (2014a). “Confidence intervals and hypothesis testing for high-dimensional regression”. *J. Mach. Learn. Res.* 15, pp. 2869–2909. MR: [3277152](#) (cit. on p. 2996).
- (2014b). “Hypothesis testing in high-dimensional regression under the Gaussian random design model: asymptotic theory”. *IEEE Trans. Inform. Theory* 60.10, pp. 6522–6554. MR: [3265038](#) (cit. on p. 2996).
- (2015). “De-biasing the Lasso: Optimal Sample Size for Gaussian Designs”. arXiv: [1508.02757](#) (cit. on p. 2996).
- Adel Javanmard, Andrea Montanari, and Federico Ricci-Tersenghi (2016). “Phase transitions in semidefinite relaxations”. *Proc. Natl. Acad. Sci. USA* 113.16, E2218–E2223. MR: [3494080](#) (cit. on p. 3002).
- Mark Jerrum (1992). “Large cliques elude the Metropolis process”. *Random Structures Algorithms* 3.4, pp. 347–359. MR: [1179827](#) (cit. on p. 3004).
- Iain M. Johnstone (2001). “On the distribution of the largest eigenvalue in principal components analysis”. *Ann. Statist.* 29.2, pp. 295–327. MR: [1863961](#) (cit. on p. 2999).
- Ravi Kannan, Santosh Vempala, and Adrian Vetta (2004). “On clusterings: good, bad and spectral”. *J. ACM* 51.3, pp. 497–515. MR: [2145863](#) (cit. on p. 2992).

- Noureddine El Karoui (2013). “Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators : rigorous results”. arXiv: [1311.2445](#) (cit. on p. [2996](#)).
- Scott Kirkpatrick and David Sherrington (1978). “Infinite-ranged models of spin-glasses”. *Physical Review B* 17.11, pp. 4384–4403 (cit. on p. [3007](#)).
- Antti Knowles and Jun Yin (2013). “The isotropic semicircle law and deformation of Wigner matrices”. *Comm. Pure Appl. Math.* 66.11, pp. 1663–1750. MR: [3103909](#) (cit. on p. [2999](#)).
- Daphne Koller and Nir Friedman (2009). *Probabilistic graphical models*. Adaptive Computation and Machine Learning. Principles and techniques. MIT Press, Cambridge, MA, pp. xxxvi+1231. MR: [2778120](#) (cit. on p. [3005](#)).
- Satish Babu Korada and Andrea Montanari (2011). “Applications of the Lindeberg principle in communications and statistical learning”. *IEEE Trans. Inform. Theory* 57.4, pp. 2440–2450. MR: [2809100](#) (cit. on p. [2996](#)).
- Lev Davidovich Landau (1937). “On the theory of phase transitions”. *Ukr. J. Phys.* 7, pp. 19–32 (cit. on p. [3007](#)).
- Marc Lelarge and Léo Miolane (2016). “Fundamental limits of symmetric low-rank matrix estimation”. arXiv: [1611.03888](#) (cit. on pp. [3001](#), [3003](#)).
- Thibault Lesieur, Florent Krzakala, and Lenka Zdeborová (2017). “Constrained low-rank matrix estimation: phase transitions, approximate message passing and applications”. *J. Stat. Mech. Theory Exp.* 7, pp. 073403, 86. arXiv: [1701.00858](#). MR: [3683819](#) (cit. on p. [3003](#)).
- H Brendan McMahan et al. (2013). “Ad click prediction: a view from the trenches”. In: *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, pp. 1222–1230 (cit. on p. [2992](#)).
- Marc Mézard and Andrea Montanari (2009). *Information, physics, and computation*. Oxford Graduate Texts. Oxford University Press, Oxford, pp. xiv+569. MR: [2518205](#) (cit. on pp. [3003](#), [3005](#)).
- Marc Mézard, Giorgio Parisi, and Miguel Angel Virasoro (1987). *Spin glass theory and beyond*. Vol. 9. World Scientific Lecture Notes in Physics. World Scientific Publishing Co., Inc., Teaneck, NJ, pp. xiv+461. MR: [1026102](#) (cit. on p. [3007](#)).
- Léo Miolane (2017). “Fundamental limits of low-rank matrix estimation: the non-symmetric case”. arXiv: [1702.00473](#) (cit. on p. [3001](#)).
- Andrea Montanari (2012). “Graphical Models Concepts in Compressed Sensing”. In: *Compressed Sensing: Theory and Applications*. Ed. by Y.C. Eldar and G. Kutyniok. Cambridge University Press (cit. on p. [2995](#)).
- Andrea Montanari and Subhabrata Sen (2016). “Semidefinite programs on sparse random graphs and their application to community detection”. In: *STOC’16—Proceedings of*

- the 48th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, pp. 814–827. MR: [3536616](#) (cit. on p. [3002](#)).
- Andrea Montanari and Ramji Venkataramanan (2017). “Estimation of Low-Rank Matrices via Approximate Message Passing”. arXiv: [1711.01682](#) (cit. on pp. [3000](#), [3003](#), [3007](#)).
- Yu. Nesterov (1998). “Semidefinite relaxation and nonconvex quadratic optimization”. *Optim. Methods Softw.* 9.1-3, pp. 141–160. MR: [1618100](#) (cit. on p. [3002](#)).
- Samet Oymak and Joel A. Tropp (2015). “Universality laws for randomized dimension reduction, with applications”. arXiv: [1511.09433](#) (cit. on p. [2996](#)).
- Dmitry Panchenko (2013). *The Sherrington-Kirkpatrick model*. Springer Monographs in Mathematics. Springer, New York, pp. xii+156. MR: [3052333](#) (cit. on p. [3007](#)).
- Giorgio Parisi (1979). “Infinite number of order parameters for spin-glasses”. *Phys. Rev. Lett.* 43.23, p. 1754 (cit. on p. [3007](#)).
- Debashis Paul (2007). “Asymptotics of sample eigenstructure for a large dimensional spiked covariance model”. *Statist. Sinica* 17.4, pp. 1617–1642. MR: [2399865](#) (cit. on p. [2999](#)).
- Galen Reeves and Henry D Pfister (2016). “The replica-symmetric prediction for compressed sensing with Gaussian matrices is exact”. In: *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, pp. 665–669 (cit. on p. [2997](#)).
- Tom Richardson and Rüdiger Urbanke (2008). *Modern coding theory*. Cambridge University Press, Cambridge, pp. xvi+572. MR: [2494807](#) (cit. on p. [3005](#)).
- Cyrille Rossant et al. (2016). “Spike sorting for large, dense electrode arrays”. *Nature neuroscience* 19.4, pp. 634–641 (cit. on p. [2992](#)).
- Cynthia Rush and Ramji Venkataramanan (2016). “Finite-sample analysis of approximate message passing”. In: *Information Theory (ISIT), 2016 IEEE International Symposium on*. IEEE, pp. 755–759 (cit. on p. [3007](#)).
- Philip Schniter, Sundeep Rangan, and Alyson K Fletcher (2016). “Vector approximate message passing for the generalized linear model”. In: *Signals, Systems and Computers, 2016 50th Asilomar Conference on*. IEEE, pp. 1525–1529 (cit. on p. [3007](#)).
- Shirish Krishnaj Shevade and S Sathiya Keerthi (2003). “A simple and efficient algorithm for gene selection using sparse logistic regression”. *Bioinformatics* 19.17, pp. 2246–2253 (cit. on p. [2992](#)).
- A. J. Stam (1959). “Some inequalities satisfied by the quantities of information of Fisher and Shannon”. *Information and Control* 2, pp. 101–112. MR: [0109101](#) (cit. on p. [2997](#)).
- Michel Talagrand (2007). “Mean field models for spin glasses: some obnoxious problems”. In: *Spin glasses*. Vol. 1900. Lecture Notes in Math. Springer, Berlin, pp. 63–80. MR: [2309598](#) (cit. on p. [3007](#)).
- David J Thouless, Philip W Anderson, and Robert G Palmer (1977). “Solution of ‘solvable model of a spin glass’”. *Philosophical Magazine* 35.3, pp. 593–601 (cit. on p. [3006](#)).

- Christos Thrampoulidis, Samet Oymak, and Babak Hassibi (2015). “Regularized linear regression: A precise analysis of the estimation error”. In: *Conference on Learning Theory*, pp. 1683–1709 (cit. on p. 2997).
- Robert Tibshirani (1996). “Regression shrinkage and selection via the lasso”. *J. Roy. Statist. Soc. Ser. B* 58.1, pp. 267–288. MR: 1379242 (cit. on p. 2992).
- Antonia M Tulino, Giuseppe Caire, Sergio Verdu, and Shlomo Shamai (2013). “Support recovery with sparsely sampled free random matrices”. *IEEE Transactions on Information Theory* 59.7, pp. 4243–4271 (cit. on p. 2996).
- Lanhui Wang and Amit Singer (2013). “Exact and stable recovery of rotations for robust synchronization”. *Inf. Inference* 2.2, pp. 145–193. MR: 3311446 (cit. on p. 2999).
- Cun-Hui Zhang and Stephanie S. Zhang (2014). “Confidence intervals for low dimensional parameters in high dimensional linear models”. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 76.1, pp. 217–242. MR: 3153940 (cit. on p. 2996).

Received 2017-12-01.

ANDREA MONTANARI  
DEPARTMENT OF ELECTRICAL ENGINEERING AND DEPARTMENT OF STATISTICS  
STANFORD UNIVERSITY  
[montanari@stanford.edu](mailto:montanari@stanford.edu)

# NONPARAMETRIC ADDITIVE REGRESSION

BYEONG U. PARK

## ABSTRACT

In this article we discuss statistical methods of estimating structured nonparametric regression models. Our discussion is mainly on the additive models where the regression function (map) is expressed as a sum of unknown univariate functions (maps), but it also covers some other non- and semi-parametric models. We present the state of the art in the subject area with the prospect of an extension to non-Euclidean data objects.

## 1 Introduction

Let  $Y$  be a scalar random variable and  $\mathbf{X} \equiv (X_1, \dots, X_d)$  be a  $d$ -dimensional random vector. Suppose that one has observations  $(\mathbf{X}_i, Y_i)$ ,  $1 \leq i \leq n$ , that are independent and identically distributed copies of  $(\mathbf{X}, Y)$ . The regression problem in statistics is to estimate the conditional mean  $f(\mathbf{x}) \equiv E(Y | \mathbf{X} = \mathbf{x})$  using the observations  $(\mathbf{X}_i, Y_i)$ . The parametric approach to this problem is to assume that the true regression function  $f$  belongs to a finite-dimensional model  $\mathcal{F}$ . The simplest example of  $\mathcal{F}$  is a linear model  $\mathcal{F} = \{f(\cdot, \boldsymbol{\theta}) : \boldsymbol{\theta} \in \mathbb{R}^{d+1}\}$ , where  $f(\mathbf{x}, \boldsymbol{\theta}) = \theta_0 + \theta_1 x_1 + \dots + \theta_d x_d$ . This is certainly restrictive excluding many important realities. The nonparametric approach, on the contrary, is to allow the unknown  $f$  to lie in an infinite-dimensional function space. The problem is clearly ‘ill-posed’ since one is given only a finite number of observations  $(\mathbf{X}_i, Y_i)$ . One way, called method of sieves, is to reduce  $\mathcal{F}$  to a subspace  $\mathcal{F}_n$  in such a way that the sequence of sieve spaces  $\mathcal{F}_n$  grows as  $n$  increases and one searches for an estimator among functions in  $\mathcal{F}_n$ . Another way of solving the ill-posed inverse problem is through penalization, putting more penalties for functions that are more complex to enforce smoothness

---

This research was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIP) (NRF-2015R1A2A1A05001753).

*MSC2010:* primary 62G08; secondary 62G20.

*Keywords:* Additive models, smooth backfitting, varying coefficient models, partially linear models, errors-in variables, Hilbertian responses, Bochner integral.

for the resulting estimator. The approach termed as ‘kernel smoothing’ has quite a different nature and is based on localization. It basically converts the infinite-dimensional problem to solving locally finite-dimensional problems with the localization being finer for larger sample size  $n$ . In this paper we discuss nonparametric regression, focusing on kernel smoothing.

There is another problem of dimensionality. When the dimension  $d$  of  $\mathbf{X}$  gets high, all nonparametric estimation techniques fail theoretically. For instance, if  $\mathcal{F}$  is a class of functions with two continuous (partial) derivatives, then one cannot get an estimator  $\hat{f}$  that has a rate faster than  $n^{-2/(d+4)}$  for  $\|\hat{f} - f\|_2$ . Nonparametric methods fail practically as well when  $d$  is high. In the case of local kernel smoothing one basically takes  $\mathcal{X}_h(\mathbf{x}) \equiv \{(\mathbf{X}_i, Y_i) : \mathbf{X}_i \text{ are within distance } h \text{ from } \mathbf{x}\}$  for each  $\mathbf{x}$ , where  $h > 0$  is termed as ‘window width’ or ‘bandwidth’, and then estimate  $f(\mathbf{x})$  using those  $(\mathbf{X}_i, Y_i) \in \mathcal{X}_h(\mathbf{x})$ . The practical difficulty one encounters here is that one cannot choose  $h$  small enough for a fine local approximation of  $f$  since the number of  $(\mathbf{X}_i, Y_i)$  in  $\mathcal{X}_h(\mathbf{x})$ , which is asymptotic to  $nh^d$ , gets smaller very fast as  $h$  decreases when  $d$  is high. Note that one needs  $nh^d \geq \ell$  for the corresponding locally  $\ell$ -dimensional problem to be well-posed. This phenomenon, referred to as ‘the curse of dimensionality’, is present in other nonparametric methods such as sieves and penalization techniques.

Structured nonparametric models have been studied to circumvent the curse of dimensionality. A structured nonparametric model is defined as a known function of lower-dimensional unknown underlying functions, see [Mammen and J. P. Nielsen \[2003\]](#) for discussion on generalized structured models. They typically allow reliable estimation when a full nonparametric model does not work. The simplest example is the additive model

$$(1-1) \quad E(Y | \mathbf{X} = \mathbf{x}) = f_1(x_1) + \cdots + f_d(x_d),$$

where  $f_j$  are unknown univariate smooth functions. This model was first introduced by [Friedman and Stuetzle \[1981\]](#). Various nonparametric regression problems reduce to the estimation of this model. Examples include nonparametric regression with time series errors or with repeated measurements, panels with individual effects and semiparametric GARCH models, see [Mammen, Park, and Schienle \[2014\]](#).

Three main techniques of fitting the model (1-1) are ordinary backfitting ([Buja, Hastie, and Tibshirani \[1989\]](#)), marginal integration ([Linton and J. P. Nielsen \[1995\]](#)) and smooth backfitting (SBF, [Mammen, Linton, and J. Nielsen \[1999\]](#)). A difficulty with the ordinary backfitting technique is that the estimator of (1-1) is defined only when the backfitting iteration converges, as its limit. It is known that the backfitting iteration converges under rather strong conditions on the joint distribution of the covariates, see [Opsomer and Ruppert \[1997\]](#) and [Opsomer \[2000\]](#). For marginal integration, the main drawback is that it

does not resolve the dimensionality issue since it requires consistent estimation of the full-dimensional density of  $\mathbf{X}$ , see [Y. K. Lee \[2004\]](#). Smooth backfitting, on the other hand, is not subject to these difficulties. The method gives a well-defined estimator of the model and the iterative algorithm converges always under weak conditions. Furthermore, it has been shown for many structured nonparametric models that smooth backfitting estimators have univariate rates of convergence regardless of the dimension  $d$ .

In this paper, we revisit the theory of smooth backfitting for the additive regression model (1-1). We discuss some important extensions that include varying coefficient models, the case of errors-in-variables, some structured models for functional response and/or predictors and a general framework with Hilbertian response. Our discussion is primarily on the i.i.d. case where  $(\mathbf{X}_i, Y_i)$  are independent across  $1 \leq i \leq n$  and identically distributed, and for Nadaraya-Watson (locally constant) kernel smoothing since the theory is best understood under this setting.

## 2 Additive regression models

Let the distributions of  $X_j$  have densities  $p_j$  with respect to the Lebesgue measure on  $\mathbb{R}$ , and  $\mathbf{X}$  have a joint density  $p$  with respect to the Lebesgue measure on  $\mathbb{R}^d$ . We assume that  $p_j$  are commonly supported on the unit interval  $[0, 1]$ , for simplicity. In the original theory of [Mammen, Linton, and J. Nielsen \[1999\]](#), it is assumed that the joint density  $p$  is bounded away from zero on  $[0, 1]^d$ . Here, we relax this condition to requiring only that each marginal density  $p_j$  is bounded away from zero on  $[0, 1]$ .

**2.1 SBF estimation.** Let  $p_{jk}$  denote the two-dimensional joint densities of  $(X_j, X_k)$  for  $1 \leq j \neq k \leq d$ . From the model (1-1) we get a system of  $d$  integral equations,

$$(2-1) \quad f_j(x_j) = E(Y|X_j = x_j) - \sum_{k \neq j} \int_0^1 f_k(x_k) \frac{p_{jk}(x_j, x_k)}{p_j(x_j)} dx_k, \quad 1 \leq j \leq d.$$

The smooth backfitting method is nothing else than to replace the unknown marginal regression functions  $m_j \equiv E(Y|X_j = \cdot)$  and the marginal and joint densities  $p_j$  and  $p_{jk}$  by suitable estimators, and then to solve the resulting system of estimated integral equations. It is worthwhile to note here that the system of equations (2-1) only identifies  $f_+(\mathbf{x}) \equiv \sum_{j=1}^d f_j(x_j)$ , not the individual component functions  $f_j$ . We discuss the estimation of  $f_j$  later in [Section 2.3](#).

For simplicity, we consider Nadaraya-Watson type estimators of  $m_j$ ,  $p_j$  and  $p_{jk}$ . For a projection interpretation of SBF estimation, we use a normalized kernel scheme as described below. The projection interpretation is crucial for the success of SBF estimation.



Let  $K$  be a baseline symmetric, bounded and nonnegative kernel function supported on  $[-1, 1]$  such that  $\int K = 1$ . The conventional kernel weight scheme for the variable  $X_j$  based on  $K$  is to give the weight  $K_{h_j}(x - u) \equiv h_j^{-1} K((x - u)/h_j)$  to an observed value  $u$  of  $X_j$  locally at each point  $x_j \in [0, 1]$ , where  $h_j > 0$  is called the bandwidth and determines the degree of localization for  $X_j$ . The normalized kernel function based on  $K$  is defined by

$$(2-2) \quad K_{h_j}(x_j, u) = \left[ \int_0^1 K_{h_j}(v - u) dv \right]^{-1} K_{h_j}(x_j - u), \quad 0 \leq x_j, u \leq 1.$$

Then, it holds that  $K_{h_j}(x_j, u) = K_{h_j}(x_j - u)$  for all  $(x_j, u) \in [2h_j, 1 - 2h_j] \times [0, 1]$  or  $(x_j, u) \in [0, 1] \times [h_j, 1 - h_j]$ . Furthermore,

$$(2-3) \quad \begin{aligned} \int_0^1 K_{h_j}(x_j, u) dx_j &= 1, \quad \text{for all } u \in [0, 1], \\ \mu_{j,\ell}(x_j) &= \int_{-1}^1 t^\ell K(t) dt, \quad \text{for all } x_j \in [2h_j, 1 - 2h_j], \\ |\mu_{j,\ell}(x_j)| &\leq 2 \int_{-1}^1 |t|^\ell K(t) dt, \quad \text{for all } x_j \in [0, 1], \end{aligned}$$

where and below  $\mu_{j,\ell}(x_j) = \int_0^1 h_j^{-\ell} (x_j - u)^\ell K_{h_j}(x_j, u) du$ . We set  $I_j = [2h_j, 1 - 2h_j]$  and refer to them as interior regions.

We write  $X_{ij}$  for the  $j$ th entry of  $\mathbf{X}_i$ . With the normalized kernel function  $K_{h_j}(\cdot, \cdot)$  we estimate the marginal and joint densities by

$$\hat{p}_j(x_j) = n^{-1} \sum_{i=1}^n K_{h_j}(x_j, X_{ij}), \quad \hat{p}_{jk}(x_j, x_k) = n^{-1} \sum_{i=1}^n K_{h_j}(x_j, X_{ij}) K_{h_k}(x_k, X_{ik}).$$

Also, by Nadaraya-Watson smoothing we estimate  $m_j$  by

$$\hat{m}_j(x_j) = \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n K_{h_j}(x_j, X_{ij}) Y_i.$$

Plugging these estimators into (2-1) gives the following system of backfitting equations to solve for  $\hat{f} : \hat{f}(\mathbf{x}) \equiv \sum_{j=1}^d \hat{f}_j(x_j)$ .

$$(2-4) \quad \hat{f}_j(x_j) = \hat{m}_j(x_j) - \sum_{k \neq j}^d \int_0^1 \hat{f}_k(x_k) \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} dx_k, \quad 1 \leq j \leq d.$$

We call it *smooth backfitting equation*. The system of equations (2-4) can identify only the sum function  $\hat{f}(\mathbf{x}) = \sum_{j=1}^d \hat{f}_j(x_j)$  as we discuss in Section 2.2.

Define  $\hat{p}(\mathbf{x}) = n^{-1} \sum_{i=1}^n \prod_{j=1}^d K_{h_j}(x_j, X_{ij})$ , the estimator of the joint density  $p$ . Let  $\mathcal{H}(\hat{p})$  denote the space of additive functions  $g \in L_2(\hat{p})$  of the form  $g(\mathbf{x}) = g_1(x_1) + \dots + g_d(x_d)$  where  $g_j$  are univariate functions. Endowed with the inner product  $\langle g, \eta \rangle_n = \int g(\mathbf{x}) \eta(\mathbf{x}) \hat{p}(\mathbf{x}) d\mathbf{x}$ , it is a Hilbert space. By considering the Fréchet differentials of functionals defined on  $\mathcal{H}(\hat{p})$  and from the first property of (2-3), we may show that

$$(2-5) \quad \hat{f} = \arg \min_{g \in \mathcal{H}(\hat{p})} \int_{[0,1]^d} n^{-1} \sum_{i=1}^n (Y_i - g(\mathbf{x}))^2 \prod_{j=1}^d K_{h_j}(x_j, X_{ij}) d\mathbf{x}$$

whenever a solution  $\hat{f}$  of (2-4) exists and is unique. To solve the system of equations (2-4) the following iterative scheme is employed. First, initialize  $\hat{f}_j^{[0]}$  for  $1 \leq j \leq d$ . In the  $r$ th cycle of the iteration, update  $\hat{f}_j^{[r-1]}$  successively for  $1 \leq j \leq d$  by

$$(2-6) \quad \begin{aligned} \hat{f}_j^{[r]}(x_j) = & \hat{m}_j(x_j) - \sum_{1 \leq k \leq j-1} \int_0^1 \hat{f}_k^{[r]}(x_k) \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} dx_k \\ & - \sum_{j+1 \leq k \leq d} \int_0^1 \hat{f}_k^{[r-1]}(x_k) \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} dx_k. \end{aligned}$$

**2.2 Convergence of SBF algorithm.** Here, we discuss the existence and uniqueness of the solution of the backfitting Equation (2-4), and also the convergence of the backfitting Equation (2-6).

Consider the subspaces of  $L_2(\hat{p})$  defined by

$$L_2(\hat{p}_j) = \{g \in L_2(\hat{p}) : g(\mathbf{x}) = g_j(x_j) \text{ for some univariate function } g_j\}.$$

Let  $\hat{\pi}_j : L_2(\hat{p}) \rightarrow L_2(\hat{p}_j)$  denote projection operators such that

$$(2-7) \quad \hat{\pi}_j(g) = \int_{[0,1]^{d-1}} g(\mathbf{x}) \frac{\hat{p}(\mathbf{x})}{\hat{p}_j(x_j)} d\mathbf{x}_{-j},$$

where  $\mathbf{x}_{-j}$  for  $\mathbf{x}$  equals  $(x_1, \dots, x_{j-1}, x_{j+1}, \dots, x_d)$ . Then, the system of equations (2-4) can be written as

$$(2-8) \quad \hat{f} = (I - \hat{\pi}_j) \hat{f} + \hat{m}_j, \quad 1 \leq j \leq d,$$

where we have used the convention that  $\hat{m}_j(\mathbf{x}) = \hat{m}_j(x_j)$ . The equivalence between (2-4) and (2-8) follows from

$$(\hat{\pi}_j f_k)(\mathbf{x}) = \int_0^1 f_k(x_k) \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} dx_k, \quad 1 \leq j \neq k \leq d,$$

which holds due to the first property of (2-3). Put

$$\hat{m}_\oplus = \hat{m}_d + (I - \hat{\pi}_d)\hat{m}_{d-1} + (I - \hat{\pi}_d)(I - \hat{\pi}_{d-1})\hat{m}_{d-2} + \cdots + (I - \hat{\pi}_d)\cdots(I - \hat{\pi}_2)\hat{m}_1$$

and  $\hat{T} = (I - \hat{\pi}_d)\cdots(I - \hat{\pi}_1)$ . Note that  $\hat{T}$  is a linear operator that maps  $\mathcal{H}(\hat{p})$  to itself. A successive application of (2-8) for  $j = d, d-1, \dots, 2, 1$  gives

$$(2-9) \quad \hat{f} = \hat{T}\hat{f} + \hat{m}_\oplus.$$

If (2-9) has a solution  $\hat{f} \in \mathcal{H}(\hat{p})$ , then solving (2-9) is equivalent to solving (2-8) and thus  $\hat{f}$  is also a solution of (2-8). To see this, consider a version of  $\hat{T}$  for which the index  $j$  takes the role of the index  $d$ . Call it  $\hat{T}_j$ . Define a version of  $\hat{m}_\oplus$  accordingly and call it  $\hat{m}_{\oplus,j}$ . Then, it holds that  $\hat{\pi}_j\hat{T}_j = 0$  and  $\hat{\pi}_j\hat{m}_{\oplus,j} = \hat{m}_j$ . Suppose that there exists  $\hat{f} \in \mathcal{H}(\hat{p})$  that satisfies (2-9). If we exchange the roles of  $j$  and  $d$ , then the solution also satisfies  $\hat{f} = \hat{T}_j\hat{f} + \hat{m}_{\oplus,j}$ . Since this holds for all  $1 \leq j \leq d$ , we may conclude

$$\hat{\pi}_j\hat{f} = \hat{\pi}_j\hat{T}_j\hat{f} + \hat{\pi}_j\hat{m}_{\oplus,j} = 0 + \hat{m}_j, \quad 1 \leq j \leq d,$$

which is equivalent to (2-8).

The existence and uniqueness of the solution of (2-9) now follows if the linear operator  $\hat{T}$  is a contraction. An application of Proposition A.4.2 of [Bickel, Klaassen, Ritov, and Wellner \[1993\]](#) to the projection operators  $\hat{\pi}_j$  gives that  $\mathcal{H}(\hat{p})$  is a closed subspace of  $L_2(\hat{p})$  and  $\|\hat{T}\|_{\text{op}} < 1$ , under the condition that

$$(2-10) \quad \int_{[0,1]^2} \left[ \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)\hat{p}_k(x_k)} \right]^2 \hat{p}_j(x_j)\hat{p}_k(x_k) dx_j dx_k < \infty \quad \text{for all } 1 \leq j \neq k \leq d.$$

An analogue of (2-9) for the backfitting [Equation \(2-6\)](#) is

$$(2-11) \quad \hat{f}^{[r]} = \hat{T}\hat{f}^{[r-1]} + \hat{m}_\oplus.$$

Assuming (2-10), we get from (2-9) that  $\hat{f} = \sum_{j=1}^{\infty} \hat{T}^j \hat{m}_\oplus$ . This and the fact that  $\hat{T}\hat{f}^{[r-1]} + \hat{m}_\oplus = \hat{T}^r \hat{f}^{[0]} + \sum_{j=0}^{r-1} \hat{T}^j \hat{m}_\oplus$  give

$$(2-12) \quad \|\hat{f}^{[r]} - \hat{f}\|_{2,n} \leq \|\hat{T}\|_{\text{op}}^r \left( \|\hat{f}^{[0]}\|_{2,n} + \frac{1}{1 - \|\hat{T}\|_{\text{op}}} \cdot \|\hat{m}_\oplus\|_{2,n} \right),$$

where  $\|\cdot\|_{2,n}$  denote the induced norm of the inner product  $\langle \cdot, \cdot \rangle_n$  defined earlier. The following theorem is a non-asymptotic version of Theorem 1 of [Mammen, Linton, and J. Nielsen \[1999\]](#).

THEOREM 2.1. Assume the condition (2-10). Then, it holds that the solution of the system of equations (2-4) exists and is unique, and that the backfitting iteration (2-6) converges to the solution.

The condition (2-10) holds with probability tending to one if  $p_j$  are continuous and bounded away from zero on  $[0, 1]$  and  $p_{jk}$  are continuous and bounded above on  $[0, 1]^2$ . This follows since under these conditions there exists a constant  $0 < C < \infty$  such that

$$\sup_{(x_j, x_k) \in [0, 1]^2} \frac{\hat{p}_{jk}(x_j, x_k)^2}{\hat{p}_j(x_j)\hat{p}_k(x_k)} \leq C$$

with probability tending to one. Thus, we can deduce that

$$(2-13) \quad P \left( \lim_{r \rightarrow \infty} \|\hat{f}^{[r]} - \hat{f}\|_{2,n} = 0 \right) \rightarrow 1$$

as  $n \rightarrow \infty$ . Below, we give a stronger result than (2-13) owing to [Mammen, Linton, and J. Nielsen \[ibid.\]](#). We make the following assumptions to be used in the subsequent discussion.

- (C1) The joint densities  $p_{jk}$  are partially continuously differentiable and  $p$  is bounded away from zero and infinity on  $[0, 1]^d$ .
- (C2) The bandwidths satisfy  $h_j \rightarrow 0$  and  $nh_j h_k / \log n \rightarrow \infty$  as  $n \rightarrow \infty$  for all  $1 \leq j \neq k \leq d$ .
- (C3) The baseline kernel function  $K$  is bounded, has compact support  $[-1, 1]$ , is symmetric about zero and Lipschitz continuous.

Define an analogue of  $\mathcal{H}(\hat{p})$  as

$$\mathcal{H}(p) \equiv \{g \in L_2(p) : g(\mathbf{x}) = g_1(x_1) + \cdots + g_d(x_d), \text{ } g_j \text{ are univariate functions} \}$$

equipped with the inner product  $\langle g, \eta \rangle = \int g(\mathbf{x})\eta(\mathbf{x})p(\mathbf{x}) d\mathbf{x}$  and its induced norm  $\|\cdot\|_2$ . We note that  $P(\mathcal{H}(\hat{p}) = \mathcal{H}(p)) \rightarrow 1$  under the condition (C1)–(C3). This follows since the conditions imply that there exist absolute constants  $0 < c < C < \infty$  such that

$$c\|g_j\|_2 \leq \|g_j\|_{2,n} \leq C\|g_j\|_2$$

with probability tending to one. Now, define  $\pi_j$  as  $\hat{\pi}_j$  with  $\hat{p}$  and  $\hat{p}_j$  being replaced by  $p$  and  $p_j$ , respectively. Let  $T = (I - \pi_d) \cdots (I - \pi_1)$ . From (C1) we get that, for all  $1 \leq j \neq k \leq d$ ,

$$\int_{[0, 1]^2} \left[ \frac{p_{jk}(x_j, x_k)}{p_j(x_j)p_k(x_k)} \right]^2 p_j(x_j)p_k(x_k) dx_j dx_k < \infty,$$

so that  $T$  is also a contraction as a map from  $\mathcal{H}(p)$  to itself. Furthermore, another application of Proposition A.4.2 of [Bickel, Klaassen, Ritov, and Wellner \[1993\]](#) gives that there is an absolute constant  $0 < c < \infty$  such that for any  $g \in \mathcal{H}(p)$  there exists a decomposition  $g = g_1 + \cdots + g_d$  with

$$(2-14) \quad \|g\|_2 \geq c \sum_{j=1}^d \|g_j\|_2.$$

For such a decomposition and from successive applications of the Minkowski and Hölder inequalities, we get

$$\begin{aligned} \|(\hat{\pi}_j - \pi_j)g\|_2 &\leq \\ &\leq \sum_{k \neq j}^d \|g_k\|_2 \left( \int \left[ \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)p_k(x_k)} - \frac{p_{jk}(x_j, x_k)}{p_j(x_j)p_k(x_k)} \right]^2 p_j(x_j)p_k(x_k) dx_j dx_k \right)^{1/2}. \end{aligned}$$

Using this and (2-14), we may prove  $\|\hat{\pi}_j - \pi_j\|_{\text{op}} = o_p(1)$  for all  $1 \leq j \leq d$  and thus  $\|\hat{T} - T\|_{\text{op}} = o_p(1)$ . This proves that there exists a constant  $0 < \gamma < 1$  such that  $P(\|\hat{T}\|_{\text{op}} < \gamma) \rightarrow 1$  as  $n \rightarrow \infty$ . The following theorem is an asymptotic version of [Theorem 2.1](#).

**THEOREM 2.2.** ([Mammen, Linton, and J. Nielsen \[1999\]](#)). *Assume the conditions (C1)–(C3). Then, with probability tending to one, the solution of the system of equations (2-4) exists and is unique. Furthermore, there exists a constant  $0 < \gamma < 1$  such that*

$$\lim_{n \rightarrow \infty} P\left(\|\hat{f}^{[r]} - \hat{f}\|_2 \leq \gamma^r (\|\hat{f}^{[0]}\|_2 + (1 - \gamma)^{-1} \|\hat{m}_{\oplus}\|_2)\right) = 1.$$

**2.3 Estimation of individual component functions.** The component functions  $f_j$  in the model (1-1) are not identified, but only their sum  $f$  is. We need put constraints on  $f_j$  to identify them. There may be various constraints. We consider the constraints

$$(2-15) \quad \int_0^1 f_j(x_j) p_j(x_j) dx_j = 0, \quad 1 \leq j \leq d.$$

With the constraints at (2-15) the model (1-1) is rewritten as

$$(2-16) \quad f(\mathbf{x}) = \mu + f_1(x_1) + \cdots + f_d(x_d),$$

for  $\mu = E(Y)$ , and each  $f_j$  is uniquely determined. The latter follows from (2-14) and the fact that, for  $c_j = \int_0^1 g_j(x_j) p_j(x_j) dx_j$ , we get

$$\|g_j\|_2^2 = \|g_j - c_j\|_2^2 + |c_j|^2 \geq \|g_j - c_j\|_2^2.$$

For the estimators of  $f_j$  we consider the following constraint.

$$(2-17) \quad \int_0^1 \hat{f}_j(x_j) \hat{p}_j(x_j) dx_j = 0, \quad 1 \leq j \leq d.$$

For the estimation of  $f_j$  that satisfy (2-15), the backfitting Equation (2-4) and the backfitting Equation (2-6) are modified by simply putting  $\hat{m}_j - \bar{Y}$  in the place of  $\hat{m}_j$ , where  $\bar{Y}$  is used as an estimator of  $\mu$ . Then, we may prove that, with probability tending to one, there exists a solution  $(\hat{f}_j : 1 \leq j \leq d)$  of the resulting backfitting equation that satisfies the constraint (2-17). In this section, we discuss the asymptotic properties of the estimators  $\hat{f}_j$ . The error of  $\bar{Y}$  as an estimator of  $\mu$  is of magnitude  $O_p(n^{-1/2})$ , which is negligible compared to nonparametric rates. In the subsequent discussion in this section, we assume  $\mu = 0$  and ignore  $\bar{Y}$  in the backfitting equation, for simplicity.

Put  $\varepsilon_i = Y_i - \sum_{j=1}^d f_j(X_{ij})$  and

$$\begin{aligned} \hat{m}_j^A(x_j) &= \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n K_{h_j}(x_j, X_{ij}) \varepsilon_i, \\ \hat{m}_j^B(x_j) &= \hat{p}_j(x_j)^{-1} n^{-1} \sum_{i=1}^n K_{h_j}(x_j, X_{ij}) [f_j(X_{ij}) - f_j(x_j)], \\ \hat{m}_{jk}^C(x_j) &= n^{-1} \sum_{i=1}^n \int_0^1 [f_k(X_{ik}) - f_k(x_k)] K_{h_j}(x_j, X_{ij}) K_{h_k}(x_k, X_{ik}) dx_k. \end{aligned}$$

Then, from the backfitting Equation (2-4) we get

$$(2-18) \quad \begin{aligned} \hat{f}_j(x_j) - f_j(x_j) &= \hat{m}_j^A(x_j) + \hat{m}_j^B(x_j) + \hat{p}_j(x_j)^{-1} \sum_{k \neq j} \hat{m}_{jk}^C(x_j) \\ &\quad - \sum_{k \neq j} \int_0^1 [\hat{f}_k(x_k) - f_k(x_k)] \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} dx_k, \quad 1 \leq j \leq d. \end{aligned}$$

The above equation is a key to deriving stochastic expansions of  $\hat{f}_j$ . To analyze the three terms  $\hat{m}_j^A$ ,  $\hat{m}_j^B$  and  $\hat{m}_{jk}^C$  in (2-18), we make the following assumptions.

(C4) The component functions  $f_j$  are twice continuously differentiable.

(C5)  $E|Y|^\alpha < \infty$  for  $\alpha > 5/2$  and  $\text{var}(Y|X_j = \cdot)$  are continuous on  $[0, 1]$

We also assume that  $h_j$  are of order  $n^{-1/5}$ , which is known to be optimal in univariate smoothing.

We write  $\mu_\ell = \int_{-1}^1 t^\ell K(t) dt$  and recall the definition of  $\mu_{j,\ell}$  given immediately after (2-3). Let  $r_j$  denote a generic sequence of stochastic terms corresponding to  $\hat{m}_j$  such that

$$(2-19) \quad \sup_{x_j \in [2h_j, 1-2h_j]} |r_j(x_j)| = o_p(n^{-2/5}), \quad \sup_{x_j \in [0,1]} |r_j(x_j)| = O_p(n^{-2/5}).$$

Using the conditions (C1), (C3) and (C4) we may verify that, for  $1 \leq j \neq k \leq d$ ,

$$(2-20) \quad \begin{aligned} \hat{m}_j^B(x_j) &= h_j \frac{\mu_{j,1}(x_j)}{\mu_{j,0}(x_j)} f'_j(x_j) + h_j^2 \mu_2 f'_j(x_j) \frac{p'_j(x_j)}{p_j(x_j)} + \frac{1}{2} h_j^2 \mu_2 f''_j(x_j) + r_j(x_j), \\ \frac{\hat{m}_{jk}^C(x_j)}{\hat{p}_j(x_j)} &= h_k^2 \mu_2 \int_0^1 f'_k(x_k) \frac{\partial p_{jk}(x_j, x_k) / \partial x_k}{p_j(x_j)} dx_k \\ &\quad + \int_0^1 \left[ h_k \frac{\mu_{k,1}(x_k)}{\mu_{k,0}(x_k)} f'_k(x_k) + \frac{1}{2} h_k^2 \mu_2 f''_k(x_k) \right] \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} dx_k + r_j(x_j). \end{aligned}$$

Define

$$(2-21) \quad \begin{aligned} \tilde{\Delta}_j(x_j) &= h_j^2 \mu_2 f'_j(x_j) \frac{p'_j(x_j)}{p_j(x_j)} + \sum_{k \neq j} h_k^2 \mu_2 \int_0^1 f'_k(x_k) \frac{\partial p_{jk}(x_j, x_k) / \partial x_k}{p_j(x_j)} dx_k, \\ \hat{\Delta}_j(x_j) &= \hat{f}_j(x_j) - f_j(x_j) - \hat{m}_j^A(x_j) - h_j \frac{\mu_{j,1}(x_j)}{\mu_{j,0}(x_j)} f'_j(x_j) - \frac{1}{2} h_j^2 \mu_2 f''_j(x_j). \end{aligned}$$

Then, the equations at (2-18) and the expansions at (2-20) give

$$(2-22) \quad \hat{\Delta}_j(x_j) = \tilde{\Delta}_j(x_j) - \sum_{k \neq j} \int \hat{\Delta}_k(x_k) \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} dx_k + r_j(x_j),$$

where we have used

$$\int \hat{m}_k^A(x_k) \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} dx_k = o_p(n^{-2/5})$$

uniformly for  $x_j \in [0, 1]$ .

Now, we consider a system of equations for  $\hat{D} \in \mathcal{H}(\hat{p})$ ,

$$(2-23) \quad \hat{D}_j(x_j) = \tilde{\Delta}_j(x_j) - \sum_{k \neq j} \int \hat{D}_k(x_k) \frac{\hat{p}_{jk}(x_j, x_k)}{\hat{p}_j(x_j)} dx_k, \quad 1 \leq j \leq d.$$

Arguing as in Section 2.2, solving this is equivalent to solving  $\hat{D} = \hat{T} \hat{D} + \tilde{\Delta}_\oplus$ , where  $\tilde{\Delta}_\oplus$  is defined as  $\hat{m}_\oplus$  with  $\tilde{\Delta}_j$  taking the roles of  $\hat{m}_j$ . Similarly, solving (2-22) is equivalent to

solving for  $\hat{\Delta} \in \mathcal{H}(\hat{p})$  such that  $\hat{\Delta} = \hat{T}\hat{\Delta} + \tilde{\Delta}_{\oplus} + r_{\oplus}$  with  $r_{\oplus}$  being defined accordingly. Then, under the condition (2-10) or with probability tending to one under the condition (C1) it holds that  $\hat{\Delta} = \hat{D} + \sum_{r=0}^{\infty} \hat{T}^r r_{\oplus}$ . Since  $\hat{\pi}_j r_k = o_p(n^{-2/5})$  uniformly over  $[0, 1]$ , we get  $r_{\oplus} = r_+$  with the generic  $r_+$  such that  $r_+(\mathbf{x}) = \sum_{j=1}^d r_j(x_j)$  for some  $r_j$  that satisfy (2-19). Also, from the observation that  $(I - \hat{\pi}_j) \cdots (I - \hat{\pi}_1) r_j = o_p(n^{-2/5})$  uniformly over  $[0, 1]$  for all  $1 \leq j \leq d$ , we have

$$\sum_{r=0}^{\infty} \hat{T}^r r_{\oplus} = r_{\oplus} + \sum_{r=1}^{\infty} \hat{T}^r r_{\oplus} = r_+.$$

This proves

$$(2-24) \quad \hat{\Delta} = \hat{D} + r_+.$$

To identify the limit of  $\hat{D}$  we consider the system of integral equations for  $\Delta \in \mathcal{H}(p)$ ,

$$(2-25) \quad \Delta_j(x_j) = \tilde{\Delta}_j(x_j) - \sum_{k \neq j} \int \Delta_k(x_k) \frac{p_{jk}(x_j, x_k)}{p_j(x_j)} dx_k, \quad 1 \leq j \leq d.$$

Again, arguing as in Section 2.2, solving (2-25) is equivalent to solving  $\Delta = T\Delta + \Delta_{\oplus}$ , where  $\Delta_{\oplus}$  is defined as  $\tilde{\Delta}_{\oplus}$  but with  $\hat{\pi}_j$  being replaced by  $\pi_j$ . Since  $\|T\|_{\text{op}} < 1$  under (C1), the latter equation has a unique solution  $\Delta = \sum_{r=0}^{\infty} T^r \Delta_{\oplus}$ . A careful analysis of the operators  $T$  and  $\hat{T}$  gives that  $(\hat{T} - T) \sum_{r=1}^{\infty} \hat{T}^{r-1} \tilde{\Delta}_{\oplus} = r_+$  and that

$$T \sum_{r=2}^{\infty} \sum_{j=0}^{r-2} T^j (\hat{T} - T) \hat{T}^{r-2-j} \tilde{\Delta}_{\oplus} = o_p(n^{-2/5}),$$

$$\sum_{r=1}^{\infty} T^r (\tilde{\Delta}_{\oplus} - \Delta_{\oplus}) = o_p(n^{-2/5})$$

uniformly over  $[0, 1]^d$ . From these calculations it follows that  $\hat{D} = \Delta + r_+$ . This with (2-24) entails

$$(2-26) \quad \hat{\Delta} = \Delta + r_+.$$

To get expansions for each component  $\hat{f}_j$  satisfying the constraint (2-17), we put the following constraints on  $\Delta_j$ .

$$(2-27) \quad \int \Delta_j(x_j) p_j(x_j) dx_j = \mu_2 h_j^2 \int f'_j(x_j) p'_j(x_j) dx_j, \quad 1 \leq j \leq d.$$

Then, using (2-17) and (2-26) with the definition of  $\hat{\Delta}_j$  at (2-21), we may prove  $\hat{\Delta}_j = \Delta_j + r_j$  for  $1 \leq j \leq d$ , establishing the following theorem.



**THEOREM 2.3.** *Assume that the conditions (C1)–(C5) and that the bandwidths  $h_j$  are asymptotic to  $n^{-1/5}$ . Then,*

$$\hat{f}_j(x_j) = f_j(x_j) + \hat{m}_j^A(x_j) + h_j \frac{\mu_{j,1}(x_j)}{\mu_{j,0}(x_j)} f'_j(x_j) + \frac{1}{2} h_j^2 \mu_{j,2} f''_j(x_j) + \Delta_j(x_j) + r_j(x_j),$$

where  $r_j$  satisfy (2-19).

For fixed  $x_j \in (0, 1)$ , all  $\mu_{j,1}(x_j) = 0$  for sufficiently large  $n$ , and  $(nh_j)^{1/2} \hat{m}_j^A(x_j)$  are asymptotically normal with mean zero and variance

$$\text{var}(Y|X_j = x_j) p_j(x_j)^{-1} \int K^2(u) du$$

. Thus, the asymptotic distributions of  $(nh_j)^{1/2}(\hat{f}_j(x_j) - f_j(x_j))$  of  $\hat{f}_j$  are readily obtained from the stochastic expansion in the above theorem.

Although we have not discussed here, [Mammen, Linton, and J. Nielsen \[1999\]](#) also developed a local linear version of the smooth backfitting technique. However, the original proposal does not have easy interpretation as the Nadaraya-Watson estimator that we have discussed, and its implementation is more complex than the latter. [Mammen and Park \[2006\]](#) suggested a new smooth backfitting estimator that has the simple structure of the Nadaraya-Watson estimator while maintaining the nice asymptotic properties of the local linear smooth backfitting estimator.

**2.4 Bandwidth selection and related models.** In nonparametric function estimation, selection of smoothing parameters is essential for the accuracy of the estimation. It is well known that one should not choose these tuning parameters by minimizing a measure of fit, such as the residual sum of squares  $n^{-1} \sum_{i=1}^n (Y_i - \hat{f}(\mathbf{X}_i))^2$ , since it tends to choose  $h_j$  that give ‘overfitting’. [Mammen and Park \[2005\]](#) tackled this problem by deriving higher-order stochastic expansions of the residual sum of squares and proposed a penalized least squares method of choosing  $h_j$ . They also proposed two plug-in bandwidth selectors that rely on expansions of the average square errors  $n^{-1} \sum_{i=1}^n (\hat{f}(\mathbf{X}_i) - f(\mathbf{X}_i))^2$ . [J. P. Nielsen and Sperlich \[2005\]](#) considered a cross-validated bandwidth selector and discussed some other practical aspects of the smooth backfitting algorithm.

A very important extension of the additive mean regression model at (1-1) or (2-16) is to a generalized additive model,

$$(2-28) \quad g(E(Y|\mathbf{X} = \mathbf{x})) = f_1(x_1) + \cdots + f_d(x_d),$$

where  $g$  is a known link function. This model accommodates discrete-type responses  $Y$  such as Bernoulli and Poisson random variables. [Yu, Park, and Mammen \[2008\]](#) extended

the idea of smooth backfitting to generalized additive models. The estimation of the additive function  $f = f_1 + \cdots + f_d$  based on observations of  $(\mathbf{X}, Y)$  involves a nonlinear optimization problem due to the presence of the link  $g$ . To resolve the difficulty, [Yu, Park, and Mammen \[ibid.\]](#) introduced the so called ‘smoothed likelihood’ and studied an innovative idea of double iteration to maximize the smoothed likelihood. They proved that the double iteration algorithm converges and developed a complete theory for the smooth backfitting likelihood estimators of  $f_j$ .

Varying coefficient models are another important class of structured nonparametric regression models. The models arise in many real applications, see [Hastie and Tibshirani \[1993\]](#), [Yang, Park, Xue, and Härdle \[2006\]](#) and [Park, Mammen, Y. K. Lee, and E. R. Lee \[2015\]](#). Their structure is similar to classical linear models, but they are more flexible since the regression coefficients are allowed to be functions of other predictors. There are two types of varying coefficient models that have been studied most. One type is to let all regression coefficients depend on a single predictor, say  $Z$ :  $E(Y|\mathbf{X} = \mathbf{x}, Z = z) = f_1(z)x_1 + \cdots + f_d(z)x_d$ . The estimation of this type of models is straightforward. For each given  $z$ , we may estimate  $\mathbf{f}(z) \equiv (f_1(z), \dots, f_d(z))$  by

$$\mathbf{f}(z) = \arg \min_{(\theta_1, \dots, \theta_d) \in \mathbb{R}^d} \sum_{i=1}^n \left( Y_i - \sum_{j=1}^d \theta_j X_{ij} \right)^2 K_h(z, Z_i).$$

There have been a large body of literature on this model, see [Fan and W. Zhang \[1999\]](#) and [Fan and W. Zhang \[2000\]](#), for example. The second type is to let different regression coefficients be functions of different predictors, say  $\mathbf{Z} \equiv (Z_1, \dots, Z_d)$ :

$$(2-29) \quad E(Y|\mathbf{X} = \mathbf{x}, \mathbf{Z} = \mathbf{z}) = f_1(z_1)x_1 + \cdots + f_d(z_d)x_d.$$

Fitting the model (2-29) is completely different from fitting the first type. The standard kernel smoothing that minimizes

$$\sum_{i=1}^n \left( Y_i - \sum_{j=1}^d \theta_j X_{ij} \right)^2 \prod_{j=1}^d K_{h_j}(z_j, Z_{ij})$$

for each  $\mathbf{z}$  would give multivariate function estimators of  $f_j(z_j)$  that also depend on other values of predictors  $z_k$  for  $k \neq j$ . [Yang, Park, Xue, and Härdle \[2006\]](#) studied the estimation of the latter model based on the marginal integration technique. Later, [Y. K. Lee, Mammen, and Park \[2012b\]](#) extended the idea of smooth backfitting to estimating the model.

Two limitations in the application of the model (2-29) are that the number of predictors  $X_j$  should be the same as that of  $Z_j$  and that in a modeling stage it is rather difficult

to determine which predictors we choose to be the ‘smoothing variables’  $Z_j$  and which to be ‘regressors’  $X_j$ . [Y. K. Lee, Mammen, and Park \[2012a\]](#) removed the limitations completely by studying a very general form of varying coefficient models. With a link function  $g$  and a given set of  $d$  predictors, they introduced the model

$$(2-30) \quad g(E(Y|\mathbf{X} = \mathbf{x})) = x_1 \left( \sum_{k \in I_1} f_{1k}(x_k) \right) + \cdots + x_q \left( \sum_{k \in I_q} f_{qk}(x_k) \right),$$

where  $q \leq d$  and the index sets  $I_j$  are known subsets of  $\{1, \dots, d\}$  and allowed to overlap with each other, but not to include  $j$ . If each  $I_j$  consists of a single index different from each other, then (2-30) reduces to the model (2-29), while taking  $X_1 \equiv 1$ ,  $q = 1$  and  $I_1 = \{2, \dots, d\}$  gives the generalized additive model (2-28). [Y. K. Lee, Mammen, and Park \[ibid.\]](#) proved that the component functions  $f_{jk}$  are identifiable under weak conditions, developed a powerful technique of fitting the model and presented its theory.

Other related works include [Y. K. Lee, Mammen, and Park \[2010\]](#), [Y. K. Lee, Mammen, and Park \[2014\]](#), [Yu, Mammen, and Park \[2011\]](#) and [Y. K. Lee \[2017\]](#), to list a few. Among them, [Y. K. Lee, Mammen, and Park \[2010\]](#) considered the estimation of additive quantile models,  $Y = f_1(X_1) + \cdots + f_d(X_d) + \varepsilon$ , where  $\varepsilon$  satisfies  $P(\varepsilon \leq 0|\mathbf{X}) = \alpha$  for  $0 < \alpha < 1$ . They successfully explored the theory for both the ordinary and smooth backfitting by devising a theoretical mean regression model under which the least squares ordinary and smooth backfitting estimators are asymptotically equivalent to the corresponding quantile estimators under the original model. [Y. K. Lee, Mammen, and Park \[2014\]](#) further extended the idea to the estimation of varying coefficient quantile models. [Yu, Mammen, and Park \[2011\]](#) considered a partially linear additive model. They derived the semiparametric efficiency bound in the estimation of the parametric part of the model and proposed a semiparametric efficient estimator based on smooth backfitting estimation of the additive nonparametric part. Finally, [Y. K. Lee \[2017\]](#) studied the estimation of bivariate additive regression models based on the idea of smooth backfitting.

### 3 Errors-in-variable additive models

In this section we consider the situation where the predictors  $X_j$  are not directly observed in the additive model (1-1), but contaminated  $Z_j = X_j + U_j$  with measurement errors  $U_j$  are. Many people worked on errors-in-variables problems in nonparametric density and regression estimation. A few notable examples include [Carroll and Hall \[1988\]](#), [Stefanski and Carroll \[1990\]](#), [Fan and Truong \[1993\]](#), [Delaigle, Hall, and Meister \[2008\]](#), [Delaigle, Fan, and Carroll \[2009\]](#), [Delaigle and Hall \[2016\]](#) and [Han and Park \[2018\]](#). Among them, [Han and Park \[ibid.\]](#) is considered as the first attempt dealing with errors-in-variables in

structured nonparametric regression. In this section we outline the work of [Han and Park \[ibid.\]](#) on the model (1-1) and discuss its extensions.

**3.1 Normalized deconvolution kernel.** Suppose that we observe  $Z_{ij} = X_{ij} + U_{ij}$  for  $1 \leq i \leq n$  and  $1 \leq j \leq d$ , where we assume that  $\mathbf{U}_i \equiv (U_{i1}, \dots, U_{id})$  are independent of  $\mathbf{X}_i$ . We also assume that  $U_{ij}$  are independent and have known densities. Write  $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{id})$ . The task is to estimate the mean regression function  $f = E(Y|\mathbf{X} = \cdot)$  with the additive structure  $f(\mathbf{x}) = f_1(x_1) + \dots + f_d(x_d)$  using the contaminated data  $(\mathbf{Z}_i, Y_i)$ ,  $1 \leq i \leq n$ . The very core of the difficulty is that the observed responses  $Y_i$  for  $\mathbf{Z}_i$  near a point of interest, say  $\mathbf{x}$ , may not contain relevant information about the true function  $f(\mathbf{x})$  because of the measurement errors  $\mathbf{U}_i \equiv (U_{i1}, \dots, U_{id})$ . Thus, local smoothing of  $Y_i$  with a conventional kernel weighting scheme that acts on  $\mathbf{Z}_i$  fails.

In the estimation of a density  $p_0$  of a random variable  $X$  taking values in  $\mathbb{R}$ , one uses a special kernel scheme to effectively deconvolute irrelevant information contained in the contaminated  $Z = X + U$ . For a baseline kernel function  $K \geq 0$ , define

$$(3-1) \quad \tilde{K}_h(u) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itu} \frac{\phi_K(ht)}{\phi_U(t)} dt,$$

where  $\phi_W$  for a random variable  $W$  denotes the characteristic function of  $W$ . This kernel has the so called ‘unbiased scoring’ property that

$$(3-2) \quad E(\tilde{K}_h(x - Z)|X) = K_h(x - X).$$

The property (3-2) basically tells that the bias of the deconvolution kernel density estimator  $\hat{p}_0(x) = n^{-1} \sum_{i=1}^n \tilde{K}_h(x - Z_i)$  is the same as the ‘oracle’ estimator  $\hat{p}_{0,\text{ora}}(x) = n^{-1} \sum_{i=1}^n K_h(x - X_i)$  that is based on unobservable  $X_i$  and the conventional kernel scheme  $K_h$ .

In [Section 2](#) we have seen that the first property of (2-3) is essential in the estimation of the additive model (1-1). One may think of normalizing the deconvolution kernel as defined at (3-1) as in (2-2) with  $\tilde{K}_{h_j}(x_j - u)$  taking the role of  $K_{h_j}(x_j - u)$ . But, it turns out that the resulting kernel violates the corresponding version of the unbiased scoring property (3-2). [Han and Park \[ibid.\]](#) noted that

$$\tilde{K}_{h_j}(x_j - z) = \frac{1}{2\pi h_j} \int_{-\infty}^{\infty} e^{-it(x_j - z)/h_j} \frac{\varphi_K(t, x_j, h_j)}{\phi_{U_j}(t/h_j)} dt,$$

where  $\varphi_K(t, x_j; h_j) = \int_0^1 e^{it(x_j - v)/h_j} K_{h_j}(x_j - v) dv$ . The basic idea was then to replace  $K_{h_j}(x_j - \cdot)$  in the definition of  $\varphi_K(t, x_j; h_j)$  by the normalized kernel  $K_{h_j}(x_j, \cdot)$  as defined at (2-2). The resulting kernel is not well-defined, however, for  $x_j$  on the boundary

region  $[0, h_j] \cup (1 - h_j, 1]$ . To remedy this, [Han and Park \[2018\]](#) proposed a new kernel scheme  $\tilde{K}_{h_j}^*$  defined by

$$(3-3) \quad \tilde{K}_{h_j}^*(x_j, z) = \frac{1}{2\pi h_j} \int_{-\infty}^{\infty} e^{-it(x_j - z)/h_j} \frac{\phi_K(t, x_j; h_j) \phi_K(t)}{\phi_{U_j}(t/h_j)} dt,$$

where  $\phi_K(t, x_j; h_j) = \int_0^1 e^{it(x_j - v)/h_j} K_{h_j}(x_j, v) dv$ . [Han and Park \[ibid.\]](#) proved that  $\tilde{K}_{h_j}^*$  has both the properties of normalization and unbiased scoring under the following condition (A1). Let  $\lfloor \gamma \rfloor$  denote the largest integer that is less than or equal to  $\gamma$ , and  $K^{(\ell)}$  the  $\ell$ -th derivative of  $K$ .

(A1) There exist constants  $\beta \geq 0$  and  $0 < c < C < \infty$  such that  $c(1 + |t|)^{-\beta} \leq |\phi_{U_j}(t)| \leq C(1 + |t|)^{-\beta}$  for all  $t \in \mathbb{R}$  and for all  $1 \leq j \leq d$ . For such constant  $\beta$  the baseline kernel  $K$  is  $\lfloor \beta + 1 \rfloor$ -times continuously differentiable and  $K^{(\ell)}(-1) = K^{(\ell)}(1) = 0$  for  $0 \leq \ell \leq \lfloor \beta \rfloor$ .

**THEOREM 3.1.** ([Han and Park \[ibid.\]](#)). *Under the conditions (A1) and (C3), the integral in (3-3) exists for all  $x_j \in [0, 1]$  and  $z \in \mathbb{R}$ . Furthermore,  $\int_0^1 \tilde{K}_{h_j}^*(x_j, z) dx_j = 1$  for all  $z \in \mathbb{R}$  and*

$$\mathbb{E} \left( \tilde{K}_{h_j}^*(x_j, Z_j) \mid X_j = u_j \right) = K_{h_j}(x_j, \cdot) * K_{h_j}(u_j) \text{ for all } x_j, u_j \in [0, 1].$$

**3.2 Theory of smooth backfitting.** With the normalized and smoothed deconvolution kernel  $\tilde{K}_{h_j}^*$  introduced in [Section 3.1](#), we simply replace  $\hat{p}_j$ ,  $\hat{p}_{jk}$  and  $\hat{m}_j$  in (2-4), respectively, by

$$\begin{aligned} \hat{p}_j^*(x_j) &= n^{-1} \sum_{i=1}^n \tilde{K}_{h_j}^*(x_j, Z_{ij}), \\ \hat{p}_{jk}^*(x_j, x_k) &= n^{-1} \sum_{i=1}^n \tilde{K}_{h_j}^*(x_j, Z_{ij}) \tilde{K}_{h_k}^*(x_k, Z_{ik}), \\ \hat{m}_j^*(x_j) &= \hat{p}_j^*(x_j)^{-1} n^{-1} \sum_{i=1}^n \tilde{K}_{h_j}^*(x_j, Z_{ij}) Y_i. \end{aligned}$$

Define  $\hat{p}^*(\mathbf{x}) = n^{-1} \sum_{i=1}^n \prod_{j=1}^d \tilde{K}_{h_j}^*(x_j, X_{ij})$  and  $\hat{\pi}_j^*$  as  $\hat{\pi}_j$  at (2-7) with  $\hat{p}$  and  $\hat{p}_j$  being replaced by  $\hat{p}^*$  and  $\hat{p}_j^*$ , respectively. Let  $\hat{T}^* = (I - \hat{\pi}_d^*) \cdots (I - \hat{\pi}_1^*)$ . We can express the resulting backfitting equation as equations

$$(3-4) \quad \hat{f}^* = (I - \hat{\pi}_j^*) \hat{f}^* + \hat{m}_j^*, \quad 1 \leq j \leq d.$$

As we argued in [Section 2.2](#), solving this system of equations is equivalent to solving  $\hat{f}^* = \hat{T}^* \hat{f}^* + \hat{m}_\oplus^*$ , where  $\hat{m}_\oplus^*$  is defined as  $\hat{m}_\oplus$  with  $\hat{\pi}_j$  and  $\hat{m}_j$  being replaced by  $\hat{\pi}_j^*$  and  $\hat{m}_j^*$ , respectively. The corresponding version of the backfitting algorithm as at (2-11) is given by  $\hat{f}^{*[r]} = \hat{T}^* \hat{f}^{*[r-1]} + \hat{m}_\oplus^*$ ,  $r \geq 1$ . It holds that  $\hat{f}^{*[r]}$  converges to  $\hat{f}^*$  as  $r \rightarrow \infty$  under the condition that

(3-5)

$$\int_{[0,1]^2} \left[ \frac{\hat{p}_{jk}^*(x_j, x_k)}{\hat{p}_j^*(x_j) \hat{p}_k^*(x_k)} \right]^2 \hat{p}_j^*(x_j) \hat{p}_k^*(x_k) dx_j dx_k < \infty \quad \text{for all } 1 \leq j \neq k \leq d.$$

An analogue of [Theorem 2.2](#) also holds. We make the following additional assumptions for this.

(A2) For the constant  $\beta \geq 0$  in the condition (A1),  $|t^{\beta+1} \phi'_{U_j}(t)| = O(1)$  as  $|t| \rightarrow \infty$  and  $\int |t^\beta \phi_K(t)| dt < \infty$ .

(A3) For the constant  $\beta \geq 0$  in the condition (A1),  $h_j \rightarrow 0$  and  $n(h_j h_k)^{1+2\beta} / \log n \rightarrow \infty$  as  $n \rightarrow \infty$  for all  $1 \leq j \neq k \leq d$ .

**THEOREM 3.2.** ([Han and Park \[ibid.\]](#)). *Assume the conditions (C1), (C3) and (A1)–(A3). Then, with probability tending to one, the solution of the system of equations (3-4) exists and is unique. Furthermore, there exists a constant  $0 < \gamma < 1$  such that*

$$\lim_{n \rightarrow \infty} P\left(\|\hat{f}^{*[r]} - \hat{f}^*\|_2 \leq \gamma^r (\|\hat{f}^{*[0]}\|_2 + (1 - \gamma)^{-1} \|\hat{m}_\oplus^*\|_2)\right) = 1.$$

Now we discuss the asymptotic properties of  $\hat{f}^*$  and its components. To identify the individual components  $\hat{f}_j^*$ , we use the constraints

$$(3-6) \quad \int_0^1 \hat{f}_j^*(x_j) \hat{p}_j^*(x_j) dx_j = 0, \quad 1 \leq j \leq d.$$

As in [Section 2](#), we assume  $EY = 0$  for simplicity so that  $f(\mathbf{x}) = f_1(x_1) + \cdots + f_d(x_d)$  with  $f_j$  satisfying the constraints (2-15). We also set  $h_j \asymp h$ .

The asymptotic analysis of  $\hat{f}_j^*$  is much more complex than in the case of no measurement error. To explain the main technical challenges, we note that

$$(3-7) \quad \hat{f}_j^* - f_j = \hat{\delta}_j - \sum_{k \neq j} \hat{\pi}_k^*(\hat{f}_k^* - f_k), \quad 1 \leq j \leq d,$$

where  $\hat{\delta}_j = \hat{m}_j^* - \hat{\pi}_j^*(f)$ . Since

$$\pi_j(f) = \int_{[0,1]^{d-1}} E(Y|\mathbf{X} = \mathbf{x}) \frac{p(\mathbf{x})}{p_j(x_j)} d\mathbf{x}_{-j} = E(Y|X_j = x_j) = m_j,$$

$\hat{\delta}_j$  basically represent the errors of  $\hat{m}_j^*$  as an estimator of  $m_j$ . Each  $\hat{\delta}_j$  corresponds to  $\hat{m}_j^A + \hat{m}_j^B + \hat{p}_j^{-1} \sum_{k \neq j} \hat{m}_{jk}^C$  in the no measurement error case. Consider the same decomposition  $\hat{\delta}_j = \hat{m}_j^{*A} + \hat{m}_j^{*B} + \hat{p}_j^{*-1} \sum_{k \neq j} \hat{m}_{jk}^{*C}$ , where  $\hat{m}_j^{*A}$ ,  $\hat{m}_j^{*B}$  and  $\hat{p}_j^{*-1} \sum_{k \neq j} \hat{m}_{jk}^{*C}$  are defined in the same way as  $\hat{m}_j^A$ ,  $\hat{m}_j^B$  and  $\hat{p}_j^{-1} \sum_{k \neq j} \hat{m}_{jk}^C$ , respectively, with  $K_{h_j}(x_j, X_{ij})$  and  $K_{h_k}(x_k, X_{ik})$  being replaced by  $\tilde{K}_{h_j}^*(x_j, Z_{ij})$  and  $\tilde{K}_{h_k}^*(x_k, Z_{ik})$ , respectively. We have seen in [Section 2](#) that the error components  $\hat{m}_j^B$  and  $\hat{m}_{jk}^C$  of  $\hat{m}_j - \hat{\pi}_j(f)$  are spread, through the backfitting operation, into the errors of the other component function estimators  $\hat{f}_k, k \neq j$ , to the first order. In the present case, the errors are of two types. One type is for the replacement of  $K_{h_j}$  with  $X_{ij}$  by  $\tilde{K}_{h_j}^*$  with contaminated  $Z_{ij}$ , and the other is for those one would have when one uses  $K_{h_j}$  with  $X_{ij}$  in the estimation of  $f$ . The analysis of the first type is more involved. It has an additional complexity that we need to analyze whether the first type of errors in  $\hat{m}_j^{*B}$  and  $\hat{m}_{jk}^{*C}$  are spread into the errors of  $\hat{f}_k^*$  for  $k \neq j$ , through the backfitting operation.

[Han and Park \[2018\]](#) solved this problem and proved the following theorem. To state the theorem, let

$$\tau_n(\beta) = \begin{cases} 1 & \beta < 1/2 \\ \sqrt{\log h^{-1}} & \beta = 1/2 \\ h^{1/2-\beta} & \beta > 1/2. \end{cases}$$

Also, let  $r_j^*$  be generic stochastic terms such that

$$\sup_{x_j \in [2h_j, 1-2h_j]} |r_j^*(x_j)| = o_p(h^2), \quad \sup_{x_j \in [0, 1]} |r_j^*(x_j)| = O_p(h^2).$$

**THEOREM 3.3.** ([Han and Park \[ibid.\]](#)). Assume the conditions (C1), (C3)–(C5), (A1) and (A2). Assume also that  $nh^{3+4\beta}/\log n$  is bounded away from zero. Then, uniformly for  $x_j \in [0, 1]$ ,

$$\begin{aligned} \hat{f}_j^*(x_j) &= f_j(x_j) + h_j \frac{\mu_{1,j}(x_j)}{\mu_{0,j}(x_j)} f_j'(x_j) + \frac{1}{2} h_j^2 \mu_{2,j} f_j''(x_j) + \Delta_j(x_j) \\ &\quad + r_j^*(x_j) + O_p \left( \sqrt{\frac{\log n}{nh^{1+2\beta}}} \cdot \tau_n(\beta) \right), \quad 1 \leq j \leq d, \end{aligned}$$

where  $\Delta_j$  are the same as those in [Theorem 2.3](#).

The rates of convergence of  $\hat{f}_j^*$  to their targets  $f_j$  are readily obtained from [Theorem 3.3](#). For example, in case  $\beta < 1/2$  we may get

$$\sup_{x_j \in [2h_j, 1-2h_j]} |\hat{f}_j^*(x_j) - f_j(x_j)| = O_p \left( n^{-2/(5+2\beta)} \sqrt{\log n} \right),$$

$$\sup_{x_j \in [0,1]} |\hat{f}_j^*(x_j) - f_j(x_j)| = O_p \left( n^{-1/(5+2\beta)} \right)$$

by choosing  $h \asymp n^{-1/(5+2\beta)}$ . The uniform rate in the interior is known to be the optimal rate that one can achieve in one-dimensional deconvolution problems, see [Fan \[1991\]](#). For other cases where  $\beta \geq 1/2$ , see Corollary 3.5 of [Han and Park \[2018\]](#).

**3.3 Extension to partially linear additive models.** In this subsection we consider the model

$$(3-8) \quad Y = \boldsymbol{\theta}^\top \mathbf{X} + f_1(Z_1) + \cdots + f_d(Z_d) + \varepsilon,$$

where  $\varepsilon$  is independent of the predictor vectors  $\mathbf{X} \equiv (X_1, \dots, X_p)^\top$  and  $\mathbf{Z} \equiv (Z_1, \dots, Z_d)^\top$ ,  $\boldsymbol{\theta}$  are unknown and  $f_j$  are unknown univariate functions. We do not observe  $\mathbf{X}$  and  $\mathbf{Z}$ , but the contaminated  $\mathbf{X}^* = \mathbf{X} + \mathbf{U}$  and  $\mathbf{Z}^* = \mathbf{Z} + \mathbf{V}$  for measurement error vectors  $\mathbf{U}$  and  $\mathbf{V}$ . We assume that  $\varepsilon$  is also independent of  $(\mathbf{U}, \mathbf{V})$ ,  $\mathbf{U}$  has mean zero and a known variance  $\boldsymbol{\Sigma}_\mathbf{U}$  and is independent of  $\mathbf{V}$ ,  $V_j$  are independent across  $j$  and have known densities, and  $(\mathbf{U}, \mathbf{V})$  is independent of  $\mathbf{X}$  and  $\mathbf{Z}$ . Below, we outline the work of [E. R. Lee, Han, and Park \[2018\]](#) that studies the estimation of  $\boldsymbol{\theta}$  and  $f_j$  in the model (3-8) based on independent and identically distributed observations  $(\mathbf{X}_i^*, \mathbf{Z}_i^*, Y_i)$ ,  $1 \leq i \leq n$ .

Let  $\mathcal{H}$  be the space of square integrable functions  $g : \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $g(\mathbf{z}) = g_1(z_1) + \cdots + g_d(z_d)$ . Let  $\Pi(\cdot|\mathcal{H})$  denote the projection operator onto  $\mathcal{H}$ . Define  $\xi = \Pi(E(Y|\mathbf{Z} = \cdot)|\mathcal{H})$  and  $\eta_j = \Pi(E(X_j|\mathbf{Z} = \cdot)|\mathcal{H})$ ,  $1 \leq j \leq p$ . We write  $\boldsymbol{\eta} = (\eta_1, \dots, \eta_p)^\top$ . Under the condition that  $\mathbf{D} := E(\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z}))(\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z}))^\top$  is positive definite, it holds that

$$(3-9) \quad \boldsymbol{\theta} = \mathbf{D}^{-1}E(\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z}))(Y - \xi(\mathbf{Z})) \stackrel{\text{let}}{=} \mathbf{D}^{-1}\mathbf{c}.$$

If we observe  $\mathbf{X}_i$  and  $\mathbf{Z}_i$ , then the estimation of  $\boldsymbol{\theta}$  is straightforward from the [Equation \(3-9\)](#). If we observe  $\mathbf{Z}_i$  and  $\mathbf{X}_i^*$  but not  $\mathbf{X}_i$ , then we may employ the standard technique that corrects ‘attenuation effect’ due to the measurement errors  $\mathbf{U}_i$  in the estimation of  $\mathbf{D}$ , see [Liang, Härdle, and Carroll \[1999\]](#).

In our setting where both  $\mathbf{X}_i$  and  $\mathbf{Z}_i$  are not available, we may estimate  $\boldsymbol{\eta}$  and  $\xi$  by the technique we have discussed in [Section 3.1](#) with the normalized deconvolution kernel scheme. Call them  $\hat{\boldsymbol{\eta}}^*$  and  $\hat{\xi}^*$ , respectively. A further complication here is that we may



not use  $\hat{\boldsymbol{\eta}}^*(\mathbf{Z}_i)$  and  $\hat{\xi}^*(\mathbf{Z}_i)$  in a formula for estimating  $\boldsymbol{\theta}$  that basically replaces the expectations in (3-9) by the corresponding sample average, since  $\mathbf{Z}_i$  are not available. [E. R. Lee, Han, and Park \[2018\]](#) successfully addressed this problem by observing the following identities.

$$\begin{aligned}\mathbf{D} &= \int_{[0,1]^d} \mathbb{E} \left( (\mathbf{X}^* - \boldsymbol{\eta}(\mathbf{z}))(\mathbf{X}^* - \boldsymbol{\eta}(\mathbf{z}))^\top \middle| \mathbf{Z} = \mathbf{z} \right) p_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z} - \boldsymbol{\Sigma}_{\mathbf{U}}, \\ \mathbf{c} &= \int_{[0,1]^d} \mathbb{E} \left( (\mathbf{X}^* - \boldsymbol{\eta}(\mathbf{z}))(Y - \xi(\mathbf{z}))^\top \middle| \mathbf{Z} = \mathbf{z} \right) p_{\mathbf{Z}}(\mathbf{z}) d\mathbf{z},\end{aligned}$$

where  $p_{\mathbf{Z}}$  denote the joint density of  $\mathbf{Z}$ . Using the normalized deconvolution kernel function  $\tilde{K}_{b_j}^*$  introduced in [Section 3.1](#) with bandwidth  $b_j$  being possibly different from  $h_j$  that are used to estimate  $\boldsymbol{\eta}$  and  $\xi$ , we may estimate  $\mathbf{D}$  and  $\mathbf{c}$  by

$$\begin{aligned}\hat{\mathbf{D}} &= n^{-1} \sum_{i=1}^n \int_{[0,1]^d} (\mathbf{X}_i^* - \hat{\boldsymbol{\eta}}^*(\mathbf{z}))(\mathbf{X}_i^* - \hat{\boldsymbol{\eta}}^*(\mathbf{z}))^\top \prod_{j=1}^d \tilde{K}_{b_j}^*(z_j, Z_{ij}^*) d\mathbf{z} - \boldsymbol{\Sigma}_{\mathbf{U}}, \\ \hat{\mathbf{c}} &= n^{-1} \sum_{i=1}^n \int_{[0,1]^d} (\mathbf{X}_i^* - \hat{\boldsymbol{\eta}}^*(\mathbf{z}))(Y_i - \hat{\xi}^*(\mathbf{z}))^\top \prod_{j=1}^d \tilde{K}_{b_j}^*(z_j, Z_{ij}^*) d\mathbf{z}.\end{aligned}$$

These gives an estimator  $\hat{\boldsymbol{\theta}} = \hat{\mathbf{D}}^{-1} \hat{\mathbf{c}}$  of  $\boldsymbol{\theta}$ .

We may then estimate the additive function  $f = f_1 + \cdots + f_d$  and its component  $f_j$  by applying the technique discussed in [Section 3](#). In this application we takes  $Y_i - \hat{\boldsymbol{\theta}}^\top \mathbf{X}_i^*$  as responses and  $\mathbf{Z}_i^*$  as the contaminated predictor values. Since the rate of convergence of the parametric estimator  $\hat{\boldsymbol{\theta}}$  is faster than the nonparametric rate, as we will see in the following theorem, the resulting estimators of  $f$  and its components  $f_j$  have the same first-order asymptotic properties as the corresponding oracle estimators that use  $Y_i - \boldsymbol{\theta}^\top \mathbf{X}_i^*$  as responses. The asymptotic properties of the oracle estimators are the same as in [Theorem 3.3](#). [Theorem 3.4](#) below demonstrates the best possible rates that  $\hat{\boldsymbol{\theta}}$  can achieves in the three ranges of  $\beta$ , the index for the smoothness of measurement error distribution in the condition (A1). To state the theorem for  $\hat{\boldsymbol{\theta}}$ , we make the following additional assumptions.

(B1)  $\mathbb{E}(X_j^2 \mid \mathbf{Z} = \cdot)$  are bounded on  $[0, 1]^d$ .

(B2) For  $1 \leq j \leq p$ , the component functions of the additive function  $\eta_j$  are twice continuously differentiable on  $[0, 1]$ .

(B3)  $\mathbb{E}(\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z}))(\mathbf{X} - \boldsymbol{\eta}(\mathbf{Z}))^\top$  is positive definite.

(B4) There exist constants  $C > 0$  such that  $\mathbb{E}e^{uW} \leq \exp(Cu^2/2)$  for all  $u$ , for  $W = U_j, X_j$  and  $\varepsilon$ .

**THEOREM 3.4.** (E. R. Lee, Han, and Park [ibid.]). Assume the condition (C1) holds for the marginal and joint densities of  $Z_j$  and  $Z_{jk}$  for all  $1 \leq j \neq k \leq d$ . Also, assume the conditions (C3), (A1), (A2) and (B1)–(B4) hold. Then, (i)  $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p(n^{-1/2})$  when  $\beta < 1/2$  if  $h_j \asymp n^{-\alpha_1}$  and  $b_j \asymp n^{-\alpha_2}$  with  $1/4 \leq \alpha_2 < \alpha_1/(2\beta)$  and  $\max\{1/6, \beta/2\} < \alpha_1 < 1/(3 + 2\beta)$ ; (ii)  $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p(n^{-1/2} \log n)$  when  $\beta = 1/2$  if  $h_j \asymp b_j \asymp n^{-1/4} \sqrt{\log n}$ ; (iii)  $\hat{\boldsymbol{\theta}} - \boldsymbol{\theta} = O_p(n^{-1/(1+2\beta)} \sqrt{\log n})$  when  $\beta > 1/2$  if  $h_j \asymp b_j \asymp n^{-1/(2\beta+4)} (\log n)^{1/4}$ .

## 4 Hilbertian additive models

The analysis of non-Euclidean data objects is an emerging area in modern statistics. A well-known and most studied example is functional data analysis. There have been a few attempts for nonparametric models in this area. These include Dabo-Niang and Rhomari [2009], Ferraty, Laksaci, Tadj, and Vieu [2011] and Ferraty, Van Keilegom, and Vieu [2012]. They studied Nadaraya-Watson estimation of the full-dimensional regression function  $E(Y|\mathbf{X} = \cdot)$  without any structure when the response is in a separable Hilbert or Banach space. The full-dimensional estimator suffers from the curse of dimensionality. More recently, X. Zhang, Park, and Wang [2013], Han, Müller, and Park [2018] and Park, Chen, Tao, and Müller [2018] considered the estimation of structured nonparametric models for functional data, but their studies were either for SBF methods applied to  $Y(t)$  for each  $t$  or for models based on finite number of functional principal/singular components of predictors and responses. Thus, the structured nonparametric models and the methods of estimating them were actually for finite-dimensional Euclidean variables.

In this section we introduce an additive model with response taking values in a Hilbert space and discuss briefly some statistical notions that lay the foundations for estimating the model. This discussion is largely based on the recent work in progress by Jeon [2018]. Let  $\mathbf{Y}$  be a random element taking values in a separable Hilbert space  $\mathbb{H}$ . We confine our discussion to the case where the predictor  $\mathbf{X} = (X_1, \dots, X_d)^\top$  takes values in  $[0, 1]^d$ , however. This is mainly because SBF methods discussed in the previous sections require the marginal and joint densities of  $X_j$  and  $(X_j, X_k)$ , which generally do not exist in infinite-dimensional non-Euclidean cases. For the case where the predictors do not have densities, one may employ ‘surrogate probability density functions’ as discussed in Delaigle and Hall [2010]. Let us denote a vector addition and a real-scalar multiplication by  $\oplus$  and  $\odot$ , respectively. For Borel measurable maps  $\mathbf{f}_j : [0, 1] \rightarrow \mathbb{H}$  as additive components, an additive model for  $E(\mathbf{Y}|\mathbf{X})$  may be written as

$$(4-1) \quad E(\mathbf{Y}|\mathbf{X}) = \mathbf{f}_1(X_1) \oplus \cdots \oplus \mathbf{f}_d(X_d).$$

Below we introduce the notion of Bochner integral, and then discuss briefly its applications to some important statistical notions for the SBF theory.

**4.1 Bochner integration.** Bochner integral is defined for Banach space-valued maps. We start with the classical definition. Let  $(\mathcal{Z}, \mathcal{Q}, \mu)$  be a measure space and  $\mathbb{B}$  be a Banach space with a norm denoted by  $\|\cdot\|$ . We say a map  $\mathbf{m} : \mathcal{Z} \rightarrow \mathbb{B}$  is  $\mu$ -simple if  $\mathbf{m} = \bigoplus_{i=1}^n \mathbf{b}_i \odot 1_{A_i}$  for  $\mathbf{b}_i \in \mathbb{B}$  and disjoint  $A_i \in \mathcal{Q}$  with  $\mu(A_i) < \infty$ . In this case, the Bochner integral of  $\mathbf{m}$  is defined by

$$\int \mathbf{m} d\mu = \bigoplus_{i=1}^n \mathbf{b}_i \odot \mu(A_i).$$

A map  $\mathbf{m} : \mathcal{Z} \rightarrow \mathbb{B}$  is called  $\mu$ -measurable if  $\mathbf{m}$  is an almost everywhere limit of  $\mu$ -simple maps. A  $\mu$ -measurable map  $\mathbf{m}$  is called Bochner integrable if  $\int \|\mathbf{m}\| d\mu < \infty$ . In this case, the Bochner integral of  $\mathbf{m}$  is defined by

$$\int \mathbf{m} d\mu = \lim_{n \rightarrow \infty} \int \mathbf{m}_n d\mu$$

for a sequence of  $\mu$ -simple maps  $\mathbf{m}_n$  such that  $\mathbf{m}_n \rightarrow \mathbf{m}$  a.e.  $[\mu]$ .

In statistical applications of Bochner integrals, the measure  $\mu$  in a measure space  $(\mathcal{Z}, \mathcal{Q}, \mu)$  is the distribution of a random variable. In the case of the additive maps  $\mathbf{f}_j$  in (4-1),  $\mu$  corresponds to  $PX_j^{-1}$  where  $P$  is the probability measure of the probability space  $(\Omega, \mathcal{F}, P)$  where  $X_j$  is defined. The classical definition given above for  $\mu$ -measurable maps is not appropriate since  $PX_j^{-1}$ -measurability of  $\mathbf{f}_j$  is not equivalent to Borel-measurability of  $\mathbf{f}_j$ . In the model (4-1), we implicitly assume that  $\mathbf{f}_j(X_j)$  are random elements, i.e., Borel-measurable with respect to  $\mathcal{F}$ , as is usual in all statistical problems. For this reason we assume in the model (4-1) that each  $\mathbf{f}_j$  is Borel-measurable with respect to the Borel  $\sigma$ -field of  $[0, 1]$ .

The notion of Bochner integral may be extended to Borel-measurable maps. We introduce it briefly here. We refer to Cohn [2013] for more details. For a Banach space  $\mathbb{B}$ , a map  $\mathbf{m} : \mathcal{Z} \rightarrow \mathbb{B}$  is called simple if  $\mathbf{m}$  takes only finitely many values. A map  $\mathbf{m} : \mathcal{Z} \rightarrow \mathbb{B}$  is called strongly measurable if  $\mathbf{m}$  is Borel-measurable and  $\mathbf{m}(\mathcal{Z})$  is separable. A map  $\mathbf{m} : \mathcal{Z} \rightarrow \mathbb{B}$  is called strongly integrable if  $\mathbf{m}$  is strongly measurable and  $\int_{\mathcal{Z}} \|\mathbf{m}\| d\mu < \infty$ . If  $\mathbf{m}$  is strongly integrable, then there exists a Cauchy sequence of strongly integrable simple maps  $\mathbf{m}_n$  such that  $\lim_{n, m \rightarrow \infty} \int \|\mathbf{m}_n - \mathbf{m}_m\| d\mu \rightarrow 0$  and  $\lim_{n \rightarrow \infty} \mathbf{m}_n(z) = \mathbf{m}(z)$  for all  $z \in \mathcal{Z}$ . In this case,  $\int \mathbf{m} d\mu$  is defined as  $\lim_{n \rightarrow \infty} \int \mathbf{m}_n d\mu$ .

**4.2 Statistical properties of Bochner integrals.** Since the notion of Bochner integral is new in statistics, statistical properties of this integral have been rarely studied. There are many statistical notions and properties one needs to define and derive to develop relevant theory for estimating the model (4-1). It was only recent that Jeon [2018] studied such basic ingredients. Below, we present two formulas regarding the notions of expectation

and of conditional expectation that are essential in developing further theory for the SBF estimation of (4-1).

Let  $\mathbb{B}$  be a separable Banach space. Let  $\mathbf{Z}$  and  $\mathbf{W}$  be random elements taking values in  $\sigma$ -finite measure spaces  $(\mathcal{Z}, \mathcal{G}, \mu)$  and  $(\mathcal{W}, \mathcal{B}, \nu)$ , respectively. We assume  $P\mathbf{Z}^{-1} \ll \mu$ ,  $P\mathbf{W}^{-1} \ll \nu$  and  $P(\mathbf{Z}, \mathbf{W})^{-1} \ll \mu \otimes \nu$ , where  $P\mathbf{Z}^{-1}$ ,  $P\mathbf{W}^{-1}$  and  $P(\mathbf{Z}, \mathbf{W})^{-1}$  are the probability distributions of  $\mathbf{Z}$ ,  $\mathbf{W}$  and  $(\mathbf{Z}, \mathbf{W})$ , respectively, so that there exist densities of  $\mathbf{Z}$ ,  $\mathbf{W}$  and  $(\mathbf{Z}, \mathbf{W})$ , denoted by  $p_{\mathbf{Z}}$ ,  $p_{\mathbf{W}}$  and  $p_{\mathbf{Z}, \mathbf{W}}$ , respectively. We first introduce a general expectation formula, and then a conditional expectation formula, in terms of the densities of  $\mathbf{Z}$ ,  $\mathbf{W}$  and  $(\mathbf{Z}, \mathbf{W})$ .

**PROPOSITION 4.1.** (Jeon [ibid.]) *Assume that  $\mathbf{f} : \mathcal{Z} \rightarrow \mathbb{B}$  is a strongly measurable map such that  $E(\|\mathbf{f}(\mathbf{Z})\|) < \infty$ . Then,  $E(\mathbf{f}(\mathbf{Z})) = \int_{\mathcal{Z}} \mathbf{f}(\mathbf{z}) \odot p_{\mathbf{Z}}(\mathbf{z}) d\mu$ .*

**PROPOSITION 4.2.** (Jeon [ibid.]) *Assume that  $p_{\mathbf{W}} \in (0, \infty)$  on  $\mathcal{W}$  and that  $\mathbf{f} : \mathcal{Z} \rightarrow \mathbb{B}$  is a strongly measurable map such that  $E(\|\mathbf{f}(\mathbf{Z})\|) < \infty$ . Let  $\mathbf{g} : \mathcal{W} \rightarrow \mathbb{B}$  be a map defined by*

$$\mathbf{g}(\mathbf{w}) = \begin{cases} \int_{\mathcal{Z}} \mathbf{f}(\mathbf{z}) \odot \frac{p_{\mathbf{Z}, \mathbf{W}}(\mathbf{z}, \mathbf{w})}{p_{\mathbf{W}}(\mathbf{w})} d\mu, & \text{if } \mathbf{w} \in D_{\mathcal{W}} \\ \mathbf{g}_0(\mathbf{w}), & \text{otherwise} \end{cases}$$

where  $D_{\mathcal{W}} = \{\mathbf{w} \in \mathcal{W} : \int_{\mathcal{Z}} \|\mathbf{f}(\mathbf{z})\| p_{\mathbf{Z}, \mathbf{W}}(\mathbf{z}, \mathbf{w}) d\mu < \infty\}$  and  $\mathbf{g}_0 : \mathcal{W} \rightarrow \mathbb{B}$  is any strongly measurable map. Then,  $\mathbf{g}$  is strongly measurable and  $\mathbf{g}(\mathbf{W})$  is a version of  $E(\mathbf{f}(\mathbf{Z})|\mathbf{W})$ .

**4.3 Discussion.** The additive regression model (4-1) for Hilbertian response have many important applications. Non-Euclidean data objects often take values in Hilbert spaces. Examples include functions, images, probability densities and simplices. Among them, the latter two data objects have certain constraints. A density is non-negative and its integral over the corresponding domain where it is defined equals 1. A simplex data object,  $(v_1, \dots, v_D)^{\top}$  with  $v_k > 0$  for  $1 \leq k \leq D$  and  $\sum_{k=1}^D v_k = 1$ , has similar constraints. Analyzing such data objects with standard Euclidean regression techniques would give estimates that are off the space where the data objects take values. The approach based on the model (4-1) with the corresponding Hilbertian operations  $\oplus$  and  $\odot$  would give a proper estimate of the regression map that forces its values lie in the space of the data objects. It also avoids the curse of dimensionality when  $d$  is high. This way would lead us to a powerful nonparametric technique that unifies various statistical methods for analyzing non-Euclidean data objects.

## References

- E. J. Beltrami (1967). “On infinite-dimensional convex programs”. *J. Comput. System Sci.* 1, pp. 323–329. MR: [0232603](#).
- Peter J. Bickel, Chris A. J. Klaassen, Ya’acov Ritov, and Jon A. Wellner (1993). *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, p. 560. MR: [1245941](#) (cit. on pp. [3018](#), [3020](#)).
- Andreas Buja, Trevor Hastie, and Robert Tibshirani (1989). “Linear smoothers and additive models”. *Ann. Statist.* 17.2, pp. 453–555. MR: [994249](#) (cit. on p. [3014](#)).
- Raymond J. Carroll and Peter Hall (1988). “Optimal rates of convergence for deconvolving a density”. *J. Amer. Statist. Assoc.* 83.404, pp. 1184–1186. MR: [997599](#) (cit. on p. [3026](#)).
- Donald L. Cohn (2013). *Measure theory*. Second. Birkhäuser Advanced Texts: Basler Lehrbücher. [Birkhäuser Advanced Texts: Basel Textbooks]. Birkhäuser/Springer, New York, pp. xxi+457. MR: [3098996](#) (cit. on p. [3034](#)).
- Sophie Dabo-Niang and Noureddine Rhomari (2009). “Kernel regression estimation in a Banach space”. *J. Statist. Plann. Inference* 139.4, pp. 1421–1434. MR: [2485136](#) (cit. on p. [3033](#)).
- Aurore Delaigle, Jianqing Fan, and Raymond J. Carroll (2009). “A design-adaptive local polynomial estimator for the errors-in-variables problem”. *J. Amer. Statist. Assoc.* 104.485, pp. 348–359. MR: [2504382](#) (cit. on p. [3026](#)).
- Aurore Delaigle and Peter Hall (2010). “Defining probability density for a distribution of random functions”. *Ann. Statist.* 38.2, pp. 1171–1193. MR: [2604709](#) (cit. on p. [3033](#)).
- (2016). “Methodology for non-parametric deconvolution when the error distribution is unknown”. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* 78.1, pp. 231–252. MR: [3453654](#) (cit. on p. [3026](#)).
- Aurore Delaigle, Peter Hall, and Alexander Meister (2008). “On deconvolution with repeated measurements”. *Ann. Statist.* 36.2, pp. 665–685. MR: [2396811](#) (cit. on p. [3026](#)).
- Jianqing Fan (1991). “On the optimal rates of convergence for nonparametric deconvolution problems”. *Ann. Statist.* 19.3, pp. 1257–1272. MR: [1126324](#) (cit. on p. [3031](#)).
- Jianqing Fan and Young K. Truong (1993). “Nonparametric regression with errors in variables”. *Ann. Statist.* 21.4, pp. 1900–1925. MR: [1245773](#) (cit. on p. [3026](#)).
- Jianqing Fan and Wenyang Zhang (1999). “Statistical estimation in varying coefficient models”. *Ann. Statist.* 27.5, pp. 1491–1518. MR: [1742497](#) (cit. on p. [3025](#)).
- (2000). “Simultaneous confidence bands and hypothesis testing in varying-coefficient models”. *Scand. J. Statist.* 27.4, pp. 715–731. MR: [1804172](#) (cit. on p. [3025](#)).

- F. Ferraty, I. Van Keilegom, and P. Vieu (2012). “Regression when both response and predictor are functions”. *J. Multivariate Anal.* 109, pp. 10–28. MR: [2922850](#) (cit. on p. [3033](#)).
- Frédéric Ferraty, Ali Laksaci, Amel Tadj, and Philippe Vieu (2011). “Kernel regression with functional response”. *Electron. J. Stat.* 5, pp. 159–171. MR: [2786486](#) (cit. on p. [3033](#)).
- Jerome H. Friedman and Werner Stuetzle (1981). “Projection pursuit regression”. *J. Amer. Statist. Assoc.* 76.376, pp. 817–823. MR: [650892](#) (cit. on p. [3014](#)).
- K. Han and Byeong U. Park (2018). “Smooth backfitting for errors-in-variables additive models”. To appear in *Annal of Statistics* (cit. on pp. [3026–3031](#)).
- Kyunghee Han, Hans-Georg Müller, and Byeong U. Park (2018). “Smooth backfitting for additive modeling with small errors-in-variables, with an application to additive functional regression for multiple predictor functions”. *Bernoulli* 24.2, pp. 1233–1265. MR: [3706793](#) (cit. on p. [3033](#)).
- Trevor Hastie and Robert Tibshirani (1993). “Varying-coefficient models”. *J. Roy. Statist. Soc. Ser. B* 55.4. With discussion and a reply by the authors, pp. 757–796. MR: [1229881](#) (cit. on p. [3025](#)).
- J. M. Jeon (2018). “Additive regression with Hilbertian responses”. PhD thesis. Seoul National University (cit. on pp. [3033–3035](#)).
- E. R. Lee, K. Han, and Byeong U. Park (2018). “Estimation of errors-in-variables partially linear additive models”. To appera in *Statistica Sinica* (cit. on pp. [3031–3033](#)).
- Young K. Lee, Enno Mammen, and Byeong U. Park (2012a). “Flexible generalized varying coefficient regression models”. *The Annals of Statistics* 40.3, pp. 1906–1933. MR: [3015048](#) (cit. on p. [3026](#)).
- (2012b). “Projection-type estimation for varying coefficient regression models”. *Bernoulli* 18.1, pp. 177–205. MR: [2888703](#) (cit. on p. [3025](#)).
  - (2014). “Backfitting and smooth backfitting in varying coefficient quantile regression”. *Econom. J.* 17.2, S20–S38. MR: [3219148](#) (cit. on p. [3026](#)).
- Young Kyung Lee (2004). “On marginal integration method in nonparametric regression”. *J. Korean Statist. Soc.* 33.4, pp. 435–447. MR: [2126371](#) (cit. on p. [3015](#)).
- (2017). “Nonparametric estimation of bivariate additive models”. *J. Korean Statist. Soc.* 46.3, pp. 339–348. MR: [3685573](#) (cit. on p. [3026](#)).
- Young Kyung Lee, Enno Mammen, and Byeong U. Park (2010). “Backfitting and smooth backfitting for additive quantile models”. *Ann. Statist.* 38.5, pp. 2857–2883. MR: [2722458](#) (cit. on p. [3026](#)).
- Hua Liang, Wolfgang Härdle, and Raymond J. Carroll (1999). “Estimation in a semiparametric partially linear errors-in-variables model”. *Ann. Statist.* 27.5, pp. 1519–1535. MR: [1742498](#) (cit. on p. [3031](#)).

- Oliver Linton and Jens Perch Nielsen (1995). “A kernel method of estimating structured nonparametric regression based on marginal integration”. *Biometrika* 82.1, pp. 93–100. MR: [1332841](#) (cit. on p. [3014](#)).
- E. Mammen, O. Linton, and J. Nielsen (1999). “The existence and asymptotic properties of a backfitting projection algorithm under weak conditions”. *Ann. Statist.* 27.5, pp. 1443–1490. MR: [1742496](#) (cit. on pp. [3014](#), [3015](#), [3018–3020](#), [3024](#)).
- Enno Mammen and Jens Perch Nielsen (2003). “Generalised structured models”. *Biometrika* 90.3, pp. 551–566. MR: [2006834](#) (cit. on p. [3014](#)).
- Enno Mammen and Byeong U. Park (2005). “Bandwidth selection for smooth backfitting in additive models”. *Ann. Statist.* 33.3, pp. 1260–1294. MR: [2195635](#) (cit. on p. [3024](#)).
- (2006). “A simple smooth backfitting method for additive models”. *Ann. Statist.* 34.5, pp. 2252–2271. MR: [2291499](#) (cit. on p. [3024](#)).
- Enno Mammen, Byeong U. Park, and Melanie Schienle (2014). “Additive models: extensions and related models”. In: *The Oxford handbook of applied nonparametric and semiparametric econometrics and statistics*. Oxford Univ. Press, Oxford, pp. 176–211. MR: [3306926](#) (cit. on p. [3014](#)).
- Jens Perch Nielsen and Stefan Sperlich (2005). “Smooth backfitting in practice”. *J. R. Stat. Soc. Ser. B Stat. Methodol.* 67.1, pp. 43–61. MR: [2136638](#) (cit. on p. [3024](#)).
- Jean D. Opsomer (2000). “Asymptotic properties of backfitting estimators”. *J. Multivariate Anal.* 73.2, pp. 166–179. MR: [1763322](#) (cit. on p. [3014](#)).
- Jean D. Opsomer and David Ruppert (1997). “Fitting a bivariate additive model by local polynomial regression”. *Ann. Statist.* 25.1, pp. 186–211. MR: [1429922](#) (cit. on p. [3014](#)).
- Byeong U. Park, C.-J. Chen, W. Tao, and H.-G. Müller (2018). “Singular additive models for function to function regression”. To appear in *Statistica Sinica* (cit. on p. [3033](#)).
- Byeong U. Park, Enno Mammen, Young K. Lee, and Eun Ryung Lee (2015). “Varying coefficient regression models: a review and new developments”. *Int. Stat. Rev.* 83.1, pp. 36–64. MR: [3341079](#) (cit. on p. [3025](#)).
- Leonard Stefanski and Raymond J. Carroll (1990). “Deconvoluting kernel density estimators”. *Statistics* 21.2, pp. 169–184. MR: [1054861](#) (cit. on p. [3026](#)).
- Lijian Yang, Byeong U. Park, Lan Xue, and Wolfgang Härdle (2006). “Estimation and testing for varying coefficients in additive models with marginal integration”. *J. Amer. Statist. Assoc.* 101.475, pp. 1212–1227. MR: [2328308](#) (cit. on p. [3025](#)).
- Kyusang Yu, Enno Mammen, and Byeong U. Park (2011). “Semi-parametric regression: efficiency gains from modeling the nonparametric part”. *Bernoulli* 17.2, pp. 736–748. MR: [2787613](#) (cit. on p. [3026](#)).
- Kyusang Yu, Byeong U. Park, and Enno Mammen (2008). “Smooth backfitting in generalized additive models”. *Ann. Statist.* 36.1, pp. 228–260. MR: [2387970](#) (cit. on pp. [3024](#), [3025](#)).

Xiaoke Zhang, Byeong U. Park, and Jane-Ling Wang (2013). “[Time-varying additive models for longitudinal data](#)”. *J. Amer. Statist. Assoc.* 108.503, pp. 983–998. MR: [3174678](#) (cit. on p. [3033](#)).

Received 2017-12-06.

BYEONG U. PARK  
[bupark@stats.snu.ac.kr](mailto:bupark@stats.snu.ac.kr)





# A SELECTIVE SURVEY OF SELECTIVE INFERENCE

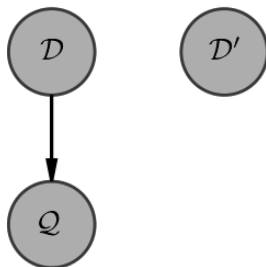
JONATHAN E. TAYLOR

The idea of a scientist, struck, as if by lightning with a question, is far from the truth. – Tukey [1980].

## Abstract

It is not difficult to find stories of a crisis in modern science, either in the popular press or in the scientific literature. There are likely multiple sources for this crisis. It is also well documented that one source of this crisis is the misuse of statistical methods in science, with the  $P$ -value receiving its fair share of criticism. It could be argued that this misuse of statistical methods is caused by a shift in how data is used in 21st century science compared to its use in the mid-20th century which presumed scientists had formal statistical hypotheses before collecting data. With the advent of sophisticated statistical software available to anybody this paradigm has been shifted to one in which scientists *collect data first and ask questions later*.

## 1 The new (?) scientific paradigm



We are all familiar with a paradigm that does allow scientists to collect data first and ask questions later: the classical scientific method illustrated in Figure 1. A scientist collects data  $\mathcal{D}$ , generates questions of interest  $Q(\mathcal{D})$ , then collects fresh data  $\mathcal{D}'$  for confirmation and perhaps to discover additional questions of interest. The problem with this new paradigm is that it seeks to use  $\mathcal{D}$  to answer these questions and may not have access to  $\mathcal{D}'$ .

We pause here to note that Tukey used the term *question* rather than the more precise term *hypothesis* which statisticians might reasonably impute to be a statistical hypothesis. Given the computing capabilities of modern

Figure 1: A simplified version of the scientific method.

MSC2010: primary 62-02; secondary 62J15, 62J05, 62-07.

statistical software, it is not really clear that data analysis produces statistical hypotheses:

In practice, of course, hypotheses often emerge after the data have been examined; patterns seen in the data combine with subject-matter knowledge in a mix that has so far defied description. – P. Diaconis “[Theories of Data Analysis](#)” [n.d.]

We continue to distinguish a question  $Q$  from a *statistical object* such as a hypothesis test, i.e. a pair  $(\mathfrak{M}, H_0)$  with  $\mathfrak{M}$  a statistical model (a collection of distributions on some measurable space) and  $H_0 \subset \mathfrak{M}$ ; or a pair  $(\mathfrak{M}, \theta)$  with  $\theta : \mathfrak{M} \rightarrow \mathbb{R}^k$  a parameter for which we might form a region or point estimate. An example of a Bayesian statistical might be a triple  $(\pi, \ell, \mathcal{T})$  with  $\pi$  a prior,  $\ell$  a likelihood and  $\mathcal{T}$  some functional of the posterior. The transformation from questions to statistical objects is up to the scientist, perhaps in partnership with a statistician.

Returning to the new paradigm in science, whether the statistics community feel that this is the correct way to run experiments and advance a particular field of science, it is difficult to ignore the fact that it is how (at least some) modern science is practiced. We feel it is imperative to provide scientists with tools that provide some of the guarantees of the classical methods but are applicable in this new paradigm. These are the problems that the area of *selective inference* attempts to address. The term *selective* refers to the fact that the results reported in a scientific study (e.g.  $P$ -values, confidence intervals) are selected through some mechanism guided by the scientist. When the mechanism of selection is known, it is often possible to mitigate this selection bias.

**1.1 Two prototypical settings with many questions.** We describe two prototypical problems occurring in many modern scientific disciplines, from genomic studies to neuroscience and many others. Both involve a response  $y \in \mathbb{R}$  (which we take to be real-valued simply for concreteness) and a set of features  $X \in \mathbb{R}^p$ .

**1.1.1 Large scale inference.** Often of interest are the  $p$  questions

$$(1) \quad Q_j^L : \text{Is feature } j \text{ associated with outcome } y? \quad 1 \leq j \leq p$$

This problem is often referred to as *large-scale inference* [Efron \[2012\]](#) ( $L$  for large) and has brought about a renewed interest in empirical Bayes methodology and multiple comparisons in general.

A canonical experimental design in this problem samples  $n$  pairs IID from some law  $F$  in a statistical model  $\mathfrak{M}$ . Having these pre-determined set of questions allows the statistician, given the model  $\mathfrak{M}$ , to transform each question to a parameter in model  $\mathfrak{M}$ :

$Q_j^L \mapsto \theta_j^L \in \mathbb{R}^{\mathfrak{M}}$  where  $\theta_j^L$  measures a marginal association between  $y$  and feature  $j$  on  $\mathfrak{M}$ . Parameters in hand, the statistician can then use the formal methods of statistical inference to “answer” these questions.

**1.1.2 Feature selection in regression.** The second problem also involves response  $y$  and features  $X$ , though in this case the scientist seeks to build a predictive model of  $y$  from  $X$ . At first glance, the natural questions are

(2)  $Q_j^R$ : Is feature  $j$  important when trying to predict  $y$  from  $X$ ?  $1 \leq j \leq p$ .

As is clear to any student after a course in linear regression ( $R$  for regression) the above questions are ill posed. This point is emphasized in [Berk, Brown, Buja, K. Zhang, and Zhao \[2013\]](#) which then posed the following questions

(3)  $\bar{Q}_{j|E}^R$ : Is feature  $j$  correlated with the residual when trying predict  $y$  from  $X_{E \setminus j}$ ?

Such questions are indexed by  $j \in E, E \subset \{1, \dots, p\}$ . These questions are also posed before data collection as soon as the scientist decided to collect these  $p$  features to build a predictive model for  $y$  from  $X$ .

These two problems have inspired much work in selective inference: with the large scale inference problem drawing intense focus in the early part of the 21st century [Efron \[2012\]](#) and [Storey \[2003\]](#) building on the seminal work of [Benjamini and Hochberg \[1995\]](#). The regression problem is an area of more recent interest [Berk, Brown, Buja, K. Zhang, and Zhao \[2013\]](#), [Hurvich and Tsai \[1990\]](#), [Lee, D. L. Sun, Y. Sun, and J. E. Taylor \[2016\]](#), [Lockhart, J. Taylor, R. J. Tibshirani, and R. Tibshirani \[2014\]](#), and [Wasserman and Roeder \[2009\]](#) perhaps due to the added complexity of the set of possible questions under consideration.

## 2 Selective inference

We now describe some proposed methods to address this new scientific paradigm. The first set of methods, based on multiple comparisons, generally ignore the possible existence of  $\mathfrak{D}'$  (formally equivalent to setting  $\mathfrak{D}' = 0$ , a constant random variable) while the conditional methods (of which the classical scientific method is one) do acknowledge that  $\mathfrak{D}$  likely does not exist in a vacuum. A scientist probably will have run previous experiments and will (subject to restrictions) run future experiments.

**2.1 Multiple comparisons and simultaneous inference.** Some of the earliest references to selective inference [Benjamini \[2010\]](#) and [Benjamini and Yekutieli \[2005\]](#) come

from the field of *multiple comparisons* in the large scale inference problem, particularly through the extraordinarily influential work of [Benjamini and Hochberg \[1995\]](#) and its introduction of the False Discovery Rate (FDR) as a more liberal error rate than the Family Wise Error Rate (FWER).

The goal in multiple comparisons procedures is to construct procedures  $T$  that control an error rate such as the FDR or FWER (defined below) over some statistical model. For concreteness, suppose that in the large scale inference problem we have access to an estimate  $Z_j$  of association  $\theta_j$  between feature  $j$  and response  $y$  we might take  $\mathfrak{D} = Z$  and our statistical model to be

$$(4) \quad \mathfrak{M}^L = \{N(\theta, \Sigma) : \theta \in \mathbb{R}^p, \Sigma \in \mathbb{R}^{p \times p} \geq 0, \text{diag}(\Sigma) = 1\}$$

with  $\Sigma$  known or unknown (but hopefully estimable). In this problem, a multiple comparisons procedure is a map  $T : \mathbb{R}^p \rightarrow \{0, 1\}^p$  that makes a decision whether each hypothesis  $H_{0,j}$  is true or false. A procedure  $T$  that controls the FWER at level  $\alpha$  satisfies

$$(5) \quad FWER(T, F) = \mathbb{P}_F(V(T, Z) > 0) \leq \alpha, \quad \forall F \in \mathfrak{M}^L$$

with

$$(6) \quad V(T, Z) = \{j : \theta_j \neq 0, T_j(Z) = 1\}$$

the number of false positives the procedure  $T$  selected on outcomes  $Z$  where  $T_j(Z) = 1$  signals a positive decision, i.e. that  $H_{0,j}$  is false. When  $p = 1$  and only one hypothesis is under consideration,  $FWER$  reduces to Type I error, 0 for any  $F \notin H_0$ . A test that controls the Type I error at level  $\alpha$  satisfies

$$(7) \quad \mathbb{P}_F(T(Z) = 1) \leq \alpha, \quad \forall F \in H_0.$$

The  $FDR$  of procedure  $T$  is also expressible as an expectation under the law  $F$ . Note that controlling FDR or FWER are *marginal* properties of each  $F \in \mathfrak{M}^L$ . This will be contrasted below with *conditional* properties.

The prototypical example of a procedure that controls the FWER is the Bonferroni procedure which uses a simple bound on the law of the largest  $(Z_j)_{1 \leq j \leq p}$ :

$$(8) \quad \mathbb{P}_F \left( \max_{1 \leq j \leq p} |Z_j - \theta_j| > t \right) \leq p \cdot \bar{\Phi}(t), \quad \forall F \in \mathfrak{M}^L$$

with  $\bar{\Phi}$  the tail of a standard normal random variable. Tighter approximations of the left-hand valid over some  $\mathfrak{M}$  can be used to get an improvement over Bonferroni. This area of research is sometimes referred to as *simultaneous inference*. The late 20th century

saw its own golden era in research in this area [Adler and J. E. Taylor \[2007\]](#), [Azaïs and Wschebor \[2009\]](#), [Siegmund and Worsley \[1995\]](#), [J. Sun \[1993\]](#), and [Takemura and Kuriki \[2002\]](#) in which the feature space was modelled as approximating a continuum with  $\Sigma$  a representation of the covariance function of some Gaussian process.

It is well known that bounds for the left hand side of (8) translate to (*simultaneous*) *coverage guarantees* for confidence intervals for the  $\theta_j$ . For example, suppose  $t$  is such that some bound for the left hand side of (8) is less than  $\alpha$ . Then,

$$(9) \quad \mathbb{P}_F (\exists j : \theta_j \notin [Z_j - t, Z_j + t]) \leq \alpha, \quad \forall F \in \mathfrak{M}^L.$$

This simultaneous approach was considered in [Berk, Brown, Buja, K. Zhang, and Zhao \[2013\]](#). Formally (considering  $X$  fixed) the authors set

$$(10) \quad \mathfrak{M}^{\text{POSI}} = \{N(\mu, I_{n \times n}) : \mu \in \mathbb{R}^n\}.$$

and transform the questions  $\bar{Q}_{j|E}^R \mapsto \theta_{j|E}^R \in (\mathbb{R}^n)^*$  to parameters taken to be the linear functionals

$$(11) \quad \theta_{j|E}^R(\mu) = e_j^T (X_E^T X_E)^{-1} X_E^T \mu$$

where  $e_j : \mathbb{R}^E \rightarrow \mathbb{R}$  is projection onto the  $j$  coordinate. The authors of [Berk, Brown, Buja, K. Zhang, and Zhao \[ibid.\]](#) then note that  $\mathfrak{M}^{\text{POSI}}$  can be embedded in a model of the form  $\mathfrak{M}^L$  indexed by  $(j, E)$  with  $j \in E$  and  $E \subset \{1, \dots, p\}$  and taking  $\mathfrak{D}$  to be (some subset of) the collection of corresponding  $Z$  statistics. The authors propose finding a bound better than Bonferroni using simulation.

We end this section with *knockoffs* [Barber and Candès \[2014\]](#), a different approach to the regression problem within the framework of multiple comparisons. Being a regression problem, questions must specify  $E$ . The knockoff setting fixes  $E = \{1, \dots, p\}$  in which case the questions of interest are

$$(12) \quad \bar{Q}_j^F : \text{Is feature } j \text{ correlated with the residual when trying predict } y \text{ from } X_{-j}?$$

with  $F$  above standing for the *full* model. The authors consider  $X$  fixed and take  $\mathfrak{D} = y$  and the statistical model to be

$$(13) \quad \mathfrak{M}^K = \{N(X\beta, \sigma^2 I) : \beta \in \mathbb{R}^p, \sigma^2 > 0\}.$$

There is again a natural transformation from questions to statistical hypotheses  $\bar{Q}_j^F \mapsto H_{0,j} : \beta_j = 0$ . The usual  $t$  or  $Z$ -statistics for the least-squares estimates in the above regression model could be used to test each of these hypotheses. Rather than use this embedding, the authors choose alternative statistics based on constructing a pseudo-feature

for each feature  $j$  constructing a procedure that controls (a slight modification of) the FDR. Their demonstration of (modified) FDR control through counting processes has reinvigorated methodological work in FDR and has led to, among other things, work on other more adaptive procedures for FDR control [Lei and Fithian \[2016\]](#), [Li and Barber \[2015\]](#), [Lei, Ramdas, and Fithian \[2017\]](#), and [Barber and Ramdas \[2017\]](#). Other interesting work in FDR control includes work on hierarchically arranged families of hypotheses [Benjamini and Bogomolov \[2014\]](#).

The authors of [Barber and Candes \[2014\]](#) demonstrate empirically that this construction can be more powerful than the natural embedding based on the usual  $t$  or  $Z$  statistics. The knockoffs framework has been extended [Candes, Fan, Janson, and Lv \[2016\]](#) to settings under which the law of  $X$  is assumed known and it is feasible to construct swap-exchangeable pseudo-features  $\tilde{X}$ , expanding the applicability of knockoffs when such assumptions are reasonable in which the questions  $\bar{Q}_j^F$  are transformed to hypotheses of conditional independence.

**2.2 Does this match science’s new paradigm?** It seems that both the large-scale inference problem as well as the regression problem can be embedded into multiple comparisons problems (though technical considerations certainly still remain). We have also unfortunately ignored Bayesian methods up to this point, though we return to this briefly below.

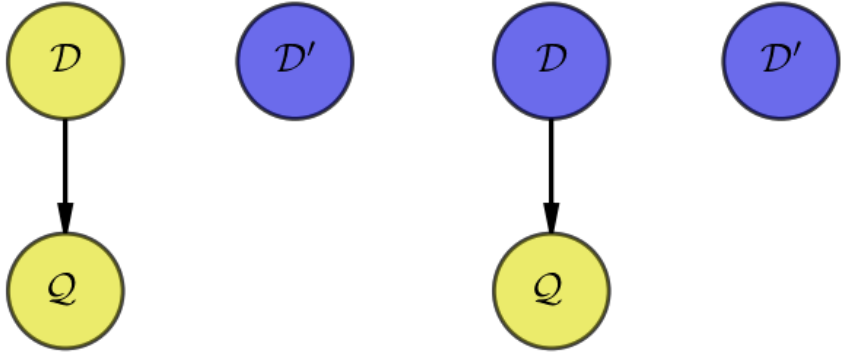
Recall our original goal: to provide tools for inference in a paradigm where people collect data first and ask questions later. A quick look the examples show that the questions were actually pre-determined<sup>1</sup>.

Not only were the questions pre-determined, the questions were formally transformed to statistical objects. This transformation is what allows statisticians to apply the formal methods of multiple comparisons to “answer” these questions. While these transformations of questions to statistical objects seem quite natural, they are not exhaustive. Why not just let the scientist look at the data to generate their own questions and have them pick the statistical objects for reporting?

**2.3 Conditional inference.** In this section, we describe a *conditional* approach to selective inference that allows the scientist to look at their data to generate questions of interest and corresponding statistical objects for final reporting. We do not have to look far for an example of the conditional approach. Indeed, our simple iteration of the scientific method as presented in [Figure 1](#) provides an example.

---

<sup>1</sup>Arguably, one exception to this is the simultaneous approach of [Berk, Brown, Buja, K. Zhang, and Zhao \[2013\]](#) as it allows a researcher to choose which of some prespecified list of  $E$  to use in the report. But what is the scientist to do if she discovers her chosen  $E$  (after inspection) was not in the list specified before the analysis?



(a) The scientist conditions on  $(\mathcal{D}, \mathcal{Q}(\mathcal{D}))$  and must posit a model for the law of  $\mathcal{D}'|\mathcal{D}$ .

(b) The scientist conditions on  $\mathcal{Q}(\mathcal{D})$ , the minimal sigma algebra used to generate her statistical objects, and must posit a model for the joint law of  $(\mathcal{D}, \mathcal{D}')$ .

Figure 2: Two different conditional models.

**2.3.1 The scientific method.** Having collected data  $\mathcal{D}$ , the scientist's data analysis can be represented as a function  $\mathcal{Q}(\mathcal{D})$ , quite literally the functions the scientist used in their exploratory data analysis, typically in some statistical software package. At this point, the scientist need not have attached any statistical model to the data as  $\mathcal{Q}(\mathcal{D})$  are simply patterns.

Based on  $\mathcal{Q}(\mathcal{D})$ , the scientist is free to posit a statistical model  $\mathcal{M}$  (subject to defending this model to their peers) with corresponding statistical objects which will be used for formal statistical inference on  $\mathcal{D}'$ . As the results are only evaluated on  $\mathcal{D}'$  we can view  $\mathcal{D}$  as fixed, fixing  $\mathcal{Q}(\mathcal{D})$  as well. This fixing of  $\mathcal{Q}(\mathcal{D})$  allows the scientist to transform these patterns into statistical objects such as hypothesis tests, point estimators or confidence intervals.

Fixing  $\mathcal{D}$  is equivalent to conditioning on it – any honest accounting of how  $\mathcal{D}$  and  $\mathcal{D}'$  came to be must acknowledge that  $\mathcal{D}$  is random so there certainly exists some joint distribution for  $(\mathcal{D}, \mathcal{D}')$ . Formal inference is applied only to  $\mathcal{D}'$  hence the scientist's model  $\mathcal{M}$  is really a model for the law of  $\mathcal{D}'|\mathcal{D}$ . This fixing of  $\mathcal{D}$  is illustrated in Figure 2a, denoting fixed variables by yellow and variables modeled by the scientist in blue.

**2.3.2 Conditional approach in general.** If  $\mathcal{Q}(\mathcal{D})$  is enough information for the scientist to posit a model  $\mathcal{M}$  for the law  $\mathcal{D}'|\mathcal{D}$ , it is often reasonable to assume it is enough



information for the scientist to posit a model for the joint law of  $(\mathfrak{D}, \mathfrak{D}')$ . For example, when both  $\mathfrak{D}$  and  $\mathfrak{D}'$  are IID samples from some population then any model for  $\mathfrak{D}'$  is similarly a model for  $\mathfrak{D}$ . In this setting, if a model for  $\mathfrak{D}'$  is defensible to their peers, the same model must be certainly defensible for  $\mathfrak{D}$ .

Since it is  $\mathbb{Q}(\mathfrak{D})$  that led the scientist to the questions that were then transformed into statistical objects, it is sufficient to only fix  $\mathbb{Q}(\mathfrak{D})$ . This is the basis of the conditional approach to selective inference [Bi, Markovic, Xia, and J. Taylor \[2017\]](#), [Fithian, D. Sun, and J. Taylor \[2014\]](#), and [Lee, D. L. Sun, Y. Sun, and J. E. Taylor \[2016\]](#). Conditioning only on  $\mathbb{Q}(\mathfrak{D})$  we apply the formal tools of statistical inference to the remaining randomness in  $(\mathfrak{D}, \mathfrak{D}')$ . In turn, this means that the appropriate Type I error to consider is the *selective Type I error*, requiring the *conditional guarantee*

$$(14) \quad \mathbb{P}_F(T(\mathfrak{D}, \mathfrak{D}') = 1 | \mathbb{Q}(\mathfrak{D}) = q) \leq \alpha, \quad \forall F \in H_0.$$

Coverage for an interval estimate is replaced with the notion of *selective coverage*. This setting is depicted in [Figure 2b](#), in which only  $\mathbb{Q}(\mathfrak{D})$  is conditioned on. If  $\mathfrak{D}'$  is unavailable, it is still possible to use these tools as long as the scientist is able to defend a model for the law of  $\mathfrak{D}$  to their peers.

Conditioning on the event  $\{\mathbb{Q}(\mathfrak{D}) = q\}$  transforms any model  $\mathfrak{M}$  for the joint law of  $(\mathfrak{D}, \mathfrak{D}')$  to a new model

$$(15) \quad \mathfrak{M}_q^* = \left\{ F^* : \frac{dF^*}{dF}(d, d') \propto 1_{\{\mathbb{Q}^{-1}(q)\}}(d), F \in \mathfrak{M} \right\}$$

where  $q$  is the value of  $\mathbb{Q}(\mathfrak{D})$  observed by the scientist. We should note that, as in [Figure 2a](#), the model itself has been selected *after* the scientist has observed the patterns  $\mathbb{Q}(\mathfrak{D})$ .

What has this approach bought us? For one thing, we have freed the scientist from the “natural” transformations of questions to statistical objects we saw in our discussion of simultaneous methods. The scientist is free to transform the observed patterns into statistical objects how they see fit.

At what cost has this benefit come? The first cost is that conditional rather than marginal guarantees are required. Conditional guarantees are generally stronger than marginal guarantees, though they may need stricter assumptions to hold. Exploration of the gap between assumptions required for selective and marginal guarantees is certainly an interesting problem.

A second cost is the cost of exploration itself. In the classical scientific method, the scientist is faced with the cost of collecting a second data set  $\mathfrak{D}'$  in order to apply the formal methods of statistical inference. In [Figure 2b](#) the scientist is able to reuse some of the data for inference. How much is available for reuse will depend very much on  $\mathbb{Q}$  – if this is the identity map, then clearly fixing  $\mathbb{Q}(\mathfrak{D})$  is equivalent to fixing  $\mathfrak{D}$  and no data

remains after exploration. If  $\mathfrak{M} = \{f_\theta : \theta \in \Theta\}$  is a parametric model, then the amount of information for estimating  $\theta$  can be quantified by the Fisher information. The model  $\mathfrak{M}_q^*$  will have its own Fisher information, referred to as *leftover Fisher information* in [Fithian, D. Sun, and J. Taylor \[2014\]](#). Ideally, the scientist is able to explore their data in such a way that they can discovering interesting questions while preserving leftover information.

### 3 Examples and implications of the conditional approach

In the setting of [Figure 2b](#) we allow the scientist to posit a model  $\mathfrak{M}$  after having observed  $\mathbb{Q}(\mathfrak{D})$ , though we require that the scientist posit a model for the joint law of  $(\mathfrak{D}, \mathfrak{D}')$  rather than the usual  $\mathfrak{D}'|\mathfrak{D}$ . The scientist, perhaps with the assistance of a statistician, will then declare some statistical objects of interest defined on  $\mathfrak{M}$ : hypothesis tests, point estimates, confidence intervals, etc.

To simplify our presentation we make the assumption that  $(\mathfrak{D}, \mathfrak{D}')$  are (perhaps asymptotically) jointly Gaussian, implying that the patterns the scientist sees are formed by inspecting some approximately linear statistic. As the Gaussian family is parametric, this also implies there is a well-defined notion of leftover information. Of course, many statistical models (and selection procedures) can be reduced (asymptotically) to this setting. We acknowledge that allowing the scientist to view more complicated statistics, such as scatterplots does not obviously fit into this framework and certainly this is worthy of further study. These two observations bring us to the first of several challenges in the conditional setting.

**Challenge 1** (Selective Central Limit Theorem). Without the effect of selection, there is an extensive literature on uses of the CLT to justify approximations in statistical inference. Sequences of models such as  $\mathfrak{M}_{q,n}^*$  do not fit into this classical setting, though sometimes uniform consistency in  $L^p$  and in an appropriate weak sense (i.e. avoiding the impossibility results of [Leeb and Pötscher \[2006\]](#)) along sequences of models  $\mathfrak{M}_n$  can be transferred over to sequences  $\mathfrak{M}_{q,n}^*$  [Tian and J. E. Taylor \[2015\]](#) and [Markovic and J. Taylor \[2016\]](#). We acknowledge that these results are likely suboptimal.

**Challenge 2** (Rich Interactive Selection). Scatterplots are standard tools in exploratory analyses, as are other summaries. Are there realistic mechanisms to release similar information to scientists that are not wasteful in leftover information?

**3.1 The scientific method is inadmissible.** Our first example makes a rather bold claim. In this setting, the scientist has access to  $\mathfrak{D}'$  and the mechanism by which  $\mathbb{Q}(\mathfrak{D})$  is fixed is by fixing (or conditioning) on all of  $\mathfrak{D}$ . This is a finer sigma algebra than that of  $\mathbb{Q}(\mathfrak{D})$ , which means we are conditioning on more than we need to. Statistical objects constructed conditional on  $\mathfrak{D}$  are often inadmissible with concrete dominating procedures. A

more precise statement in terms of hypothesis tests can be found in Theorem 9 of [Fithian, D. Sun, and J. Taylor \[2014\]](#). Though this theorem is stated in terms of data splitting [Cox \[1975\]](#) in which  $\mathfrak{D}$  is part of a full data set and  $\mathfrak{D}'$  the remaining data, it is clearly applicable to the setting where a scientist collects fresh data  $\mathfrak{D}'$ .

For concreteness, consider a simple model for the *file-drawer* filter (in which only large positive  $Z$  statistics are reported) along with a replication study. We can model this as

$$(\mathfrak{D}, \mathfrak{D}') \sim N \left( \begin{pmatrix} \mu \\ \mu \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)$$

being suitably normalized sample means from some population and

$$(16) \quad \mathcal{Q}(\mathfrak{D}) = \begin{cases} 1 & \mathfrak{D} > 2 \\ 0 & \text{otherwise.} \end{cases}$$

and the scientist has observed  $\mathcal{Q}(\mathfrak{D}) = 1$ . The natural replication estimate is

$$\hat{\mu}(\mathfrak{D}, \mathfrak{D}') = \mathfrak{D}'.$$

It is evident that the sufficient statistic for  $\mu$  in  $\mathfrak{M}$  is  $(\mathfrak{D} + \mathfrak{D}')/2$ . It is also clear that this holds conditional on  $\mathcal{Q}(\mathfrak{D})$  as well. Hence, one can Rao-Blackwellize  $\hat{\mu}$

$$\hat{\mu}_{RB}(\mathfrak{D} + \mathfrak{D}') = \mathbb{E}_F(\mathfrak{D}' | \mathcal{Q}(\mathfrak{D}), \mathfrak{D} + \mathfrak{D}').$$

Simple calculations show that for  $\mu \gg 2$  the Rao-Blackwellized estimator is essentially  $(\mathfrak{D} + \mathfrak{D}')/2$  which has variance  $1/2$  compared to 1, the variance of  $\hat{\mu}$ . Confidence intervals for  $\mu$  and tests of hypotheses of the form  $H_0 : \mu = \mu_0$  are also relatively straightforward to construct in the conditional model

$$\mathfrak{M}_1^* = \left\{ F^* : \frac{dF^*}{dF}(d, d') \propto 1_{(2, \infty)}(d), F \in \mathfrak{M} \right\}$$

Such procedures have been proposed in similar but not identical settings [Cohen and Sackrowitz \[1989\]](#) and [Sampson and Sill \[2005\]](#) in which follow-up data  $\mathfrak{D}'$  is available through a designed experiment.

We see a clear gap in performance between the simple estimator arrived at following the classical scientific method and one which conditions on less. We note that this statement of inadmissibility is relative to the sigma algebra we condition on. If it is asserted that the correct sigma algebra conditions on  $\mathfrak{D}$  then  $\hat{\mu}$  is in fact the UMVU and this domination disappears. It is argued in [Bi, Markovic, Xia, and J. Taylor \[2017\]](#) that the minimal sigma algebra to condition is that generated by the patterns the scientist observed as this is the sigma algebra with which the statistical objects is determined. This is very mildly

in contrast to the formal theory laid out in [Fithian, D. Sun, and J. Taylor \[2014\]](#) which presumes that  $\mathbb{Q}(\mathfrak{D})$  is already in the form of statistical objects, though the theory of itself needs no essential change.

Often, it will be convenient to condition on more, perhaps to ensure that computations are feasible. This is allowed for in the formal theory of [Fithian, D. Sun, and J. Taylor \[ibid.\]](#) through a selection variable, and was used in the first (non-data splitting) conditional approach to the regression problem in [Lee, D. L. Sun, Y. Sun, and J. E. Taylor \[2016\]](#). It is not hard to imagine exploratory analyses  $\mathbb{Q}$  whose sigma algebra is simply too complex to describe so that restricting models to events  $\mathbb{Q}^{-1}(q)$  may be computationally infeasible resulting in a costly waste of leftover information. This suggests the seemingly uncontroversial principle of not using too complex an exploratory analysis to generate questions. We see no reason a priori that more complex analyses will lead the scientist to more interesting questions.

**3.1.1 Dominating the scientific method in the regression problem.** [Figure 3](#) from [Tian and J. E. Taylor \[2015\]](#) (which reproduces and extends an example in [Fithian, D. Sun, and J. Taylor \[2014\]](#)) demonstrates inadmissibility in the regression setting in which the LASSO [R. Tibshirani \[1996\]](#) is used to select variables. In this example, the total number of samples is held fixed and the portion allocated to  $\mathfrak{D}$  and  $\mathfrak{D}'$  varies. The vertical axis is Type II error, the complement of statistical power. The horizontal axis is the probability of the selection mechanism discovering all of the true effects in this regression problem, meant to be a proxy for the quality of the patterns revealed to the scientist. We carried out inference for partial correlations in the Gaussian model with features  $E$  selected by  $\mathfrak{D}$ . This seems the natural model a scientist would use for  $\mathfrak{D}'$  if they decided to replicate the study, collecting only features  $E$  discovered in the pilot study. Other statistical objects are certainly reasonable, such as the  $E$  coordinates of the full model (a subset of the targets in model (13)) or the debiased LASSO targets [Javanmard and Montanari \[2013\]](#), [T. Sun and C.-H. Zhang \[2012\]](#), and [Dezeure, Bühlmann, Meier, and Meinshausen \[2015\]](#) if  $n < p$ . The curve labelled *data splitting* follows the classical scientific method, conditioning on  $\mathfrak{D}$ . The curve labelled *data carving* conditions only on  $\mathbb{Q}(\mathfrak{D})$  but makes decisions about exactly the same statistical objects as data splitting. It is clear that data carving dominates data splitting with data splitting having Type II error of different magnitude to the data carving curve and red curve. This red curve brings us to our next example.

**3.2 Noisier is better? Randomized selection algorithms.** The initial description of [Figure 1](#) had  $\mathfrak{D}'$  as fresh data while pilot  $\mathfrak{D}$  was used to discover patterns  $\mathbb{Q}(\mathfrak{D})$ . This is an artificial constraint: the scientist uses  $\mathfrak{D}$  to discover patterns and simply must posit a model for the joint law of  $(\mathfrak{D}, \mathfrak{D}')$ . Inference is then carried out after restricting this

model to the event  $\mathbb{Q}^{-1}(q)$ . In particular,  $\mathfrak{D}$  could be a randomized version of  $\mathfrak{D}'$  with the randomization chosen by the scientist perhaps with the assistance of a statistician [Tian and J. E. Taylor \[2015\]](#). Indeed, our file drawer replication study can easily be seen in this light: let  $Z_1$  denote the pilot data and  $Z_2$  the replication data. We take  $\mathfrak{D}' = (Z_1 + Z_2)/2 \sim N(\mu, \frac{1}{2})$  and

$$\mathfrak{D}|\mathfrak{D}' \sim N\left(\mathfrak{D}', \frac{1}{2}\right).$$

Unsurprisingly, by sufficiency, the law of  $\mathfrak{D}|\mathfrak{D}'$  does not depend on the unknown  $\mu$  and it is enough to consider the law of  $\mathfrak{D}'$  after marginalizing over  $\mathfrak{D}$ . The statistical problem can be reduced to inference in the model

$$(17) \quad \mathfrak{M}_q^* = \left\{ F^* : \frac{dF^*}{dF}(d') \propto \int 1_{\{\mathbb{Q}(\cdot)=q\}}(u) G(du|d'), F \in \mathfrak{M} \right\}$$

with  $G(\cdot|d')$  the kernel representing the conditional law of  $\mathfrak{D}$  given  $\mathfrak{D}'$ .

It is apparent that the Radon-Nikodym derivative or likelihood ratio relating  $F^*$  to its corresponding  $F$  will often be a smooth function. In the case that  $G(\cdot|d') = \delta_{d'}$ , in which case  $\mathfrak{D} \stackrel{\text{a.s.}}{=} \mathfrak{D}'$ , this will in fact be an indicator function. In the setting of Gaussian randomization, the smoothness of this function can be directly related to the leftover information, explaining why both data carving and the additive noise model in [Figure 3](#) show an improvement in power after addition of at least some randomness into the pattern generation stage. With no randomization, the (rescaled) leftover Fisher information can rapidly approach 0 for some parameter values while the corresponding information

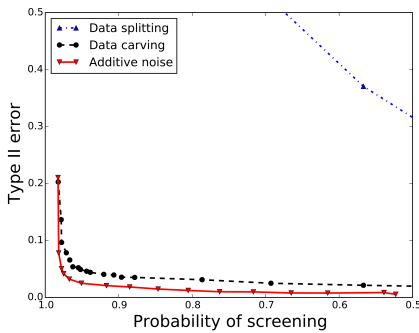


Figure 3: Comparison of inference in additive noise randomization vs. data carving.

after randomization can be bounded below. In this sense some noise is better than no noise, though [Figure 3](#) demonstrates there is a tradeoff between quality of patterns and statistical power.

The red curve in [Figure 3](#) uses the same data generating mechanism as the LASSO in the regression problem with  $\mathfrak{D}' = y$  and  $\mathfrak{D} = y + \omega$  with  $\omega \sim N(0, \tau^2 I)$ . The curve traces out the Type II error and probability of screening as  $\tau$  varies. It seems as if this particular randomization does better than data carving in this figure. However, as

$\mathfrak{D}$  and hence  $\mathbb{Q}(\mathfrak{D})$  differ between the two curves, a direct comparison of data carving to the randomization above is somewhat difficult. In practice, there is no guarantee that any two scientists given access to  $\mathfrak{D}$  will use the same  $\mathbb{Q}$  or construct the same statistical objects having observed the same  $\mathbb{Q}(\mathfrak{D})$ . It is not hard to imagine settings where some scientists know the “right”  $\mathbb{Q}$  to use based on domain experience, or perhaps know the “right” statistical objects to report having observed  $\mathbb{Q}(\mathfrak{D})$ . Such scientists will likely be able to extract more interesting answers from  $\mathfrak{D}'$  and likely more money from funding agencies than others. Identifying such scientists and / or modelling their behavior, or even identifying the “right” statistical objects seems a daunting task which we decline to pursue.

**3.3 Patterns divined from convex programs.** The LASSO [R. Tibshirani \[1996\]](#) is a popular algorithm used to discover important features in the regression context. Let us remind the readers of the LASSO optimization problem

$$(18) \quad \hat{\beta}_{\lambda}(y, X) = \operatorname{argmin}_{\beta} \frac{1}{2} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1.$$

It is well known that for large enough values of  $\lambda$ , the solution will often be sparse. The non-zero entries of  $\hat{\beta}(y, X, \lambda)$  are a natural candidate for the “important” variables in predicting  $y$  from  $X$  in a linear model. Further, the event

$$\left\{ y : \operatorname{sign}(\hat{\beta}_{\lambda}(y, X)) = s \right\}$$

can be described in terms of a set of affine inequalities [Lee, D. L. Sun, Y. Sun, and J. E. Taylor \[2016\]](#). This observation demonstrated that conditional inference in the regression problem was feasible, yielding the *polyhedral lemma* subsequently used in [Heller, Meir, and Chatterjee \[2017\]](#), [R. J. Tibshirani, J. Taylor, Lockhart, and R. Tibshirani \[2016\]](#), and [R. J. Tibshirani, Rinaldo, R. Tibshirani, and Wasserman \[2015\]](#) among other places. The assumption that  $X$  be fixed is not strictly necessary with suitable modification of the covariance estimate in the polyhedral lemma [J. Taylor and R. Tibshirani \[2017\]](#).

**Challenge 3** (High Dimensional Selective Inference). High dimensional inference [Bühlmann and Geer \[2011\]](#) is a very important topic given the sheer size of  $p$  in modern science. Rigorously addressing the conditional approach in this setting is certainly challenging. While some results are available [Markovic, Xia, and J. Taylor \[2017\]](#) and [Wasserman and Roeder \[2009\]](#) much work remains.

The LASSO has inspired many other convex optimization algorithms meant to elucidate interesting structure or patterns in  $\mathfrak{D}$ . A very short list might include [Becker, Candès, and Grant \[2011\]](#), [Chen, Donoho, and Saunders \[1998\]](#), [Ming and Lin \[2005\]](#), and [Yuan and Lin \[2007\]](#). Many natural  $\mathbb{Q}$  suggest themselves from such convex programs. Convex

programs also lend themselves naturally to randomization by perturbation of the objective function. While [Section 2.3](#) describes the reason *why* we condition on  $\mathbb{Q}(\mathfrak{D})$ , it is important to describe *how* we achieve this. Here we describe some of the approach of [Tian, Panigrahi, Markovic, Bi, and J. Taylor \[2016\]](#) which gives a general idea of the *how*.

Consider the problem

$$(19) \quad \hat{\beta}(\mathfrak{D}, \omega) = \operatorname{argmin}_{\beta \in \mathbb{R}^k} \ell(\beta; \mathfrak{D}) + \mathcal{P}(\beta) - \omega^T \beta + \frac{\epsilon}{2} \|\beta\|_2^2$$

where  $\ell$  is some smooth loss involving the data (not necessarily a negative log-likelihood),  $\mathcal{P}$  is some structure inducing convex function,  $\epsilon > 0$  is some small parameter that is sometimes necessary in order to assure the program has a solution and  $\omega \sim G$  is a randomization with  $G$  (typically having a smooth density  $g$ ) chosen by the scientist. In terms of [Figure 2b](#) the randomization  $\omega$  can be modelled as part of the function  $\mathbb{Q}$  and we are free to take  $\mathfrak{D} = \mathfrak{D}'$ .

The KKT conditions of such a problem can be written as

$$(20) \quad \omega = \nabla \ell(\beta; \mathfrak{D}) + u + \epsilon \cdot \beta.$$

with  $(\beta, u)$  required to satisfy

$$(21) \quad u \in \partial \mathcal{P}(\beta).$$

Suppose that the scientist seeks for patterns in the pair  $(\beta, u)$  so that  $\mathbb{Q} = \mathbb{Q}(\mathfrak{D}, \omega) = \bar{\mathbb{Q}}(\hat{\beta}(\mathfrak{D}, \omega), \hat{u}(\mathfrak{D}, \omega))$ . It turns out that, in wide generality, there is a natural mechanism through which we can condition on events expressed in terms of  $(\beta, u)$ . Geometrically, if  $\mathcal{P}$  is a seminorm given by the support function of convex set  $K$ , then the condition is equivalent to  $(u, \beta) \in N(K)$  where  $N(K)$  is the normal bundle of  $K$  and the integral necessary to restrict to the event of interest can be expressed through a change of variables as

$$(22) \quad \mathbb{P}(\mathbb{Q}(\mathfrak{D}, \omega) = q | \mathfrak{D}) = \int_{N(K)} 1_{\{\bar{\mathbb{Q}}^{-1}(q)\}}(u, \beta) g(\phi(u, \beta; \mathfrak{D})) J_{\phi}(u, \beta; \mathfrak{D}) \mathcal{H}_k(d\beta du).$$

where  $g$  is the density of  $\omega$ ,

$$\phi(\beta, u; \mathfrak{D}) = \nabla \ell(\beta; \mathfrak{D}) + u + \epsilon \cdot \beta$$

is the change of variables introduced by inverting the KKT conditions above and  $\mathcal{H}_k$  is  $k$ -dimensional Hausdorff measure on  $N(K) \subset \mathbb{R}^{2k}$ . See [Tian, Panigrahi, Markovic, Bi, and J. Taylor \[ibid.\]](#) for further details and examples beyond the LASSO. With a little work, expanding the Jacobian in the integrals in (22) yield objects closely related to integrals

against the generalized curvature measures of  $K$  [Schneider \[1993\]](#). Much of the work cited above on Gaussian processes and simultaneous inference also involve such geometric objects through Weyl and Steiner’s tube formulae [Adler and J. E. Taylor \[2007\]](#).

In many cases of interest the rather complicated looking (22) can be expressed as

$$(23) \quad \int_{\bar{K}_q} g(A\mathfrak{D} + Bo + \eta) \, do$$

for some  $\eta$  measurable with respect to  $\mathcal{Q}(\mathfrak{D}, \omega)$  and some nice (often convex and polyhedral)  $\bar{K}_q \subset \mathbb{R}^k$  where the variable of integration  $o$  is meant to stand for “optimization variables”, i.e. the pair  $(u, \beta)$  in (22). If we presume that the randomization used by the scientist is Gaussian (e.g. data carving can be represented as asymptotically adding Gaussian randomization [Markovic and J. Taylor \[2016\]](#)), then the likelihood ratio in the model  $\mathfrak{M}_q^*$  can be expressed as

$$(24) \quad \mathbb{P}(o \in \bar{K}_q | \mathfrak{D})$$

for some  $\eta$  measurable with respect to  $\mathcal{Q}(\mathfrak{D}, \omega)$  where the pair  $(\mathfrak{D}, o)$  are jointly Gaussian with mean and covariance determined by the mean and covariance of  $\mathfrak{D}$ , the pair  $(A, B)$  and the covariance matrix of the randomization  $\omega$ .

**3.4 Benjamini-Hochberg given access to replication data.** Some selection algorithms do not involve directly solving a (randomized) convex program such as (19) yet the appropriate likelihood ratio can be described similarly. For instance, suppose  $\mathfrak{D} = Z \sim N(\mu, \Sigma)$  and the selection algorithm involves taking the top  $k$  of the  $Z$ -statistics. A randomized version might add  $\omega \sim N(0, \tau^2 I)$  to  $Z$  before ranking. We can take the map  $\mathcal{Q}(Z, \omega)$  to return the identity of the top  $k$  and perhaps their signs. Formally this is an example of (19) involved in finding the maximizer of the convex function that returns the sum of the top  $k$  order statistics, an example of SLOPE [Bogdan, Berg, Sabatti, Su, and Candès \[2015\]](#). Let  $E_k$  denote the identity of these variables and  $s_k$  their signs. The selection probability can be expressed as

$$(25) \quad \mathbb{P}(o \in \mathfrak{D} + \bar{C}(E_k, s_k) + \eta | \mathfrak{D})$$

where  $\bar{C}(E_k, s_k)$  is the convex cone identifying the top  $k$  coordinates on  $\mathbb{R}^p$  and their signs.

Another example with such a representation is a version the Benjamini-Hochberg algorithm in which  $\mathfrak{D} = Z \sim N(\mu, \Sigma)$  and  $\mathcal{Q}(\mathfrak{D}, \omega)$  identifies which effects are selected by BH using suitably normalized “new”  $Z$ -statistics  $Z + \omega$  as well as the ordering of the non-rejected null  $Z$ -statistics [Reid, J. Taylor, and R. Tibshirani \[2017\]](#). This allows a



scientist to run the BH algorithm on a randomized version of their data, preserving some information for a point estimate or perhaps an interval estimate. Recalling our file drawer replication study, we see that this is mathematically equivalent to setting  $\mathfrak{D}$  to be pilot data and  $\mathfrak{D}'$  to be a replication study. While it is possible to construct intervals for the effects of the variables selected by BH using only  $\mathfrak{D}$  [Benjamini and Yekutieli \[2005\]](#), it is not immediately obvious how one might improve these intervals given replication data. Nor is it clear how one might arrive at unbiased estimates of such parameters using only  $\mathfrak{D}$ , though a simple generalization of our file drawer examples illustrates how to Rao-Blackwellize the replicate unbiased estimate, or hypothesis tests and confidence intervals.

We mention this example to correct what seems to be a misconception in some of the selective inference literature. The conditional approach is sometimes presented as subordinate in a hierarchy of types of simultaneous guarantees [Benjamini \[2010\]](#) and [Benjamini and Bogomolov \[2014\]](#). This is not really the case, the algorithm we just described would have a marginal FDR-control property as well as unbiased estimators of the selected effects arrived at through the conditional approach. In other words, the notion of simultaneous inference in the conditional approach is certainly a well-defined topic of research, see [Hung and Fithian \[2016\]](#) for another example of use of the conditional approach in the simultaneous setting.

**Challenge 4** (Simultaneous Selective Inference). What kind of finite sample procedures can be used to control FDR for a collection of hypotheses generated from  $\mathbb{Q}(\mathfrak{D})$ ? If the selection step has produced a small set of questions, is multiplicity correction still needed?

**3.5 A tractable pseudo-likelihood.** Let us restrict our attention to the case  $\mathfrak{D} = Z \sim N(\mu, \Sigma)$  when the appropriate appropriate can be expressed as in (23). In this setting,  $\mathfrak{M}_q^*$  can be viewed as the marginal law of  $\mathfrak{D}$  under some joint Gaussian law for  $(\mathfrak{D}, o)$  truncated to the event (23) with implied mean of  $(\mathfrak{D}, o)$  some affine function of  $\mu$ . This truncated Gaussian law has normalizing constant

$$(26) \quad \mathbb{P}_\mu(o \in \bar{K}_q).$$

If this normalizing constant were known, the rich toolbox of exponential families would be at our disposal. In [Panigrahi, J. Taylor, and Weinstein \[2016\]](#) we propose using a smoothed version of a Chernoff or large deviations estimate of (26). As the sets  $\bar{K}_q$  are often simple, it is possible to solve this optimization problem quickly. This optimization procedure yields a composite or pseudo-MLE estimate that yields (approximately) conditionally unbiased estimates of  $\mu$  in this setting. Investigation of the performance of this estimator is a topic of ongoing research.

Having (approximately) normalized the likelihood ratio in model  $\mathfrak{M}_q^*$  it is apparent that one may put a prior on  $\mu$  itself. This approach was developed in the univariate setting in

[Yekutieli \[2012\]](#) in which there is often no need to approximate the normalizing constant. Use of this approximation is considered in [Panigrahi, J. Taylor, and Weinstein \[2016\]](#). One may then use all of the advantages of the Bayesian paradigm in this conditional approach, modulo the fact that the likelihood is only approximately normalized.

**3.6 Combining queries: interactive data analysis.** We have alluded to scientists posing questions or queries of  $\mathfrak{D}$  in terms of the solution to a randomized convex program. Of course, a more realistic data analysis paradigm allows the scientist more complex queries. Indeed, the result of one query may inspire a scientist to pose a new query, or perhaps to spend some grant money to collect fresh data. A satisfactory theory of data analysis should be able to handle such situations. Suppose we limit each query to those similar to (23), allowing the results of one query to influence the following queries. Then, it is not hard to see that, after two queries, the (unnormalized) appropriate likelihood ratio takes the form

$$(27) \quad \mathbb{P}(o_1 \in \bar{K}_{q_1} | \mathfrak{D}_1) \cdot \mathbb{P}(o_2 \in \bar{K}_{(q_1, q_2)} | (\mathfrak{D}_1, \mathfrak{D}_2))$$

where each  $o_i$  represent the optimization variables in each query and  $\mathbb{P}$  is some implied joint Gaussian law for the triple  $(\mathfrak{D}_1, \mathfrak{D}_2, o_1, o_2)$  with  $\mathfrak{D}_1$  the data available at time 1,  $(\mathfrak{D}_1, \mathfrak{D}_2)$  at time 2. Generalizing this to  $m$  queries is straightforward. We have named this resulting formalism for inference after allowing a scientist to interact with their data *interactive data analysis* [Bi, Markovic, Xia, and J. Taylor \[2017\]](#).

**Challenge 5** (In Silico Implementation of Interactive Data Analysis). The approximate pseudo-MLE is seen to be a separable convex optimization problem, yielding hope for scaling up to a reasonable number of queries. In the limiting Gaussian model, the relevant reference measures can formally be represented via generalizations of *estimator augmentation* [Tian, Panigrahi, Markovic, Bi, and J. Taylor \[2016\]](#) and [Zhou \[2014\]](#). Some small steps have been taken in this direction but more work is definitely needed.

Other approaches to adaptive data analysis include [Berk, Brown, Buja, K. Zhang, and Zhao \[2013\]](#) in the regression problem, as well as some very interesting work in applications of differential privacy to data analysis [Dwork, Feldman, Hardt, Pitassi, Reingold, and Roth \[2014\]](#).

**Acknowledgments.** All of our work presented here is joint work, particularly with current and former students and colleagues from the Statistical Learning Group in the Department of Statistics at Stanford run jointly with Trevor Hastie and Rob Tibshirani. The author has also profited greatly from discussions with many other colleagues having been given the chance to present this work at numerous conferences.

## References

- Robert J. Adler and Jonathan E. Taylor (2007). *Random fields and geometry*. Springer Monographs in Mathematics. New York: Springer (cit. on pp. 3041, 3051).
- J-M. Azaïs and M. Wschebor (2009). *Level sets and extrema of random processes and fields*. Wiley (cit. on p. 3041).
- Rina Foygel Barber and Emmanuel Candes (Apr. 2014). “Controlling the False Discovery Rate via Knockoffs”. arXiv: 1404.5609 (cit. on pp. 3041, 3042).
- Rina Foygel Barber and Aaditya Ramdas (Sept. 2017). “The p-filter: multilayer false discovery rate control for grouped hypotheses”. en. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 79.4, pp. 1247–1268 (cit. on p. 3042).
- Stephen R. Becker, Emmanuel J. Candès, and Michael C. Grant (2011). “Templates for convex cone problems with applications to sparse signal recovery”. English. *Mathematical Programming Computation* 3.3, pp. 165–218 (cit. on p. 3049).
- Yoav Benjamini (Dec. 2010). “Simultaneous and selective inference: Current successes and future challenges”. eng. *Biometrische Zeitschrift* 52.6, pp. 708–721. PMID: 21154895 (cit. on pp. 3039, 3052).
- Yoav Benjamini and Marina Bogomolov (Jan. 2014). “Selective inference on multiple families of hypotheses”. en. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 76.1, pp. 297–318 (cit. on pp. 3042, 3052).
- Yoav Benjamini and Yosef Hochberg (1995). “Controlling the False Discovery Rate: a Practical and Powerful Approach to Multiple Testing”. *J R Statist Soc B* 57.1, pp. 289–300 (cit. on pp. 3039, 3040).
- Yoav Benjamini and Daniel Yekutieli (Mar. 2005). “False Discovery Rate—Adjusted Multiple Confidence Intervals for Selected Parameters”. *Journal of the American Statistical Association* 100.469, pp. 71–81 (cit. on pp. 3039, 3052).
- Richard Berk, Lawrence Brown, Andreas Buja, Kai Zhang, and Linda Zhao (Apr. 2013). “Valid post-selection inference”. EN. *The Annals of Statistics* 41.2, pp. 802–837. MR: MR3099122 (cit. on pp. 3039, 3041, 3042, 3053).
- Nan Bi, Jelena Markovic, Lucy Xia, and Jonathan Taylor (July 2017). “Inferactive data analysis”. arXiv: 1707.06692 (cit. on pp. 3044, 3046, 3053).
- Małgorzata Bogdan, Ewout van den Berg, Chiara Sabatti, Weijie Su, and Emmanuel J. Candès (Sept. 2015). “SLOPE—Adaptive variable selection via convex optimization”. EN. *The Annals of Applied Statistics* 9.3, pp. 1103–1140. MR: MR3418717 (cit. on p. 3051).
- Peter Bühlmann and Sara van de Geer (June 2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. 1st Edition. Springer (cit. on p. 3049).

- Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv (Oct. 2016). “[Panning for Gold: Model-free Knockoffs for High-dimensional Controlled Variable Selection](#)”. arXiv: [1610.02351](#) (cit. on p. [3042](#)).
- Scott Chen, David Donoho, and Michael Saunders (1998). “[Atomic decomposition for basis pursuit](#)”. *SIAM Journal on Scientific Computing* 20.1, pp. 33–61 (cit. on p. [3049](#)).
- Arthur Cohen and Harold B Sackrowitz (1989). “Two stage conditionally unbiased estimators of the selected mean”. *Statistics & Probability Letters* 8.3, pp. 273–278 (cit. on p. [3046](#)).
- DR Cox (1975). “[A note on data-splitting for the evaluation of significance levels](#)”. *Biometrika* 62.2, pp. 441–444 (cit. on p. [3046](#)).
- Ruben Dezeure, Peter Bühlmann, Lukas Meier, and Nicolai Meinshausen (Nov. 2015). “[High-Dimensional Inference: Confidence Intervals,  \$p\$ -Values and R-Software hdi](#)”. EN. *Statistical Science* 30.4, pp. 533–558. MR: [MR3432840](#) (cit. on p. [3047](#)).
- Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth (Nov. 2014). “[Preserving Statistical Validity in Adaptive Data Analysis](#)”. arXiv: [1411.2664](#) (cit. on p. [3053](#)).
- Bradley Efron (Nov. 2012). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. English. Reprint edition. Cambridge, UK; New York: Cambridge University Press (cit. on pp. [3038](#), [3039](#)).
- William Fithian, Dennis Sun, and Jonathan Taylor (Oct. 2014). “[Optimal Inference After Model Selection](#)”. arXiv: [1410.2597](#) (cit. on pp. [3044–3047](#)).
- Ruth Heller, Amit Meir, and Nilanjan Chatterjee (Nov. 2017). “[Post-selection estimation and testing following aggregated association tests](#)”. arXiv: [1711.00497](#) (cit. on p. [3049](#)).
- Kenneth Hung and William Fithian (Oct. 2016). “[Rank Verification for Exponential Families](#)”. arXiv: [1610.03944](#) (cit. on p. [3052](#)).
- Clifford M Hurvich and Chih—Ling Tsai (1990). “The impact of model selection on inference in linear regression”. *The American Statistician* 44.3, pp. 214–217 (cit. on p. [3039](#)).
- Adel Javanmard and Andrea Montanari (2013). “[Hypothesis testing in high-dimensional regression under the gaussian random design model: Asymptotic theory](#)”. arXiv: [1301.4240](#) (cit. on p. [3047](#)).
- Jason D. Lee, Dennis L. Sun, Yuekai Sun, and Jonathan E. Taylor (June 2016). “[Exact post-selection inference, with application to the lasso](#)”. EN. *The Annals of Statistics* 44.3, pp. 907–927. MR: [MR3485948](#) (cit. on pp. [3039](#), [3044](#), [3047](#), [3049](#)).
- Hannes Leeb and Benedikt M. Pötscher (Oct. 2006). “[Can one estimate the conditional distribution of post-model-selection estimators?](#)” *The Annals of Statistics* 34.5, pp. 2554–2591. MR: [MR2291510](#) (cit. on p. [3045](#)).

- Lihua Lei and William Fithian (Sept. 2016). “AdaPT: An interactive procedure for multiple testing with side information”. arXiv: [1609.06035](#) (cit. on p. [3042](#)).
- Lihua Lei, Aaditya Ramdas, and William Fithian (Oct. 2017). “STAR: A general interactive framework for FDR control under structural constraints”. arXiv: [1710.02776](#) (cit. on p. [3042](#)).
- Ang Li and Rina Foygel Barber (May 2015). “Accumulation tests for FDR control in ordered hypothesis testing”. arXiv: [1505.07352](#) (cit. on p. [3042](#)).
- Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani (Apr. 2014). “A significance test for the lasso”. EN. *The Annals of Statistics* 42.2, pp. 413–468. MR: [MR3210970](#) (cit. on p. [3039](#)).
- Jelena Markovic and Jonathan Taylor (Dec. 2016). “Bootstrap inference after using multiple queries for model selection”. arXiv: [1612.07811](#) (cit. on pp. [3045](#), [3051](#)).
- Jelena Markovic, Lucy Xia, and Jonathan Taylor (Mar. 2017). “Comparison of prediction errors: Adaptive p-values after cross-validation”. arXiv: [1703.06559](#) (cit. on p. [3049](#)).
- Yuan Ming and Yi Lin (2005). “Model selection and estimation in regression with grouped variables”. *Journal of the Royal Statistical Society: Series B* 68.1, pp. 49–67 (cit. on p. [3049](#)).
- Snigdha Panigrahi, Jonathan Taylor, and Asaf Weinstein (2016). “Bayesian Post-Selection Inference in the Linear Model”. arXiv: [1605.08824](#) (cit. on pp. [3052](#), [3053](#)).
- Stephen Reid, Jonathan Taylor, and Robert Tibshirani (June 2017). “Post-selection point and interval estimation of signal sizes in Gaussian samples”. en. *Canadian Journal of Statistics* 45.2, pp. 128–148 (cit. on p. [3051](#)).
- Allan R Sampson and Michael W Sill (2005). “Drop-the-Losers Design: Normal Case”. *Biometrical Journal* 47.3, pp. 257–268 (cit. on p. [3046](#)).
- Rolf Schneider (1993). *Convex bodies: the Brunn-Minkowski theory*. Vol. 44. Encyclopedia of Mathematics and its Applications. Cambridge: Cambridge University Press (cit. on p. [3051](#)).
- D. O Siegmund and K. J Worsley (1995). “Testing for a signal with unknown location and scale in a stationary Gaussian random field”. *The Annals of Statistics* 23.2, pp. 608–639 (cit. on p. [3041](#)).
- John D. Storey (2003). “The positive false discovery rate: a Bayesian interpretation and the  $q$ -value”. *Ann Statist* 31.6, pp. 2013–2035 (cit. on p. [3039](#)).
- Jiayang Sun (Jan. 1993). “Tail Probabilities of the Maxima of Gaussian Random Fields”. *The Annals of Probability* 21.1. ArticleType: research-article / Full publication date: Jan., 1993 / Copyright © 1993 Institute of Mathematical Statistics, pp. 34–71 (cit. on p. [3041](#)).
- Tingni Sun and Cun-Hui Zhang (Dec. 2012). “Scaled sparse linear regression”. *Biometrika* 99.4, pp. 879–898 (cit. on p. [3047](#)).

- A. Takemura and S. Kuriki (2002). “Maximum of Gaussian field on piecewise smooth domain: Equivalence of tube method and Euler characteristic method.” *Ann. of Appl. Prob.* 12.2, pp. 768–796 (cit. on p. 3041).
- Jonathan Taylor and Robert Tibshirani (Mar. 2017). “Post-selection inference for  $\ell_1$ -penalized likelihood models”. en. *Canadian Journal of Statistics* (cit. on p. 3049).
- “Theories of Data Analysis” (n.d.). “Theories of Data Analysis: From Magical Thinking Through Classical Statistics”. In: (cit. on p. 3038).
- Xiaoying Tian, Snigdha Panigrahi, Jelena Markovic, Nan Bi, and Jonathan Taylor (2016). “Selective sampling after solving a convex problem”. arXiv: 1609 . 05609 (cit. on pp. 3050, 3053).
- Xiaoying Tian and Jonathan E. Taylor (July 2015). “Selective inference with a randomized response”. arXiv: 1507.06739 (cit. on pp. 3045, 3047, 3048).
- Robert Tibshirani (1996). “Regression shrinkage and selection via the lasso”. *Journal of the Royal Statistical Society: Series B* 58.1, pp. 267–288 (cit. on pp. 3047, 3049).
- Ryan J. Tibshirani, Alessandro Rinaldo, Robert Tibshirani, and Larry Wasserman (June 2015). “Uniform Asymptotic Inference and the Bootstrap After Model Selection”. arXiv: 1506.06266 (cit. on p. 3049).
- Ryan J. Tibshirani, Jonathan Taylor, Richard Lockhart, and Robert Tibshirani (Apr. 2016). “Exact Post-Selection Inference for Sequential Regression Procedures”. *Journal of the American Statistical Association* 111.514, pp. 600–620 (cit. on p. 3049).
- John W. Tukey (1980). “We Need Both Exploratory and Confirmatory”. *The American Statistician* 34.1, pp. 23–25 (cit. on p. 3037).
- Larry Wasserman and Kathryn Roeder (Oct. 2009). “High-dimensional variable selection”. EN. *The Annals of Statistics* 37.5. Zentralblatt MATH identifier: 05596898; Mathematical Reviews number (MathSciNet): MR2543689, pp. 2178–2201 (cit. on pp. 3039, 3049).
- Daniel Yekutieli (June 2012). “Adjusted Bayesian inference for selected parameters”. en. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 74.3, pp. 515–541 (cit. on p. 3053).
- Ming Yuan and Yi Lin (2007). “Model selection and estimation in the Gaussian graphical model”. *Biometrika* 94.1, pp. 19–35 (cit. on p. 3049).
- Qing Zhou (Oct. 2014). “Monte Carlo Simulation for Lasso-Type Problems by Estimator Augmentation”. *Journal of the American Statistical Association* 109.508, pp. 1495–1516. arXiv: 1401.4425 (cit. on p. 3053).

Received 2017-12-01.



# DIFFUSIVE AND SUPER-DIFFUSIVE LIMITS FOR RANDOM WALKS AND DIFFUSIONS WITH LONG MEMORY

BÁLINT TÓTH

## Abstract

We survey recent results of normal and anomalous diffusion of two types of random motions with long memory in  $\mathbb{R}^d$  or  $\mathbb{Z}^d$ . The first class consists of random walks on  $\mathbb{Z}^d$  in divergence-free random drift field, modelling the motion of a particle suspended in time-stationary incompressible turbulent flow. The second class consists of self-repelling random diffusions, where the diffusing particle is pushed by the negative gradient of its own occupation time measure towards regions less visited in the past. We establish normal diffusion (with square-root-of-time scaling and Gaussian limiting distribution) in three and more dimensions and typically anomalously fast diffusion in low dimensions (typically, one and two). Results are quoted from various papers published between 2012–2017, with some hints to the main ideas of the proofs. No technical details are presented here.

## 1 Random walks in divergence-free random drift field

**1.1 Set-up and notation.** Let  $(\Omega, \mathcal{F}, \pi, \tau_z : z \in \mathbb{Z}^d)$  be a probability space with an ergodic  $\mathbb{Z}^d$ -action. Denote by  $\mathcal{E} := \{k \in \mathbb{Z}^d : |k| = 1\}$  the set of possible steps of a nearest-neighbour walk on  $\mathbb{Z}^d$ , and let  $p_k : \Omega \rightarrow [0, s^*]$ ,  $k \in \mathcal{E}$ , be bounded measurable functions. These will be the jump rates of the RWRE considered (see (2) below) and assume they are *doubly stochastic*,

$$(1) \quad \sum_{k \in \mathcal{E}} p_k(\omega) = \sum_{k \in \mathcal{E}} p_{-k}(\tau_k \omega).$$

---

Supported by EPSRC (UK) Established Career Fellowship EP/P003656/1 and by OTKA (HU) K-109684.

MSC2010: primary 60F05; secondary 60G99, 60K35, 60K37.

Keywords: random walk in random environment, self-repelling Brownian polymer, scaling limit, central limit theorem, anomalous diffusion, martingale approximation, resolvent methods.



Given these, define the continuous time nearest neighbour random walk  $t \mapsto X(t) \in \mathbb{Z}^d$  as a Markov process on  $\mathbb{Z}^d$ , with  $X(0) = 0$  and conditional jump rates

$$(2) \quad \mathbf{P}_\omega(X(t+dt) = x+k \mid X(t) = x) = p_k(\tau_x \omega) dt,$$

where the subscript  $\omega$  denotes that the random walk  $X(t)$  is a Markov process on  $\mathbb{Z}^d$  *conditionally*, with fixed  $\omega \in \Omega$ , sampled according to  $\pi$ . The continuous time setup is for convenience only. Since the jump rates are bounded this is fully equivalent with a discrete time walk.

We will use the notation  $\mathbf{P}_\omega(\cdot)$  and  $\mathbf{E}_\omega(\cdot)$  for *quenched* probability and expectation. That is: probability and expectation with respect to the distribution of the random walk  $X(t)$ , *conditionally, with given fixed environment*  $\omega$ . The notation  $\mathbf{P}(\cdot) := \int_\Omega \mathbf{P}_\omega(\cdot) d\pi(\omega)$  and  $\mathbf{E}(\cdot) := \int_\Omega \mathbf{E}_\omega(\cdot) d\pi(\omega)$  will be used for *annealed* probability and expectation. That is: probability and expectation with respect to the random walk trajectory  $X(t)$  *and* the environment  $\omega$ , averaged out with the distribution  $\pi$ .

It is well known (and easy to check, see e.g. [Kozlov \[1985a\]](#)) that due to double stochasticity (1) the annealed set-up is stationary and ergodic in time: the process of the environment as seen from the position of the random walker

$$(3) \quad \eta(t) := \tau_{X(t)} \omega$$

is a stationary and ergodic Markov process on  $(\Omega, \pi)$  and, consequently, the random walk  $t \mapsto X(t)$  will have stationary and ergodic annealed increments.

The local *quenched* drift of the random walk is

$$\mathbf{E}_\omega(dX(t) \mid X(t) = x) = \sum_{k \in \mathcal{E}} k p_k(\tau_x \omega) dt =: \varphi(\tau_x \omega) dt.$$

It is convenient to separate the symmetric and skew-symmetric part of the jump rates: for  $k \in \mathcal{E}$ , let  $s_k : \Omega \rightarrow [0, s^*]$ ,  $v_k : \Omega \rightarrow [-s^*, s^*]$ ,

$$(4) \quad s_k(\omega) := \frac{p_k(\omega) + p_{-k}(\tau_k \omega)}{2}, \quad v_k(\omega) := \frac{p_k(\omega) - p_{-k}(\tau_k \omega)}{2}.$$

Note that from the definitions (4) it follows that

$$(5) \quad s_k(\omega) - s_{-k}(\tau_k \omega) = 0, \quad v_k(\omega) + v_{-k}(\tau_k \omega) = 0.$$

In addition, the bi-stochasticity condition (1) is equivalent to

$$(6) \quad \sum_{k \in \mathcal{E}} v_k(\omega) \equiv 0, \quad \pi\text{-a.s.}$$

The second identity in (5) and (6) jointly mean that  $(v_k(\tau_x \omega))_{k \in \mathcal{E}, x \in \mathbb{Z}^d}$  is a stationary *sourceless flow* (or, a *divergence-free lattice vector field*) on  $\mathbb{Z}^d$ . The physical interpretation of the divergence-free condition (6) is that the walk (2) models the motion of a particle suspended in stationary, *incompressible flow*, with thermal noise.

In order that the walk  $t \mapsto X(t)$  have *zero annealed mean drift* we assume that for all  $k \in \mathcal{E}$

$$(7) \quad \int_{\Omega} v_k(\omega) d\pi(\omega) = 0.$$

Our next assumption is an *ellipticity* condition for the symmetric part of the jump rates: there exists another positive constant  $s_* \in (0, s^*]$  such that for  $\pi$ -almost all  $\omega \in \Omega$  and all  $k \in \mathcal{E}$

$$(8) \quad s_k(\omega) \geq s_*, \quad \pi\text{-a.s.}$$

Note that the ellipticity condition is imposed only on the symmetric part  $s_k$  of the jump rates and not on the jump rates  $p_k$ . It may happen that  $\pi(\{\omega : \min_{k \in \mathcal{E}} p_k(\omega) = 0\}) > 0$ , as it is the case in some of the examples given in Section 1.4.

Finally, we formulate the notorious  $\mathcal{H}_{-1}$ -condition which plays a key role in diffusive scaling limits. Denote for  $i, j = 1, \dots, d$ ,  $x \in \mathbb{Z}^d$ ,  $p \in [-\pi, \pi)^d$ ,

$$(9) \quad C_{ij}(x) := \int_{\Omega} \varphi_i(\omega) \varphi_j(\tau_x \omega) d\pi(\omega), \quad \widehat{C}_{ij}(p) := \sum_{x \in \mathbb{Z}^d} e^{\sqrt{-1}x \cdot p} C_{ij}(x).$$

That is:  $C_{ij}(x)$  is the covariance matrix of the drift field, and  $\widehat{C}_{ij}(p)$  is its Fourier-transform.

By Bochner's theorem, the Fourier transform  $\widehat{C}$  is positive definite  $d \times d$ -matrix-valued-measure on  $[-\pi, \pi)^d$ . The no-drift condition (7) is equivalent to  $\widehat{C}_{ij}(\{0\}) = 0$ , for all  $i, j = 1, \dots, d$ . With slight abuse of notation we denote this measure formally as  $\widehat{C}_{ij}(p) dp$  even though it could be not absolutely continuous with respect to Lebesgue.

The  $\mathcal{H}_{-1}$ -condition is the following:

$$(10) \quad \int_{[-\pi, \pi)^d} \left( \sum_{j=1}^d (1 - \cos p_j) \right)^{-1} \sum_{i=1}^d \widehat{C}_{ii}(p) dp < \infty.$$

This is an *infrared bound* on the correlations of the drift field,  $x \mapsto \varphi(\tau_x \omega) \in \mathbb{R}^d$ . It implies diffusive upper bound on the annealed variance of the walk and turns out to be a natural sufficient condition for the diffusive scaling limit (that is, CLT for the annealed walk). We'll see further below some other equivalent formulations of the  $\mathcal{H}_{-1}$ -condition (10). Note that the  $\mathcal{H}_{-1}$ -condition (10) formally implies the no-drift condition (7).

For later reference we state here the closely analogous problem of diffusion in divergence-free random drift field. Let  $(\Omega, \mathcal{F}, \pi, \tau_z : z \in \mathbb{R}^d)$  be now a probability space with an ergodic  $\mathbb{R}^d$ -action, and  $F : \Omega \times \mathbb{R}^d \rightarrow \mathbb{R}^d$  a stationary vector field which is  $\pi$ -almost-surely  $C^1$  and divergence-free:

$$(11) \quad \operatorname{div} F \equiv 0, \quad \pi\text{-a.s.}$$

The diffusion considered is

$$(12) \quad dX(t) = dB(t) + F(X(t))dt.$$

The SDE (12) has unique strong solution,  $\pi$ -almost surely. The main question is the same as in the case of the random walk (2): What is the asymptotic scaling behaviour and scaling limit of  $X(t)$ , as  $t \rightarrow \infty$ ? Under what conditions does the central limit theorem with diffusive scaling and Gaussian limit distribution hold? Although the physical phenomena described by (2)-(1) and (12)-(11) are very similar, the technical details of various proofs are not always the same. In particular, PDE methods and techniques used for the diffusion problem (12)-(11) are not always easily implementable for the lattice problem (2)-(1). On the other hand, often restrictive local regularity conditions must be imposed on the diffusion problem (12)-(11).

The results reported in this section refer mainly to the random walk problem (2)-(1). The diffusion problem (12)-(11) will be tangentially mentioned in an example in [Section 1.5.3](#) and in the historical notes of [Section 1.6](#).

**1.2 The infinitesimal generator of the environment process.** All forthcoming analysis will be done in the Hilbert space  $\mathcal{H} := \{f \in \mathcal{L}^2(\Omega, \pi) : \int_{\Omega} f(\omega) d\pi(\omega) = 0\}$ . The  $\mathcal{L}^2(\Omega, \pi)$ -gradients and Laplacian are bounded operators on  $\mathcal{H}$ :

$$\nabla_k f(\omega) := f(\tau_k \omega) - f(\omega) \quad \Delta := 2 \sum_{k \in \mathcal{E}} \nabla_k = - \sum_{k \in \mathcal{E}} \nabla_{-k} \nabla_k.$$

Note that  $\Delta$  is self-adjoint and negative. Thus, the operators  $|\Delta|^{1/2}$  and  $|\Delta|^{-1/2}$  are defined in terms of the spectral theorem. The domain of the unbounded operator  $|\Delta|^{-1/2}$  is

$$\mathcal{H}_{-1} := \{\phi \in \mathcal{H} : \lim_{\lambda \searrow 0} (\phi, (\lambda I - \Delta)^{-1} \phi)_{\mathcal{H}} < \infty\}.$$

The  $\mathcal{H}_{-1}$ -condition gets its name from the fact that (10) is equivalent to requesting that for  $k \in \mathcal{E}$ ,

$$(13) \quad v_k \in \mathcal{H}_{-1}.$$

We will also use the multiplication operators  $M_k, N_k : \mathfrak{L}^2(\Omega, \pi) \rightarrow \mathfrak{L}^2(\Omega, \pi), k \in \mathfrak{E}$ ,

$$M_k f(\omega) := v_k(\omega) f(\omega), \quad N_k f(\omega) := (s_k(\omega) - s_*) f(\omega).$$

The following commutation relations are direct consequences of (in fact, equivalent with) (5) and (6)

$$(14) \quad \sum_{k \in \mathfrak{E}} M_k \nabla_k = - \sum_{k \in \mathfrak{E}} \nabla_{-k} M_k, \quad \sum_{k \in \mathfrak{E}} N_k \nabla_k = \sum_{k \in \mathfrak{E}} \nabla_{-k} N_k$$

We will denote

$$S := -\frac{s_*}{2} \Delta + \sum_{k \in \mathfrak{E}} N_k \nabla_k, = S^* \quad A := \sum_{k \in \mathfrak{E}} M_k \nabla_k, = -A^*.$$

The infinitesimal generator  $L$  of the Markovian semigroup  $P_t : \mathfrak{L}^2(\Omega, \pi) \rightarrow \mathfrak{L}^2(\Omega, \pi)$  of the environment process (3) is

$$L = -S + A.$$

Note that due to ellipticity (8) and boundedness of the jump rates the (absolute value of the) Laplacian minorizes and majorizes the self-adjoint part of the infinitesimal generator:  $s_* |\Delta| \leq 2S \leq s^* |\Delta|$ . The inequalities are meant in operator sense.

**1.3 Helmholtz's theorem and the stream tensor.** In its most classical form Helmholtz's theorem states that in  $\mathbb{R}^3$  (under suitable conditions of moderate increase at infinity) a divergence-free vector field can be realised as the curl (or rotation) of another vector field, called the vector potential. Helmholtz's theorem in our context is the following:

**Proposition 1.** *Let  $v : \Omega \rightarrow \mathbb{R}^{\mathfrak{E}}$  be such that  $v_k \in \mathcal{H}$ , and assume that (5) and (6) hold.*

(i) *There exists a zero mean, square integrable, antisymmetric tensor cocycle  $H : \Omega \times \mathbb{Z}^d \rightarrow \mathbb{R}^{\mathfrak{E} \times \mathfrak{E}}, H_{k,l}(\cdot, x) \in \mathcal{H}$ :*

$$(15) \quad H_{k,l}(\omega, y) - H_{k,l}(\omega, x) = H_{k,l}(\tau_x \omega, y - x) - H_{k,l}(\tau_x \omega, 0),$$

$$(16) \quad H_{l,k}(\omega, x) = H_{-k,l}(\omega, x + k) = H_{k,-l}(\omega, x + l) = -H_{k,l}(\omega, x),$$

such that

$$(17) \quad v_k(\tau_x \omega) = \sum_{l \in \mathfrak{E}} H_{k,l}(\omega, x).$$

The realization of the tensor field  $H$  is unique up to an additive term  $H_{k,l}^0(\omega)$ , not depending on  $x \in \mathbb{Z}^d$  (but obeying the symmetries (16)).

(ii) The  $\mathcal{H}_{-1}$ -condition (10)/(13) holds if and only if the cocycle  $H$  in (i) is stationary. That is, there exists  $h : \Omega \rightarrow \mathbb{R}^{\mathcal{E} \times \mathcal{E}}$ , with  $h_{k,l} \in \mathcal{H}$ , such that

$$(18) \quad h_{l,k}(\omega) = h_{-k,l}(\tau_k \omega) = h_{k,-l}(\tau_l \omega) = -h_{k,l}(\omega),$$

and

$$(19) \quad v_k(\omega) = \sum_{l \in \mathcal{E}} h_{k,l}(\omega).$$

The tensor field  $H$  is realized as the stationary lifting of  $h$ :  $H_{k,l}(\omega, x) = h_{k,l}(\tau_x \omega)$ .

The fact that  $v$  is expressed in (17) as the curl of the tensor field  $H$  having the symmetries (16), is essentially the lattice-version of Helmholtz's theorem. Note that (16) means that the stream tensor field  $x \mapsto H(\omega, x)$  is in fact a function of the *oriented plaquettes* of  $\mathbb{Z}^d$ . In particular, in two-dimensions  $x \mapsto H(\omega, x)$  defines a *height function* with stationary increments, on the dual lattice  $\mathbb{Z}^2 + (1/2, 1/2)$ , in three-dimensions  $x \mapsto H(\omega, x)$  defines an *oriented flow* (that is: a lattice vector field) with stationary increments on the dual lattice  $\mathbb{Z}^3 + (1/2, 1/2, 1/2)$ . In Helmholtz's theorem, if  $d > 2$ , there is much freedom in the choice of the *gauge* of  $H$ . The cocycle condition (15) is met by the *Coulomb gauge*, which makes the construction essentially uniquely determined.

In Kozma and Tóth [2017] it was shown that for a RWRE (2) whose environment satisfies conditions (1), (8) and (10) the central limit theorem holds, under diffusive scaling and Gaussian limit with finite and nondegenerate asymptotic covariance, *in probability with respect to the environment*. See Theorem 1 below.

In order to obtain the quenched version, that is central limit theorem for the displacement  $X(t)$  at late times, with frozen environment,  $\pi$ -almost surely, we impose a slightly stronger integrability condition on the stream-tensor-field,

$$(20) \quad h \in \mathcal{L}^{2+\varepsilon}(\Omega, \pi),$$

for some  $\varepsilon > 0$ , rather than being merely square integrable. This stronger integrability condition is needed in the proof of quenched tightness of the diffusively scaled displacement  $t^{-1/2}X(t)$ . We will refer to the  $\mathcal{H}_{-1}$ -condition complemented with the stronger integrability assumption (20) as the *turbo- $\mathcal{H}_{-1}$ -condition*. In Tóth [n.d.] the quenched version of the central limit theorem for the displacement of the random walker was proved under the conditions (1), (8) and (10) and (20). See Theorem 3 below.

**1.4 Examples. Bounded stream tensor:** Let  $((\chi_{ij}(x))_{1 \leq i < j \leq d})_{x \in \mathbb{Z}^d}$ , be a stationary and ergodic (with respect to  $x \in \mathbb{Z}^d$ ) sequence of bounded random variables (say,  $|\chi_{ij}(x)| \leq$

1), and extend them to  $i, j \in \{1, \dots, d\}$  as  $\chi_{ji} = -\chi_{ij}$ ,  $\chi_{ii} = 0$ . Define for  $k, l \in \mathcal{E}$ ,  $x \in \mathbb{Z}^d$ ,

$$H_{k,l}(x) := (k \cdot e_i)(l \cdot e_j)\chi_{ij}(x + (k \cdot e_i - 1)e_i/2 + (l \cdot e_j - 1)e_j/2).$$

(This formula extends the random variables  $\chi$  to a tensor field, consistent with the symmetries (16)). Define  $v_k(\omega)$  as in (17) and let  $s_k(\omega) \equiv s_* \geq 2(d-1)$ . This is the most general construction of the case when  $h \in \mathcal{L}^\infty(\Omega, \pi)$ . In particular, it covers those cases when the random environment of jump probabilities admits a *bounded cycle representation*, cf. [Kozlov \[1985a\]](#), [Komorowski and Olla \[2003\]](#), [Deuschel and K\"osters \[2008\]](#), [Komorowski, Landim, and Olla \[2012\]](#) (chapter 3.3). Due to Proposition 1 the  $\mathcal{H}_{-1}$ -condition (10)/(13) holds.

**Randomly oriented Manhattan lattice:** Let  $u_i(y)$ ,  $i \in \{1, \dots, d\}$ ,  $y \in \mathbb{Z}^{d-1}$ , be i.i.d. random variables with the common distribution  $\mathbf{P}(u = \pm 1) = 1/2$ , and define for  $x \in \mathbb{Z}^d$  and  $k \in \mathcal{E}$

$$v_k(\tau_x \omega) := \sum_{i=1}^d (k \cdot e_i) u_i(x_1, \dots, x_{i-1}, \cancel{x_i}, x_{i+1}, \dots, x_d),$$

One can easily compute the covariances (9):  $C_{ij}(x) = \delta_{i,j} \prod_{i' \neq i} \delta_{x_{i'}, 0}$ , and their Fourier transforms  $\widehat{C}_{ij}(p) = \delta_{i,j} \delta(p_i)$ . From here it follows that in this particular model the  $\mathcal{H}_{-1}$ -condition fails robustly (with power law divergence in (10)) in  $d = 2$ , fails marginally (with logarithmic divergence in (10)) in  $d = 3$ , and holds if  $d \geq 4$ .

**$\binom{2d}{d}$ -vertex models on  $\mathbb{Z}^d$ :** Let  $\Omega$  be the set of all possible orientations of the edges of  $\mathbb{Z}^d$  with the constraint that at all vertices the number of edges oriented towards the site is equal to the number of edges oriented away,  $d$  out of  $2d$ . In this way, locally at every vertex  $\binom{2d}{d}$  configurations of orientations are possible and there is a very rigid constraint on the configurations.  $\Omega$  is a compact metric space and the group of translations  $\tau_z : \Omega \rightarrow \Omega$ ,  $z \in \mathbb{Z}^d$ , acts naturally on it. Let, for  $k \in \mathcal{E}$ ,  $v_k(\omega) = \pm 1$  be the orientation of the edge  $\overline{0k}$  in the configuration  $\omega \in \Omega$ , and  $p_k(\omega) = 1 + v_k(\omega)$ . Any translation invariant ergodic measure  $\pi$  on  $\Omega$  realizes a model of our RWRE. The most natural choice is the one when  $\pi$  is the unique weak limit of the uniform distribution of the allowed  $\binom{2d}{d}$ -vertex configurations on the discrete torus  $(-L, L] \times \dots \times (-L, L]$ , with periodic boundary conditions, as  $L \rightarrow \infty$ . In 2-dimensions this is the notorious (uniform) six-vertex model. In this case - in 2-dimensions - the  $\mathcal{H}_{-1}$ -condition fails: the integral in (10) is logarithmically divergent.

## 1.5 Scaling limits.

### 1.5.1 Central limit theorem in probability w.r.t. the environment under the $\mathcal{H}_{-1}$ condition.

**Theorem 1.** (Source: [Kozma and Tóth \[2017\]](#)) *Conditions (1), (8), (10) are assumed. The asymptotic annealed covariance matrix*

$$(21) \quad (\sigma^2)_{ij} := \lim_{t \rightarrow \infty} t^{-1} \mathbf{E} (X_i(t) X_j(t))$$

*exists, and it is finite and non-degenerate. For any bounded and continuous function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ ,*

$$(22) \quad \lim_{T \rightarrow \infty} \int_{\Omega} \left| \mathbf{E}_{\omega} \left( f(T^{-1/2} X(T)) \right) - \int_{\mathbb{R}^d} \frac{e^{-\frac{|y|^2}{2}}}{(2\pi)^{\frac{d}{2}}} f(\sigma^{-1} y) dy \right| d\pi(\omega) = 0.$$

[Theorem 1](#) is proved in [Kozma and Tóth \[ibid.\]](#), and, weak convergence in the sense of (22) of all finite dimensional marginal distributions of  $t \mapsto T^{-\frac{1}{2}} X(Tt)$ , as  $T \rightarrow \infty$ , to those of a  $d$ -dimensional Brownian motion with covariance  $\sigma^2$  is established. We sketch the main points. Start with a most natural martingale decomposition

$$(23) \quad X(t) = \left\{ X(t) - \int_0^t \varphi(\eta(s)) ds \right\} + \int_0^t \varphi(\eta(s)) ds =: M(t) + I(t).$$

In this decomposition  $M(t)$  is clearly a square integrable martingale with stationary and ergodic annealed increments. The main issue is an efficient martingale approximation of the term  $I(t)$ , à la Kipnis-Varadhan.

We rely on the following general result on Kipnis-Varadhan type of martingale approximation. Let  $\eta(t)$  be a stationary and ergodic Markov process on the probability space  $(\Omega, \pi)$ , and  $L$  be the infinitesimal generator of its Markovian semigroup acting on  $\mathfrak{L}^2(\Omega, \pi)$ . Denote  $S := -(L + L^*)/2$ ,  $A := (L - L^*)/2$  and assume that the symmetric part  $S$  is minorised and majorized by a positive operator  $D \geq 0$ :  $s_* D \leq S \leq s^* D$ , with  $0 < s_* \leq s^* < \infty$ . Further, denote

$$B_{\lambda} := (\lambda I + D)^{-1/2} A (\lambda I + D)^{-1/2}.$$

**Theorem 2.** (Source: [Horváth, Tóth, and Vető \[2012b\]](#), [Kozma and Tóth \[2017\]](#)) *Assume that there exist a dense subspace  $\mathfrak{B} \subseteq \mathfrak{L}^2(\Omega, \pi)$  and a linear operator  $B : \mathfrak{B} \rightarrow \mathfrak{L}^2(\Omega, \pi)$  which is essentially skew-self-adjoint on the core  $\mathfrak{B}$  and such that for any  $\varphi \in \mathfrak{B}$  there exists a sequence  $\varphi_{\lambda} \in \mathfrak{L}^2(\Omega, \pi)$  such that*

$$\lim_{\lambda \rightarrow 0} \|\varphi_{\lambda} - \varphi\| = 0. \quad \text{and} \quad \lim_{\lambda \rightarrow 0} \|B_{\lambda} \varphi_{\lambda} - B\varphi\| = 0.$$

Then for any  $f \in \text{Dom}(D^{-1/2})$  there exists a martingale  $M_f(t)$  (on the probability space and with respect to the filtration of the Markov process  $t \mapsto \eta_t$ ) such that

$$\lim_{t \rightarrow \infty} t^{-1} \int_{\Omega} \mathbf{E}_{\omega} \left( \left| \int_0^t f(\eta(s)) ds - M_f(t) \right|^2 \right) = 0.$$

In plain and informal words: if the operator  $B = D^{-1/2}AD^{-1/2}$  makes sense as a densely defined unbounded *skew-self-adjoint* operator then integrals along the Markov process trajectory of functions in  $\mathcal{H}_{-1} \subset \mathcal{L}^2(\Omega, \pi)$  defined in terms of the positive operator  $D$  are efficiently approximated with martingales, à la Kipnis-Varadhan. As shown in Horváth, Tóth, and Vető [2012b], the condition of Theorem 2 is weaker than the graded sector condition of Sethuraman, Varadhan, and H.-T. Yau [2000] (which in turn is weaker than the strong sector condition of Varadhan [1995]).

In our particular case define  $B : \mathcal{H}_{-1} \rightarrow \mathcal{H}$  as

$$(24) \quad B := - \sum_{l \in \mathcal{E}} |\Delta|^{-1/2} \nabla_{-l} M_l |\Delta|^{-1/2}.$$

(Note that the operators  $|\Delta|^{-1/2} \nabla_{-l}$ ,  $l \in \mathcal{E}$  are bounded.) From the commutation relations (14) it follows that the operator  $B$  is *skew-symmetric* on the dense subspace  $\mathfrak{B} := \mathcal{H}_{-1}$ . It is not difficult to show that if  $h \in \mathcal{L}^{\infty}$  then  $B$  is a bounded operator and thus, its skew-self-adjointness drops out for free. (This is essentially the same as Varadhan's strong sector condition, cf. Varadhan [ibid.]) On the other hand, if  $h \notin \mathcal{L}^{\infty}$  then  $B$  is genuinely unbounded and proving its (essential) skew-self-adjointness is far from trivial.

The main technical result in Kozma and Tóth [2017] is the proof of the fact that  $B$  is in fact *essentially skew-self-adjoint* on  $\mathcal{H}_{-1}$ . By applying von Neumann's criterion for (skew-)self-adjointness this boils down to proving that the lattice PDE

$$(25) \quad \Delta \Psi(\cdot, \omega) + V(\cdot, \omega) \cdot \nabla \Psi(\cdot, \omega) = 0,$$

does not have a non-trivial cocycle solution  $\Psi(x, \omega)$ . Here now  $\nabla$  and  $\Delta$  denote the lattice gradient, respectively, the lattice Laplacian,  $V(x, \omega) = v(\tau_x \omega)$  and  $\Psi(x, \omega)$  is a zero mean cocycle to be determined.

### 1.5.2 Quenched central limit theorem under the turbo- $\mathcal{H}_{-1}$ -condition.

**Theorem 3.** (Source: Tóth [n.d.]) *Conditions (1), (8), (10), and (20) are assumed. For any bounded continuous function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ ,*

$$\lim_{T \rightarrow \infty} \mathbf{E}_{\omega} \left( f(T^{-1/2} X(T)) \right) = \int_{\mathbb{R}^d} \frac{e^{-\frac{|y|^2}{2}}}{(2\pi)^{\frac{d}{2}}} f(\sigma^{-1} y) dy, \quad \pi\text{-a.s.}$$

with the non-degenerate covariance matrix  $\sigma^2$  given in (21).



**Theorem 3** is proved in [Tóth \[n.d.\]](#), and as in the case of [Theorem 1](#), the weak convergence of all finite dimensional distributions follows. The proof consists of three major steps: (1) Proof of *quenched tightness* of the scaled displacement  $t^{-1/2}X(t)$ , as  $t \rightarrow \infty$ . (2) Construction of the *harmonic coordinates* for the walk. (3) Efficient estimate of the discrepancy between the actual position of the walker and the approximating harmonic coordinates.

**Quenched tightness of  $t^{-1/2}X(t)$ :** The proof relies on adapting Nash's moment bound on reversible diffusions with strictly elliptic and bounded dispersion coefficients, cf. [Nash \[1958\]](#), to this type of non-reversible setup. The extension in the case of  $h \in \mathcal{L}^\infty(\Omega, \pi)$  is essentially straightforward, following [Osada \[1983\]](#). (Though, adaptation to the lattice walk case needs some attention.) The extension to  $h \in \mathcal{L}^{2+\varepsilon}(\Omega, \pi)$  is tricky. An integration over time and the Chacon-Ornstein ergodic theorem help. Full details can be found in [Tóth \[n.d.\]](#). This is the only part of the proof where the stronger integrability condition (20) is used.

**Harmonic coordinates:** The idea of harmonic coordinates for random walks in random environments originates in the classical works [Kozlov \[1979\]](#), [G. C. Papanicolaou and Varadhan \[1981\]](#), [Osada \[1983\]](#), [Kozlov \[1985a\]](#). Since then it had been widely used in proving quenched central limit theorems, mostly for random walks among random conductances. That is: in time reversible cases. See, however, [Deuschel and Kösters \[2008\]](#) for a non-reversible application. The idea is very natural: find an  $\mathbb{R}^d$ -valued  $\mathcal{L}^2(\Omega, \pi)$ , zero mean random cocycle  $\Theta(x, \omega)$ , such that

$$(26) \quad \sum_{k \in \mathcal{E}} p_k(\tau_x \omega) (k + \Theta(\omega, x + k) - \Theta(\omega, x)) = 0, \quad \pi\text{-a.s.}$$

If there exists a solution  $\Theta$  to the equation (26) then the process  $t \mapsto Y(t) := X(t) + \Theta(\omega, X(t))$  is a quenched martingale (that is, a martingale in its own filtration, with the environment  $\omega \in \Omega$  frozen). It turns out that equation (26) is equivalent with the following equation in  $\mathcal{H}$ :

$$(27) \quad (I + B^*)\chi = |\Delta|^{-1/2} \varphi,$$

where  $\varphi \in \mathcal{H}_{-1}$  is given and  $\chi$  is to be determined. The operator  $B^*$  on the left hand side is exactly the adjoint of  $B$  from (24). Since it was proved that the operator  $B$  is skew-self-adjoint it follows that  $I + B^*$  is invertible and thus equation (27) has a unique solution. As a consequence, equation (26) also has a unique solution  $\Theta$  which is an  $\mathbb{R}^d$ -valued cocycle, as required.

Once the harmonic coordinates are constructed the quenched central limit theorem for  $t^{-1/2}Y(t)$  drops out via the martingale CLT, using ergodicity of the environment process (3).

**Error bound:** In order to obtain the quenched CLT for the scaled displacement  $t^{-1/2}X(t)$  it remains to prove that for all  $\delta > 0$  and  $\pi$ -almost all  $\omega \in \Omega$ ,

$$\lim_{t \rightarrow \infty} \mathbf{P}_\omega \left( |\Theta(X(t))| > \delta \sqrt{t} \right) = 0.$$

The key ingredients of this are the a priori quenched tightness of  $t^{-1/2}X(t)$  proved in the first main step, and a soft but nevertheless useful ergodic theorem for cocycles: Let  $\Omega \times \mathbb{Z}^d \ni x \mapsto \Psi(\omega, x) \in \mathbb{R}$  be a zero-mean  $\mathcal{L}^2$ -cocycle. Then

$$\lim_{N \rightarrow \infty} N^{-(d+1)} \sum_{|x| \leq N} |\Psi(x)| = 0, \quad \pi\text{-a.s.}$$

In 1-dimension this is a direct consequence of Birkhoff's ergodic theorem. For  $d > 1$ , however, the multidimensional *unconditional ergodic theorem* is invoked. See [Tóth \[n.d.\]](#) for the proof.

**1.5.3 Superdiffusive bounds when the  $\mathcal{H}_{-1}$ -condition fails.** If the  $\mathcal{H}_{-1}$ -condition (10)/(13) fails, or, equivalently, the conditions of part (ii) of Proposition 1 don't hold, then there is no a priori upper bound on  $t^{-1}\mathbf{E}(|X(t)|^2)$  and superdiffusive behaviour is expected. There is no general statement like this, but there are particular interesting cases studied. The fully worked out cases are, however, continuous-space diffusions on  $\mathbb{R}^d$  rather than random walks on  $\mathbb{Z}^d$ . Since  $d = 2$  is the most interesting we only mention the following two-dimensional example.

Let  $x \mapsto F(x)$  be a stationary Gaussian random vector field with covariances  $K_{ij}(x) := \mathbf{E}(F_i(0)F_j(x))$  as follows

$$(28) \quad K_{ij}(x) = (\partial_{ij}^2 - \delta_{i,j}(\partial_{11}^2 + \partial_{22}^2)) V * G(x), \quad \widehat{K}_{ij}(p) = \left( \delta_{i,j} - \frac{p_i p_j}{|p|^2} \right) \widehat{V}(p),$$

where  $G(x) = \log|x|$  is the two-dimensional (Laplacian) Green function and  $V : \mathbb{R}^2 \rightarrow \mathbb{R}_+$  is a  $C^\infty$  approximate identity with fast decay and positive Fourier transform,  $\widehat{V}(p) > 0$ . A good concrete choice could be  $V(x) = (2\pi\sigma^2)^{-d/2} \exp\{-|x|^2/(2\sigma^2)\}$ , with some  $\sigma \in (0, \infty)$ . In plain words:  $F$  is the rotation (curl) of the two-dimensional Gaussian free field, locally mollified by convolving with the convolution-square-root of  $V$ . As a rotation, the vector field  $F$  is divergence-free, cf. (11). Define the diffusion in this random drift field:  $t \mapsto X(t) \in \mathbb{R}^2$  as the unique strong solution of the SDE (12).

From (28) it appears that the  $\mathcal{H}_{-1}$ -condition fails marginally: the integral on the right hand side of (10) diverges logarithmically. In [Tóth and Valkó \[2012\]](#) superdiffusive bounds are proved for this diffusion in the rotation field of the two-dimensional Gaussian free field, which look formally very similar to (38) in Section 2.4 below. This extends earlier results (with power-law divergences) of [Komorowski and Olla \[2002\]](#) to the

marginal case of the two-dimensional Gaussian free field (with logarithmic divergences). The random walk on the six-vertex model (see the third example in [Section 1.4](#)) behaves phenomenologically similarly, but its superdiffusivity is not yet proved. Applying the same methods as in [Tóth and Valkó \[2012\]](#) we obtain, however, superdiffusive bounds for the variance of  $X(t)$  for the random walk on the randomly oriented Manhattan lattice (second example in [Section 1.4](#)) in dimensions  $d = 2$  (robust, power law) and  $d = 3$  (marginal, logarithmic), cf. [Ledger, Tóth, and Valkó \[n.d.\]](#).

**1.6 Historical notes.** The problems of scaling limit of diffusions in divergence-free random drift field (12)-(11) and that of the random walks in doubly stochastic random environment (2)-(1) are closely related. Although the physical phenomena modelled are very similar (tracer motion along the drift lines of incompressible turbulent flow), the technical details of various proofs are not always the same. In particular, PDE methods and techniques used for the diffusion problem (12)-(11) are not always easily implementable for the lattice problem (2)-(1). In the following list we give a summary of the main stations in the probability literature along the almost forty years history of the subject. The list is far from complete and contains only the probability results. See also the bibliographical notes of chapters 3 and 11 of [Komorowski, Landim, and Olla \[2012\]](#).

1979: Kozlov, respectively, Papanicolaou and Varadhan, independently and in parallel formulate the problem of scaling limits of diffusions in stationary random environment and prove the first CLT for the self-adjoint case under strong ellipticity condition, [Kozlov \[1979\]](#), [G. C. Papanicolaou and Varadhan \[1981\]](#).

1983: Osada proves quenched CLT for the diffusion (12) in divergence-free drift field (11), when the the stream tensor is bounded, [Osada \[1983\]](#).

1985: Kozlov formulates the problem of random walk in doubly stochastic random environment (1)-(2). An annealed CLT is stated for the case when the jump probabilities  $((p_k(\tau_x \omega))_{k \in \mathbb{E}})_{x \in \mathbb{Z}^d}$  are *finitely dependent*, [Kozlov \[1985a\]](#). Double stochasticity (1) and finite dependence of  $(p(\tau_x \omega))_{x \in \mathbb{Z}^d}$ , jointly are rather restrictive conditions. Natural examples are provided by a Bernoulli soup of bounded cycles.

1988: Oelschläger proves annealed invariance principle for the diffusion problem (12)-(11), under the optimal  $\mathcal{H}_{-1}$ -condition, and local regularity condition imposed on the drift field.

1996: Fannjiang and Papanicolaou consider the homogenisation for the parabolic problem corresponding to (12)-(11), under  $\mathcal{H}_{-1}$ -condition, [Fannjiang and G. Papanicolaou \[1996\]](#).

1997: Fannjiang and Komorowski prove quenched invariance principle for the diffusion (12)-(11), under the condition that  $h \in \mathcal{L}^p$ , with  $p > d$ , [Fannjiang and Komorowski \[1997\]](#)

2003: Komorowski and Olla prove annealed CLT for the random walk (2)-(1) when  $h \in \mathcal{L}^\infty$ , by applying Varadhan's strong sector condition, Komorowski and Olla [2003].

2008: Deuschel and Kösters prove quenched CLT for the random walk (2)-(1) when the jump probabilities admit a bounded cyclic representation, Deuschel and Kösters [2008]. This condition implies  $h \in \mathcal{L}^\infty$  and thus the strong sector condition.

2012: Komorowski, Landim and Olla publish the proof of the CLT for the random walk problem (2)-(1) in the case when  $h \in \mathcal{L}^p$ ,  $p > d$ , Komorowski, Landim, and Olla [2012].

2017: Annealed CLT for the doubly stochastic RWRE problem (2)-(1) under the optimal  $\mathcal{H}_{-1}$ -condition is proved in Kozma and Tóth [2017]. Quenched CLT under the additional integrability condition (20) is proved in Tóth [n.d.].

## 2 Self-repelling Brownian polymers

**2.1 Set-up and notation.** We consider a self-repelling random process  $t \mapsto X(t) \in \mathbb{R}^d$  which is pushed by the negative gradient of its own occupation time measure, towards regions less visited in the past. In this order, fix an approximate identity  $V : \mathbb{R}^d \rightarrow \mathbb{R}_+$ , which is infinitely differentiable, decays exponentially fast at infinity, and is of *positive type*:

$$(29) \quad \widehat{V}(p) := (2\pi)^{-d/2} \int_{\mathbb{R}^d} e^{ip \cdot x} V(x) dx \geq 0.$$

As an example, take  $V(x) = (2\pi\sigma^2)^{-d/2} e^{-|x|^2/(2\sigma^2)}$ ,  $\widehat{V}(p) = e^{-\sigma^2|p|^2/2}$ .

Let  $X(t)$  be the unique strong solution of the stochastic differential equation

$$(30) \quad dX(t) = dB(t) - \left( \int_0^t \text{grad } V(X(t) - X(u)) du \right) dt.$$

Denoting by  $\ell(t, \cdot)$  the *occupation time measure* of the process  $X$ ,

$$(31) \quad \ell(t, A) := |\{0 < s \leq t : X(s) \in A\}|,$$

where  $A \subset \mathbb{R}^d$  is any measurable set, the SDE (30) is written in the alternative form

$$(32) \quad dX(t) = dB(t) - \text{grad}(V * \ell(t, \cdot))(X(t)) dt,$$

which is more suggestive regarding the nature of the process  $t \mapsto X(t)$ : it is indeed driven by the negative gradient of an appropriate local regularization of its occupation time measure (local time). Following the terminology of the related probability literature we will refer to the process  $X(t)$  defined in (30)/(32) as the *self-repelling Brownian polymer*. The

main question is: What is the long time asymptotic behaviour of  $X$ ? How does the self-repulsion of the trajectory influence the long time scaling? The problem arose essentially in parallel, but unrelated, in the physics (random walk versions) and probability (diffusion versions) literature, cf. [Amit, Parisi, and Peliti \[1983\]](#), [Obukhov and Peliti \[1983\]](#), [Peliti and Pietronero \[1987\]](#), [Norris, Rogers, and Williams \[1987\]](#), [Durrett and Rogers \[1992\]](#), [Cranston and Mountford \[1996\]](#). Mathematically non-rigorous, nevertheless strong and compelling scaling arguments appearing in physics papers [Amit, Parisi, and Peliti \[1983\]](#), [Obukhov and Peliti \[1983\]](#), [Peliti and Pietronero \[1987\]](#) convincingly suggest the following dimension dependent asymptotic scaling behaviour, as  $t \rightarrow \infty$ :

$$(33) \quad X(t) \sim \begin{cases} t^{2/3} & \text{in } d = 1, \text{ with non-Gaussian scaling limit,} \\ t^{1/2}(\log t)^\gamma & \text{in } d = 2, \text{ with Gaussian scaling limit,} \\ t^{1/2} & \text{in } d \geq 3, \text{ with Gaussian scaling limit.} \end{cases}$$

In  $d = 2$ , the value of the exponent  $\gamma \in (0, 1/2]$  in the logarithmic correction is disputed. However, there is good scaling reason to expect  $\gamma = \frac{1}{4}$ . In the following sections we are going to present the mathematically rigorous results related to the conjectures of (33).

In one-dimension, for some particular nearest-neighbour lattice walk versions of the self-repelling motion, the conjecture (in the first line of) (33) is fully settled. In [Tóth \[1995\]](#) a limit theorem is proved for  $t^{-2/3}X(t)$ , with an intricate non-Gaussian limit distribution, believed to be universally valid for the 1-d cases. In [Tóth and Werner \[1998\]](#) the presumed scaling limit process  $t \mapsto \mathbb{X}(t)$  is constructed and fully analysed. In this note we are not going to cover those older results.

## 2.2 The environment process. Let

$$\Omega := \left\{ \omega \in C^\infty(\mathbb{R}^d \rightarrow \mathbb{R}^d) : \partial_k \omega_l = \partial_l \omega_k, \quad \|\omega\|_{k,m,r} < \infty, \right\},$$

where the seminorms are, for  $k \in \{1, \dots, d\}$ ,  $m \in \mathbb{N}^d$ ,  $r \in \mathbb{N}$ ,

$$\|\omega\|_{k,m,r} := \sup_{x \in \mathbb{R}^d} (1 + |x|)^{-1/r} \left| \partial_{m_1, \dots, m_d}^{m_1 + \dots + m_d} \omega_k(x) \right|.$$

In plain words:  $\Omega$  is the space of gradient vector fields on  $\mathbb{R}^d$ , with all partial derivatives increasing slower than any power of  $|x|$ .

It turns out that the process  $t \mapsto \eta(t, \cdot) \in \Omega$ ,

$$\eta(t, x) := \text{grad}(V * \ell(t, \cdot))(X(t) + x)$$

is a Markov process on the state space  $\Omega$  with almost surely continuous sample paths. We allow for an initial profile  $\eta(0, \cdot) \in \Omega$ . (This means an initial signed measure  $\ell(t, \cdot)$  in

(31).) The finite dimensional non-Markovian process  $t \mapsto X(t) \in \mathbb{R}^d$  is traded for the infinite dimensional Markov process  $t \mapsto \eta(t) \in \Omega$ .

Next we define a Gaussian probability measure on  $\Omega$ : the distribution of the gradient of the Gaussian free field on  $\mathbb{R}^d$ , locally regularised by convolving with the convolution-square-root of  $V$ . This is the point where positive definiteness (29) of the self-interaction potential  $V$  is essential. Let  $\pi$  be the Gaussian measure on  $\Omega$  with zero mean and covariances  $K_{kl}(y - x) := \int_{\Omega} \omega_k(x) \omega_l(y) d\pi(\omega)$ ,

$$(34) \quad K_{kl}(x) = -\partial_{kl} V * G(x), \quad \widehat{K}_{kl}(p) = \frac{p_k p_l}{|p|^2} \widehat{V}(p).$$

where  $G : \mathbb{R}^d \setminus \{0\} \mapsto \mathbb{R}$  is the (Laplacian) Green function.

The group of translations  $\tau_z : \Omega \rightarrow \Omega$ ,  $z \in \mathbb{R}^d$ , acts naturally as  $\tau_z \omega(x) := \omega(z + x)$ , and  $(\Omega, \mathcal{F}, \pi, \tau_z : z \in \mathbb{R}^d)$  is ergodic.

**Proposition 2.** (Source: Tarrès, Tóth, and Valkó [2012], Horváth, Tóth, and Vető [2012a]) *The Gaussian probability measure  $\pi$ , with zero mean and covariances (34) is stationary and ergodic for the Markov process  $\eta(t)$ .*

This fact is consequence of the harmony between the two mechanisms driving the process  $\eta$ : diffusion pushed by  $\eta(t, 0)$  and building up (gradient of) local time. Proposition 2 has two different proofs. In Tarrès, Tóth, and Valkó [2012] it is proved through careful Itô-calculus. In Horváth, Tóth, and Vető [2012a] a functional analytic proof is presented.

Tilted Gaussian measures with nonzero constant expectation and the same covariances as in (34) are also stationary and ergodic. We are not considering them because they result in ballistic behaviour (that is: nonzero overall speed) of the motion  $X(t)$ . We think (though, don't prove) that in  $d = 1, 2$  these are the only stationary and ergodic probability measures for the Markov process  $\eta(t)$ . In  $d \geq 3$ , however, other stationary distributions of totally different character do exist.

The forthcoming results are all valid in this stationary regime. That is: the initial  $\eta(0, \cdot)$  is sampled according to the distribution  $\pi$ . As in (23), the displacement of the random walker  $X(t)$  will be decomposed as sum of a martingale with stationary and ergodic increments and its compensator

$$(35) \quad X(t) = \left\{ X(t) - \int_0^t \varphi(\eta(s)) ds \right\} + \int_0^t \varphi(\eta(s)) ds =: M(t) + I(t),$$

where now  $\varphi : \Omega \rightarrow \mathbb{R}^d$ ,  $\varphi_l(\omega) = \omega_l(0)$ . The first term,  $M(t)$ , in (35) is a square integrable martingale with stationary and ergodic increments (on the probability space and with respect to the natural filtration of the Markov process  $t \mapsto \eta(t)$ ). So, that part is well understood from start: it is diffusive and the martingale central limit theorem applies to it.

In one and two dimensions *superdiffusive lower bounds* have been proved for the second term,  $I(t)$ , on the right hand side of (35), cf. [Tarrès, Tóth, and Valkó \[2012\]](#), [Tóth and Valkó \[2012\]](#) and [Theorem 4](#) below. On the other hand, in three and more dimensions an efficient martingale approximation à la Kipnis-Varadhan holds for the compensator term,  $I(t)$ , on the right hand side of (35), cf. [Horváth, Tóth, and Vető \[2012a\]](#) and [Theorem 5](#).

**2.3 The infinitesimal generator of the environment process.** All computations will be performed in the Hilbert space  $\mathcal{H} := \{f \in \mathcal{L}^2(\Omega, \pi) : \int_{\Omega} f d\pi = 0\}$ . Scalar product in the Hilbert space  $\mathcal{H}$  will be denoted  $\langle \cdot, \cdot \rangle$ . This is a *Gaussian Hilbert space* with its natural grading:

$$(36) \quad \mathcal{H} = \overline{\bigoplus_{n=1}^{\infty} \mathcal{H}_n},$$

where  $\mathcal{H}_n$  is the subspace spanned by the  $n$ -fold Wick products  $:\omega_{k_1}(x_1), \dots, \omega_{k_n}(x_n):, k_j \in \{1, \dots, d\}, x_j \in \mathbb{R}^d$ .

The shift operators  $U_z : \mathcal{H} \rightarrow \mathcal{H}$ ,  $U_z f(\omega) := f(\tau_z \omega)$ ,  $z \in \mathbb{R}^d$  form a unitary representation of  $\mathbb{R}^d$ . Denote by  $\nabla_k$ ,  $k \in \{1, \dots, d\}$  the infinitesimal generators:  $U_z = e^{\sum_{k=1}^d z_k \nabla_k}$  and  $\Delta := \sum_{k=1}^d \nabla_k^2 = -\sum_{k=1}^d \nabla_k^* \nabla_k$ . Note that the shift operators and all operators derived from them (e.g.  $\nabla_k$ ,  $\Delta$ ) preserve the grading (36).

We will also use the *creation and annihilation operators*  $a_l^* : \mathcal{H}_n \rightarrow \mathcal{H}_{n+1}$ ,  $a_l : \mathcal{H}_n \rightarrow \mathcal{H}_{n-1}$  defined on Wick monomials as follows and extended by linearity.

$$\begin{aligned} a_l^* : \omega_{k_1}(x_1), \dots, \omega_{k_n}(x_n) &:= \omega_l(0), \omega_{k_1}(x_1), \dots, \omega_{k_n}(x_n) : \\ a_l : \omega_{k_1}(x_1), \dots, \omega_{k_n}(x_n) &:= \sum_{m=1}^n K_{l k_m}(x_m) : \omega_{k_1}(x_1), \dots, \cancel{\omega_{k_m}(x_m)}, \dots, \omega_{k_n}(x_n) : \end{aligned}$$

As suggested by notation the operators  $a_l$  and  $a_l^*$  are adjoints of each other and restricted to any finite grade they are bounded:

$$\|a_l\|_{\mathcal{H}_n \rightarrow \mathcal{H}_{n-1}} = \left( \int_{\mathbb{R}^d} |p|^{-2} p_l^2 \widehat{V}(p) dp \right)^{1/2} \sqrt{n}.$$

The infinitesimal generator  $L$  of the semigroup of the Markov process  $\eta(t)$ , acting on  $\mathcal{H}$ ,  $P_t f(\omega) := \mathbf{E}(f(\eta(t)) \mid \eta(0) = \omega)$  is expressed in terms of the operators introduced above, as

$$(37) \quad L = -\frac{1}{2}\Delta + \sum_{l=1}^d a_l^* \nabla_l + \sum_{l=1}^d \nabla_l a_l = -S + A_+ + A_-.$$

The proof of this form of the infinitesimal generator relies – beside usual manipulations (integration by parts, etc.) – on *directional derivative* identities in the Gaussian Hilbert space  $\mathcal{H}$  (that is: elements of Malliavin calculus). For details see [Tarrès, Tóth, and Valkó \[2012\]](#), [Horváth, Tóth, and Vető \[2012a\]](#). The notation indicates that  $A_+ : \mathcal{H}_n \rightarrow \mathcal{H}_{n+1}$  while  $A_- : \mathcal{H}_n \rightarrow \mathcal{H}_{n+1}$ , and clearly,  $S = S^* \geq 0$ ,  $A_+ = -A_-^*$ .

## 2.4 Scaling limits.

### 2.4.1 Superdiffusive bounds in $d = 1$ and $d = 2$ . Let, for $\lambda > 0$

$$\widehat{E}(\lambda) := \int_0^\infty e^{-\lambda t} \mathbf{E}(|X(t)|^2) dt.$$

**Theorem 4.** (Source: [Tarrès, Tóth, and Valkó \[2012\]](#), [Tóth and Valkó \[2012\]](#)) In  $d \leq 2$ , the following bounds hold, with some constants  $0 < c < C < \infty$ , as  $\lambda \rightarrow 0$  :

$$\begin{aligned} \text{in } d = 1 : \quad & c\lambda^{-\frac{9}{4}} < \widehat{E}(\lambda) < C\lambda^{-\frac{5}{2}}, \\ \text{in } d = 2 : \quad & c\lambda^{-2} \log |\log \lambda| < \widehat{E}(\lambda) < C\lambda^{-2} \log |\lambda|. \end{aligned}$$

With some minimal regularity assumption on the asymptotic behaviour of  $t \mapsto \mathbf{E}(|X(t)|^2)$ , as  $t \rightarrow \infty$ , the bounds in [Theorem 4](#) imply bounds on their *Césaro means*,

$$\begin{aligned} \text{in } d = 1 : \quad & ct^{\frac{5}{4}} < \frac{1}{t} \int_0^t \mathbf{E}(|X(s)|^2) ds < Ct^{\frac{3}{2}}, \\ (38) \quad \text{in } d = 2 : \quad & ct \log \log t < \frac{1}{t} \int_0^t \mathbf{E}(|X(s)|^2) ds < Ct \log t. \end{aligned}$$

With some extra work the upper bounds can be improved to hold without Césaro averaging, see e.g. [Landim, Olla, and H. T. Yau \[1998\]](#). However, the lower bounds are the more interesting here. The bounds are consistent with but don't quite match the asymptotic behaviour conjectured in (33). Nevertheless, robust superdiffusivity in  $d = 1$  and marginal superdiffusivity in  $d = 2$  is at least established. Note also, that in  $d = 1$ , in particular cases (of self-repelling lattice walks) the scaling  $t^{2/3} X(t)$  has been rigorously established, cf. [Tóth \[n.d.\]](#), [Tóth and Werner \[1998\]](#), [Tóth and Vető \[2008\]](#).

The proof of [Theorem 4](#) follows the *resolvent method* of [Landim, Olla, and H. T. Yau \[1998\]](#), [Komorowski and Olla \[2002\]](#), [Landim, Quastel, Salmhofer, and H.-T. Yau \[2004\]](#), with new input in the variational computations for the 2-dimensional case.



Due to the martingale decomposition (35) and stationarity of the process  $\eta(s)$ , applying a straightforward Schwarz inequality we obtain

$$\int_0^t (t-s) \langle \varphi, P_t \varphi \rangle ds - 2\alpha^2 t \leq \mathbf{E} \left( |X(t)|^2 \right) \leq 4 \int_0^t (t-s) \langle \varphi, P_t \varphi \rangle ds + 2\alpha^2 t,$$

where  $\alpha^2 := t^{-1} \mathbf{E} (M(t)^2)$  is the variance rate of the first term,  $M(t)$ , in the decomposition (35), which is a square integrable martingale with stationary increments. (The value of  $\alpha^2$  is explicitly computable but does not matter.) Hence,

$$\lambda^{-2} \langle \varphi, R_\lambda \varphi \rangle - 2\alpha^2 \lambda^{-2} \leq \widehat{E}(\lambda) \leq 4\lambda^{-2} \langle \varphi, R_\lambda \varphi \rangle + 2\alpha^2 \lambda^{-2},$$

where  $R_\lambda$  is the resolvent of the semigroup  $P_t$ . Thus, lower and upper bounds on  $\widehat{E}(\lambda)$  reduce to lower and upper bounds on  $\langle \varphi, R_\lambda \varphi \rangle$ . The following variational formula, proved in Landim, Olla, and H. T. Yau [1998], is valid in the widest generality for any contraction semigroup  $P_t = e^{tL}$ , with infinitesimal generator  $L = -S + A$ :

$$(39) \quad \langle \varphi, R_\lambda \varphi \rangle = \sup_{\psi \in \mathcal{H}} \left\{ 2\langle \varphi, \psi \rangle - \langle \psi, (\lambda I + S)\psi \rangle - \langle A\psi, (\lambda I + S)^{-1} A\psi \rangle \right\}.$$

The upper bounds in Theorem 4 are obtained simply by dropping the third (negative!) term on the right hand side of (39). This is essentially for free. The lower bounds are obtained by bounding from below the variational expression on the right hand side of (39), in the subspace  $\mathcal{H}_1$ . This leads to a nontrivial variational problem in  $u : \mathbb{R}^d \mapsto \mathbb{R}^d$  ( $d = 1, 2$ ). In  $d = 2$  the solution is tricky. For details see Tóth and Valkó [2012].

#### 2.4.2 Diffusive limit in $d \geq 3$ .

**Theorem 5.** (Source: Horváth, Tóth, and Vető [2012a]) *In  $d \geq 3$ , the asymptotic covariance matrix*

$$(\sigma^2)_{ij} := \lim_{t \rightarrow \infty} t^{-1} \mathbf{E} (X_i(t) X_j(t))$$

*exists, it is bounded and non-degenerate. For any bounded and continuous function  $f : \mathbb{R}^d \rightarrow \mathbb{R}$ ,*

$$(40) \quad \lim_{T \rightarrow \infty} \int_{\Omega} \left| \mathbf{E}_{\omega} \left( f(T^{-1/2} X(T)) \right) - \int_{\mathbb{R}^d} \frac{e^{-|y|^2/2}}{(2\pi)^{d/2}} f(\sigma^{-1} y) dy \right| d\pi(\omega) = 0.$$

Theorem 5 is proved in Horváth, Tóth, and Vető [ibid.], and weak convergence in the sense of (40) of all finite dimensional marginal distributions of the diffusively scaled process  $t \mapsto T^{-\frac{1}{2}} X(Tt)$ , as  $T \rightarrow \infty$ , to those of a  $d$ -dimensional Brownian motion is established. The proof relies on the efficient martingale approximation à la Kipnis-Varadhan of

the integral term  $I(t)$  on the right hand side of (35). This is done by verifying the *graded sector condition* of Sethuraman, Varadhan, and H.-T. Yau [2000]. The graded structure of the Hilbert space  $\mathcal{H}$  and of the infinitesimal generator  $L$ , cf. (37). Technical details to be found in Horváth, Tóth, and Vető [2012a].

**Acknowledgments.** I thank Tomasz Komorowski and Stefano Olla for their help in clarifying some points related to the historical backgrounds.

## References

- Daniel J. Amit, G. Parisi, and L. Peliti (1983). “Asymptotic behavior of the “true” self-avoiding walk”. *Phys. Rev. B* (3) 27.3, pp. 1635–1645. MR: [690540](#) (cit. on p. 3070).
- M. Cranston and T. S. Mountford (1996). “The strong law of large numbers for a Brownian polymer”. *Ann. Probab.* 24.3, pp. 1300–1323. MR: [1411496](#) (cit. on p. 3070).
- Jean-Dominique Deuschel and Holger Kösters (2008). “The quenched invariance principle for random walks in random environments admitting a bounded cycle representation”. *Ann. Inst. Henri Poincaré Probab. Stat.* 44.3, pp. 574–591. MR: [2451058](#) (cit. on pp. 3063, 3066, 3069).
- R. T. Durrett and L. C. G. Rogers (1992). “Asymptotic behavior of Brownian polymers”. *Probab. Theory Related Fields* 92.3, pp. 337–349. MR: [1165516](#) (cit. on p. 3070).
- Albert Fannjiang and Tomasz Komorowski (1997). “A martingale approach to homogenization of unbounded random flows”. *Ann. Probab.* 25.4, pp. 1872–1894. MR: [1487440](#) (cit. on p. 3068).
- Albert Fannjiang and George Papanicolaou (1996). “Diffusion in turbulence”. *Probab. Theory Related Fields* 105.3, pp. 279–334. MR: [1425865](#) (cit. on p. 3068).
- Illés Horváth, Bálint Tóth, and Bálint Vető (2012a). “Diffusive limits for “true” (or myopic) self-avoiding random walks and self-repellent Brownian polymers in  $d \geq 3$ ”. *Probab. Theory Related Fields* 153.3-4, pp. 691–726. MR: [2948690](#) (cit. on pp. 3071–3075).
- (2012b). “Relaxed sector condition”. *Bull. Inst. Math. Acad. Sin. (N.S.)* 7.4, pp. 463–476. MR: [3077467](#) (cit. on pp. 3064, 3065).
- Tomasz Komorowski, Claudio Landim, and Stefano Olla (2012). *Fluctuations in Markov processes*. Vol. 345. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Time symmetry and martingale approximation. Springer, Heidelberg, pp. xviii+491. MR: [2952852](#) (cit. on pp. 3063, 3068, 3069).
- Tomasz Komorowski and Stefano Olla (2002). “On the superdiffusive behavior of passive tracer with a Gaussian drift”. *J. Statist. Phys.* 108.3-4, pp. 647–668. MR: [1914190](#) (cit. on pp. 3067, 3073).

- Tomasz Komorowski and Stefano Olla (2003). “A note on the central limit theorem for two-fold stochastic random walks in a random environment”. *Bull. Polish Acad. Sci. Math.* 51.2, pp. 217–232. MR: [1990811](#) (cit. on pp. [3063](#), [3069](#)).
- S. M. Kozlov (1979). “The averaging of random operators”. *Mat. Sb. (N.S.)* 109(151).2, pp. 188–202, 327. MR: [542557](#) (cit. on pp. [3066](#), [3068](#)).
- (1985a). “The averaging method and walks in inhomogeneous environments”. *Uspekhi Mat. Nauk* 40.2(242). Translated to English in [Kozlov \[1985b\]](#), pp. 61–120, 238. MR: [786087](#) (cit. on pp. [3058](#), [3063](#), [3066](#), [3068](#), [3076](#)).
  - (1985b). “The averaging method and walks in inhomogeneous environments”. *Russian Math. Surveys* 40. Translated from the original in Russian at [Kozlov \[1985a\]](#), pp. 73–145 (cit. on p. [3076](#)).
- Gady Kozma and Bálint Tóth (2017). “Central limit theorem for random walks in doubly stochastic random environment:  $\mathcal{H}_{-1}$  suffices”. *Ann. Probab.* 45.6B, pp. 4307–4347. MR: [3737912](#) (cit. on pp. [3062](#), [3064](#), [3065](#), [3069](#)).
- C. Landim, S. Olla, and H. T. Yau (1998). “Convection-diffusion equation with space-time ergodic random flow”. *Probab. Theory Related Fields* 112.2, pp. 203–220. MR: [1653837](#) (cit. on pp. [3073](#), [3074](#)).
- C. Landim, J. Quastel, M. Salmhofer, and H.-T. Yau (2004). “Superdiffusivity of asymmetric exclusion process in dimensions one and two”. *Comm. Math. Phys.* 244.3, pp. 455–481. MR: [2034485](#) (cit. on p. [3073](#)).
- S. Ledger, B. Tóth, and B. Valkó (n.d.). *Superdiffusive bounds for random walks on randomly oriented Manhattan lattices*. arXiv: [1802.01558](#) (cit. on p. [3068](#)).
- J. Nash (1958). “Continuity of solutions of parabolic and elliptic equations”. *Amer. J. Math.* 80, pp. 931–954. MR: [0100158](#) (cit. on p. [3066](#)).
- J. R. Norris, L. C. G. Rogers, and David Williams (1987). “Self-avoiding random walk: a Brownian motion model with local time drift”. *Probab. Theory Related Fields* 74.2, pp. 271–287. MR: [871255](#) (cit. on p. [3070](#)).
- S. P. Obukhov and L. Peliti (1983). “Renormalisation of the “true” self-avoiding walk”. *J. Phys. A* 16.5, pp. L147–L151. MR: [712594](#) (cit. on p. [3070](#)).
- Karl Oelschläger (1988). “Homogenization of a diffusion process in a divergence-free random field”. *Ann. Probab.* 16.3, pp. 1084–1126. MR: [942757](#).
- Hirofumi Osada (1983). “Homogenization of diffusion processes with random stationary coefficients”. In: *Probability theory and mathematical statistics (Tbilisi, 1982)*. Vol. 1021. Lecture Notes in Math. Springer, Berlin, pp. 507–517. MR: [736016](#) (cit. on pp. [3066](#), [3068](#)).
- G. C. Papanicolaou and S. R. S. Varadhan (1981). “Boundary value problems with rapidly oscillating random coefficients”. In: *Random fields, Vol. I, II (Esztergom, 1979)*. Vol. 27. Colloq. Math. Soc. János Bolyai. North-Holland, Amsterdam-New York, pp. 835–873. MR: [712714](#) (cit. on pp. [3066](#), [3068](#)).

- Luca Peliti and Luciano Pietronero (1987). “Random walks with memory”. *La Rivista del Nuovo Cimento* (1978-1999) 10.6, pp. 1–33 (cit. on p. 3070).
- Sunder Sethuraman, S. R. S. Varadhan, and Horng-Tzer Yau (2000). “Diffusive limit of a tagged particle in asymmetric simple exclusion processes”. *Comm. Pure Appl. Math.* 53.8, pp. 972–1006. MR: 1755948 (cit. on pp. 3065, 3075).
- Pierre Tarrès, Bálint Tóth, and Benedek Valkó (2012). “Diffusivity bounds for 1D Brownian polymers”. *Ann. Probab.* 40.2, pp. 695–713. MR: 2952088 (cit. on pp. 3071–3073).
- Bálint Tóth (n.d.). “Quenched Central Limit Theorem for Random Walks in Doubly Stochastic Random Environment”. To appear in *Ann. Probab.* in 2018. arXiv: 1704.06072 (cit. on pp. 3062, 3065–3067, 3069, 3073).
- (1986). “Persistent random walks in random environment”. *Probab. Theory Relat. Fields* 71.4, pp. 615–625. MR: 833271.
  - (1995). “The “true” self-avoiding walk with bond repulsion on  $\mathbb{Z}$ : limit theorems”. *Ann. Probab.* 23.4, pp. 1523–1556. MR: 1379158 (cit. on p. 3070).
- Bálint Tóth and Benedek Valkó (2012). “Superdiffusive bounds on self-repellent Brownian polymers and diffusion in the curl of the Gaussian free field in  $d = 2$ ”. *J. Stat. Phys.* 147.1, pp. 113–131. MR: 2922762 (cit. on pp. 3067, 3068, 3072–3074).
- Bálint Tóth and Bálint Vető (2008). “Continuous time ‘true’ self-avoiding random walk on  $\mathbb{Z}$ ”. *Electr. J. Probab.* 13 (cit. on p. 3073).
- Bálint Tóth and Wendelin Werner (1998). “The true self-repelling motion”. *Probab. Theory Related Fields* 111.3, pp. 375–452. MR: 1640799 (cit. on pp. 3070, 3073).
- S. R. S. Varadhan (1995). “Self-diffusion of a tagged particle in equilibrium for asymmetric mean zero random walk with simple exclusion”. *Ann. Inst. H. Poincaré Probab. Statist.* 31.1, pp. 273–285. MR: 1340041 (cit. on p. 3065).

Received 2017-12-09.

BÁLINT TÓTH  
SCHOOL OF MATHEMATICS  
UNIVERSITY OF BRISTOL  
BRISTOL, BS8 1TW  
UNITED KINGDOM

and

RÉNYI INSTITUTE  
BUDAPEST

[balint.toth@bristol.ac.uk](mailto:balint.toth@bristol.ac.uk)  
[toth.balint@renyi.mta.hu](mailto:toth.balint@renyi.mta.hu)

# THE METHOD OF HYPERGRAPH CONTAINERS

JÓZSEF BALOGH, ROBERT MORRIS AND WOJCIECH SAMOTIJ

## Abstract

In this survey we describe a recently-developed technique for bounding the number (and controlling the typical structure) of finite objects with forbidden substructures. This technique exploits a subtle clustering phenomenon exhibited by the independent sets of uniform hypergraphs whose edges are sufficiently evenly distributed; more precisely, it provides a relatively small family of ‘containers’ for the independent sets, each of which contains few edges. We attempt to convey to the reader a general high-level overview of the method, focusing on a small number of illustrative applications in areas such as extremal graph theory, Ramsey theory, additive combinatorics, and discrete geometry, and avoiding technical details as much as possible.

## 1 Introduction

Numerous well-studied problems in combinatorics concern families of discrete objects which avoid certain forbidden configurations, such as the family of  $H$ -free graphs<sup>1</sup> or the family of sets of integers containing no  $k$ -term arithmetic progression. The most classical questions about these families relate to the size and structure of the extremal examples; for example, Turán [1941] determined the unique  $K_r$ -free graph on  $n$  vertices with the most edges and Szemerédi [1975] proved that every set of integers of positive upper density contains arbitrarily long arithmetic progressions. In recent decades, partly motivated by applications to areas such as Ramsey theory and statistical physics, there has been increasing interest in problems relating to the typical structure of a (e.g., uniformly chosen) member of one of these families and to extremal questions in (sparse) random graphs and random sets of integers. Significant early developments in this direction include the seminal results obtained by Erdős, Kleitman, and Rothschild [1976], who proved that almost all

---

JB is partially supported by NSF Grant DMS-1500121 and by the Langan Scholar Fund (UIUC); RM is partially supported by CNPq (Proc. 303275/2013-8), by FAPERJ (Proc. 201.598/2014), and by ERC Starting Grant 680275 MALIG; WS is partially supported by the Israel Science Foundation grant 1147/14.

*MSC2010:* primary 05-02; secondary 05C30, 05C35, 05C65, 05D10, 05D40.

<sup>1</sup>A graph is  $H$ -free if it does not contain a subgraph isomorphic to  $H$ .

triangle-free graphs are bipartite, by [Kleitman and Winston \[1982\]](#), who proved that there are  $2^{\Theta(n^{3/2})}$   $C_4$ -free graphs on  $n$  vertices, and by [Frankl and Rödl \[1986\]](#), who proved that if  $p \gg 1/\sqrt{n}$ , then with high probability every 2-colouring of the edges of  $G(n, p)$  contains a monochromatic triangle.

An important recent development in this area was the discovery that, perhaps surprisingly, it is beneficial to consider such problems in the more abstract (and significantly more general) setting of independent sets in hypergraphs. This approach was taken with stunning success by [Conlon and Gowers \[2016\]](#), [Friedgut, Rödl, and Schacht \[2010\]](#), and [Schacht \[2016\]](#) in their breakthrough papers on extremal and Ramsey-type results in sparse random sets. To give just one example of the many important conjectures resolved by their work, let us consider the random variable

$$\text{ex}(G(n, p), H) = \max \{e(G) : H \not\subset G \subset G(n, p)\},$$

which was first studied (in the case  $H = K_3$ ) by [Frankl and Rödl \[1986\]](#). The following theorem was conjectured by [Haxell, Kohayakawa, and Łuczak \[1996, 1995\]](#) and proved (independently) by [Conlon and Gowers \[2016\]](#) and by [Schacht \[2016\]](#).

**Theorem 1.1.** *Let  $H$  be a graph with at least two edges and suppose that  $p \gg n^{-1/m_2(H)}$ , where  $m_2(H)$  is the so-called 2-density<sup>2</sup> of  $H$ . Then*

$$\text{ex}(G(n, p), H) = \left(1 - \frac{1}{\chi(H) - 1} + o(1)\right) p \binom{n}{2}$$

*asymptotically almost surely (a.a.s.), that is, with probability tending to 1 as  $n \rightarrow \infty$ .*

It is not hard to show that  $\text{ex}(G(n, p), H) = (1 + o(1))p \binom{n}{2}$  a.a.s. if  $n^{-2} \ll p \ll n^{-1/m_2(H)}$  and so the assumption on  $p$  in [Theorem 1.1](#) is optimal. We remark that in the case when  $H$  is a clique even more precise results are known, due to work of [DeMarco and Kahn \[2015, n.d.\]](#), who proved that if  $p \gg n^{-1/m_2(H)} (\log n)^{2/(r+1)(r-2)}$ , then with high probability the largest  $K_{r+1}$ -free subgraph of  $G(n, p)$  is  $r$ -partite, which is again essentially best possible. We refer the reader to an excellent recent survey of [Rödl and Schacht \[2013\]](#) for more details on extremal results in sparse random sets.

In this survey we will describe an alternative approach to the problem of understanding the family of independent sets in a hypergraph, whose development was inspired by the work in [Conlon and Gowers \[2016\]](#), [Friedgut, Rödl, and Schacht \[2010\]](#), and [Schacht \[2016\]](#) and also strongly influenced by that of [Kleitman and Winston \[1982\]](#) and [Sapozhenko \[2001, 2003, 2005\]](#). This technique, which was developed independently by [Balogh, Morris, and Samotij \[2015\]](#) and by [Saxton and Thomason \[2015\]](#), has turned out

<sup>2</sup>To be precise,  $m_2(H) = \max \left\{ \frac{e(F)-1}{v(F)-2} : F \subset H, v(F) \geq 3 \right\}$ .

to be surprisingly powerful and flexible. It allows one to prove enumerative, structural, and extremal results (such as [Theorem 1.1](#)) in a wide variety of settings. It is known as the *method of hypergraph containers*.

To understand the essence of the container method, it is perhaps useful to consider as an illustrative example the family  $\mathcal{F}_n(K_3)$  of triangle-free graphs on (a given set of)  $n$  vertices. Note that the number of such graphs is at least  $2^{\lfloor n^2/4 \rfloor}$ , since every bipartite graph is triangle-free.<sup>3</sup> However, it turns out that there exists a vastly smaller family  $\mathcal{G}_n$  of graphs on  $n$  vertices, of size  $n^{O(n^{3/2})}$ , that forms a set of *containers* for  $\mathcal{F}_n(K_3)$ , which means that for every  $H \in \mathcal{F}_n(K_3)$ , there exists a  $G \in \mathcal{G}_n$  such that  $H \subset G$ . A remarkable property of this family of containers is that each graph  $G \in \mathcal{G}_n$  is ‘almost triangle-free’ in the sense that it contains ‘few’ triangles. It is not difficult to use this family of containers, together with a suitable ‘supersaturation’ theorem, to prove [Theorem 1.1](#) in the case  $H = K_3$  or to show, using a suitable ‘stability’ theorem, that almost all triangle-free graphs are ‘almost bipartite’. We will discuss these two properties of the family of triangle-free graphs in much more detail in [Section 2](#).

In order to generalize this container theorem for triangle-free graphs, it is useful to first restate it in the language of hypergraphs. To do so, consider the 3-uniform hypergraph  $\mathcal{H}$  with vertex set  $V(\mathcal{H}) = E(K_n)$  and edge set

$$E(\mathcal{H}) = \{\{e_1, e_2, e_3\} \subset E(K_n) : e_1, e_2, e_3 \text{ form a triangle}\}.$$

We shall refer to  $\mathcal{H}$  as the ‘hypergraph that encodes triangles’ and emphasize that (somewhat confusingly) the vertices of this hypergraph are the edges of the complete graph  $K_n$ . Note that  $\mathcal{F}_n(K_3)$  is precisely the family  $\mathcal{I}(\mathcal{H})$  of independent sets of  $\mathcal{H}$ , so we may rephrase our container theorem for triangle-free graphs as follows:

There exists a relatively small family  $\mathcal{C}$  of subsets of  $V(\mathcal{H})$ , each containing only few edges of  $\mathcal{H}$ , such that every independent set  $I \in \mathcal{I}(\mathcal{H})$  is contained in some member of  $\mathcal{C}$ .

There is nothing special about the fine structure of the hypergraph encoding triangles that makes the above statement true. On the contrary, the method of containers allows one to prove that a similar phenomenon holds for a large class of  $k$ -uniform hypergraphs, for each  $k \in \mathbb{N}$ . In the case  $k = 3$ , a sufficient condition is the following assumption on the distribution of the edges of a 3-uniform hypergraph  $\mathcal{H}$  with average degree  $d$ : each vertex of  $\mathcal{H}$  has degree at most  $O(d)$  and each pair of vertices lies in at most  $O(\sqrt{d})$  edges of  $\mathcal{H}$ . For the hypergraph that encodes triangles, both conditions are easily satisfied, since each edge of  $K_n$  is contained in exactly  $n - 2$  triangles and each pair of edges is contained

<sup>3</sup>In particular, every subgraph of the complete bipartite graph with  $n$  vertices and  $\lfloor n^2/4 \rfloor$  edges is triangle-free.



in at most one triangle. The conclusion of the container lemma (see Sections 2 and 3) is that each independent set  $I$  in a 3-uniform hypergraph  $\mathcal{H}$  satisfying these conditions has a *fingerprint*  $S \subset I$  of size  $O(v(\mathcal{H})/\sqrt{d})$  that is associated with a set  $X(S)$  of size  $\Omega(v(\mathcal{H}))$  which is disjoint from  $I$ . The crucial point is that the set  $X(S)$  depends only on  $S$  (and not on  $I$ ) and therefore the number of sets  $X(S)$  is bounded from above by the number of subsets of the vertex set  $V(\mathcal{H})$  of size  $O(v(\mathcal{H})/\sqrt{d})$ . In particular, each independent set of  $\mathcal{H}$  is contained in one of at most  $v(\mathcal{H})^{O(v(\mathcal{H})/\sqrt{d})}$  sets of size at most  $(1 - \delta)v(\mathcal{H})$ , for some constant  $\delta > 0$ . By iterating this process, that is, by applying the container lemma repeatedly to the subhypergraphs induced by the containers obtained in earlier applications, one can easily prove the container theorem for triangle-free graphs stated (informally) above.

Although the hypergraph container lemma (see Section 3) was discovered only recently (see Balogh, Morris, and Samotij [2015] and Saxton and Thomason [2015]), several theorems of the same flavour (though often in very specific settings) appeared much earlier in the literature. The earliest container-type argument of which we are aware appeared (implicitly) over 35 years ago in the work of Kleitman and Winston [1980, 1982] on bounding the number of lattices, and of  $C_4$ -free graphs, which already contained some of the key ideas needed for the proof in the general setting; see Samotij [2015] for details. Nevertheless, it was not until almost 20 years later that Sapozhenko [2001, 2003, 2005] made a systematic study of containers for independent sets in graphs (and coined the name *containers*). Around the same time, Green and Ruzsa [2004] obtained (using Fourier analysis) a container theorem for sum-free subsets of  $\mathbb{Z}/p\mathbb{Z}$ .

More recently, Balogh and Samotij [2011a,b] generalized the method of Kleitman and Winston [1982] to count  $K_{s,t}$ -free graphs, using what could be considered to be the first container theorem for hypergraphs of uniformity larger than two. Finally, Alon, Balogh, Morris, and Samotij [2014a,b] proved a general container theorem for 3-uniform hypergraphs and used it to prove a sparse analogue of the Cameron–Erdős conjecture. Around the same time, Saxton and Thomason [2012] developed a simpler version of the method and applied it to the problem of bounding the list chromatic number of hypergraphs. In particular, the articles Alon, Balogh, Morris, and Samotij [2014b] and Saxton and Thomason [2012] can be seen as direct predecessors of Balogh, Morris, and Samotij [2015] and Saxton and Thomason [2015].

The rest of this survey is organised as follows. In Section 2, we warm up by stating a container lemma for 3-uniform hypergraphs, giving three simple applications to problems involving triangle-free graphs and a more advanced application to a problem in discrete geometry that was discovered recently by Balogh and Solymosi [n.d.]. Next, in Section 3, we state the main container lemma and provide some additional motivation and discussion of the statement and in Section 4 we describe an application to counting  $H$ -free graphs.

Finally, in Sections 5–8, we state and discuss a number of additional applications, including to multi-coloured structures (e.g., metric spaces), asymmetric structures (e.g., sparse members of a hereditary property), hypergraphs of unbounded uniformity (e.g., induced Ramsey numbers,  $\varepsilon$ -nets), number-theoretic structures (e.g., Sidon sets, sum-free sets, sets containing no  $k$ -term arithmetic progression), sharp thresholds in Ramsey theory, and probabilistic embedding in sparse graphs.

## 2 Basic applications of the method

In this section we will provide the reader with a gentle introduction to the container method, focusing again on the family of triangle-free graphs. In particular, we will state a version of the container lemma for 3-uniform hypergraphs and explain (without giving full details) how to deduce from it bounds on the largest size of a triangle-free subgraph of the random graph  $G(n, p)$ , statements about the typical structure of a (sparse) triangle-free graph, and how to prove that every  $r$ -colouring of the edges of  $G(n, p)$  contains a monochromatic triangle. To give a simple demonstration of the flexibility of the method, we will also describe a slightly more complicated application to a problem in discrete geometry.

In order to state the container lemma, we need a little notation. Given a hypergraph  $\mathcal{H}$ , let us write  $\Delta_\ell(\mathcal{H})$  for the maximum degree of a set of  $\ell$  vertices of  $\mathcal{H}$ , that is,

$$\Delta_\ell(\mathcal{H}) = \max \{d_{\mathcal{H}}(A) : A \subset V(\mathcal{H}), |A| = \ell\},$$

where  $d_{\mathcal{H}}(A) = |\{B \in E(\mathcal{H}) : A \subset B\}|$ , and  $\mathfrak{I}(\mathcal{H})$  for the collection of independent sets of  $\mathcal{H}$ .

**The hypergraph container lemma for 3-uniform hypergraphs.** *For every  $c > 0$ , there exists  $\delta > 0$  such that the following holds. Let  $\mathcal{H}$  be a 3-uniform hypergraph with average degree  $d \geq \delta^{-1}$  and suppose that*

$$\Delta_1(\mathcal{H}) \leq c \cdot d \quad \text{and} \quad \Delta_2(\mathcal{H}) \leq c \cdot \sqrt{d}.$$

*Then there exists a collection  $\mathcal{C}$  of subsets of  $V(\mathcal{H})$  with*

$$|\mathcal{C}| \leq \binom{v(\mathcal{H})}{v(\mathcal{H})/\sqrt{d}}$$

*such that*

- (a) *for every  $I \in \mathfrak{I}(\mathcal{H})$ , there exists  $C \in \mathcal{C}$  such that  $I \subset C$ ,*
- (b)  *$|C| \leq (1 - \delta)v(\mathcal{H})$  for every  $C \in \mathcal{C}$ .*

In order to help us understand the statement of this lemma, let us apply it to the hypergraph  $\mathcal{H}$  that encodes triangles in  $K_n$ , defined in the Introduction. Recall that this hypergraph satisfies

$$v(\mathcal{H}) = \binom{n}{2}, \quad \Delta_2(\mathcal{H}) = 1, \quad \text{and} \quad d_{\mathcal{H}}(v) = n - 2$$

for every  $v \in V(\mathcal{H})$ . We may therefore apply the container lemma to  $\mathcal{H}$ , with  $c = 1$ , to obtain a collection  $\mathcal{C}$  of  $n^{O(n^{3/2})}$  subsets of  $E(K_n)$  (that is, graphs on  $n$  vertices) with the following properties:

- (a) Every triangle-free graph is a subgraph of some  $C \in \mathcal{C}$ .
- (b) Each  $C \in \mathcal{C}$  has at most  $(1 - \delta)e(K_n)$  edges.

Now, if there exists a container  $C \in \mathcal{C}$  with at least  $\varepsilon n^3$  triangles, then take each such  $C$  and apply the container lemma to the subhypergraph  $\mathcal{H}[C]$  of  $\mathcal{H}$  induced by  $C$ , i.e., the hypergraph that encodes triangles in the graph  $C$ . Note that the average degree of  $\mathcal{H}[C]$  is at least  $6\varepsilon n$ , since each triangle in  $C$  corresponds to an edge of  $\mathcal{H}[C]$  and  $v(\mathcal{H}[C]) = |C| \leq e(K_n)$ . Since (trivially)  $\Delta_{\ell}(\mathcal{H}[C]) \leq \Delta_{\ell}(\mathcal{H})$ , it follows that we can apply the lemma with  $c = 1/\varepsilon$  and replace  $C$  by the collection of containers for  $\mathcal{H}[C]$  given by the lemma.

Let us iterate this process until we obtain a collection  $\mathcal{C}$  of containers, each of which has fewer than  $\varepsilon n^3$  triangles. How large is the final family  $\mathcal{C}$  that we obtain? Note that we apply the lemma only to hypergraphs with at most  $\binom{n}{2}$  vertices and average degree at least  $6\varepsilon n$  and therefore produce at most  $n^{O(n^{3/2})}$  new containers in each application, where the implicit constant depends only on  $\varepsilon$ . Moreover, each application of the lemma shrinks a container by a factor of  $1 - \delta$ , so after a bounded (depending on  $\varepsilon$ ) number of iterations every container will have fewer than  $\varepsilon n^3$  triangles (since  $\Delta_1(\mathcal{H}) < n$ , then every graph with at most  $\varepsilon n^2$  edges contains fewer than  $\varepsilon n^3$  triangles).

The above argument yields the following container theorem for triangle-free graphs.

**Theorem 2.1.** *For each  $\varepsilon > 0$ , there exists  $C > 0$  such that the following holds. For each  $n \in \mathbb{N}$ , there exists a collection  $\mathcal{G}$  of graphs on  $n$  vertices, with*

$$(1) \quad |\mathcal{G}| \leq n^{Cn^{3/2}},$$

*such that*

- (a) *each  $G \in \mathcal{G}$  contains fewer than  $\varepsilon n^3$  triangles;*
- (b) *each triangle-free graph on  $n$  vertices is contained in some  $G \in \mathcal{G}$ .*

In order to motivate the statement of [Theorem 2.1](#), we will next present three simple applications: bounding the largest size of a triangle-free subgraph of the random graph  $G(n, p)$ , determining the typical structure of a (sparse) triangle-free graph, and proving that  $G(n, p)$  cannot be partitioned into a bounded number of triangle-free graphs.

**2.1 Mantel's theorem in random graphs.** The oldest result in extremal graph theory, which states that every graph on  $n$  vertices with more than  $n^2/4$  edges contains a triangle, was proved by [Mantel \[1907\]](#). The corresponding problem in the random graph  $G(n, p)$  was first studied by [Frankl and Rödl \[1986\]](#), who proved the following theorem (cf. [Theorem 1.1](#)).

**Theorem 2.2.** *For every  $\alpha > 0$ , there exists  $C > 0$  such that the following holds. If  $p \geq C/\sqrt{n}$ , then a.a.s. every subgraph  $G \subset G(n, p)$  with*

$$e(G) \geq \left(\frac{1}{2} + \alpha\right)p \binom{n}{2}$$

*contains a triangle.*

As a simple first application of [Theorem 2.1](#), let us use it to prove [Theorem 2.2](#) under the marginally stronger assumption that  $p \gg \log n / \sqrt{n}$ . The proof exploits the following crucial property of  $n$ -vertex graphs with  $o(n^3)$  triangles: each such graph has at most  $(\frac{1}{2} + o(1))\binom{n}{2}$  edges. This statement is made rigorous in the following supersaturation lemma for triangles, which can be proved by simply applying Mantel's theorem to each induced subgraph of  $G$  with  $O(1)$  vertices.

**Lemma 2.3** (Supersaturation for triangles). *For every  $\delta > 0$ , there exists  $\varepsilon > 0$  such that the following holds. If  $G$  is a graph on  $n$  vertices with*

$$e(G) \geq \left(\frac{1}{4} + \delta\right)n^2,$$

*then  $G$  has at least  $\varepsilon n^3$  triangles.*

Applying [Lemma 2.3](#) with  $\delta = \alpha/2$  and [Theorem 2.1](#) with  $\varepsilon = \varepsilon(\delta)$  given by the lemma, we obtain a family of containers  $\mathcal{G}$  such that each  $G \in \mathcal{G}$  has fewer than  $\varepsilon n^3$  triangles and thus

$$e(G) \leq \left(\frac{1 + \alpha}{2}\right)\binom{n}{2}$$

for every  $G \in \mathcal{G}$ . Since every triangle-free graph is a subgraph of some container, if  $G(n, p)$  contains a triangle-free graph with  $m$  edges, then in particular  $e(G \cap G(n, p)) \geq$

$m$  for some  $G \in \mathcal{G}$ . Noting that  $e(G \cap G(n, p)) \sim \text{Bin}(e(G), p)$ , standard estimates on the tail of the binomial distribution yield

$$\mathbb{P}\left(e(G \cap G(n, p)) \geq \left(\frac{1}{2} + \alpha\right)p \binom{n}{2}\right) \leq e^{-\beta p n^2},$$

for some constant  $\beta = \beta(\alpha) > 0$ . Therefore, taking a union bound over all containers  $G \in \mathcal{G}$  and using the bound (1), we have (using the notation of Theorem 1.1)

$$(2) \quad \mathbb{P}\left(\text{ex}(G(n, p), K_3) \geq \left(\frac{1}{2} + \alpha\right)p \binom{n}{2}\right) \leq n^{O(n^{3/2})} \cdot e^{-\beta p n^2} \rightarrow 0$$

as  $n \rightarrow \infty$ , provided that  $p \gg \log n / \sqrt{n}$ . This gives the conclusion of Theorem 2.2 under a slightly stronger assumption on  $p$ . In Section 3, we show how to remove the extra factor of  $\log n$ .

We remark here that Theorem 2.2, as well as numerous results of this type that now exist in the literature, cannot be proved using standard first moment estimates. Indeed, since there are at least  $\binom{\lfloor n^2/4 \rfloor}{m}$  triangle-free graphs with  $n$  vertices and  $m$  edges, then letting  $X_m$  denote the number of such graphs that are contained in  $G(n, p)$ , we have

$$\mathbb{E}[X_m] \geq p^m \binom{\lfloor n^2/4 \rfloor}{m} = \left( \frac{(e/2 + o(1))p \binom{n}{2}}{m} \right)^m \gg 1$$

if  $m \leq (e/2 + o(1))p \binom{n}{2} = o(n^2)$ . This means that a first moment estimate would yield an upper bound on  $\text{ex}(G(n, p), K_3)$  that is worse than the trivial upper bound of  $(1 + o(1))p \binom{n}{2}$ .

**2.2 The typical structure of a sparse triangle-free graph.** A seminal theorem of Erdős, Kleitman, and Rothschild [1976] states that almost all triangle-free graphs are bipartite. Our second application of Theorem 2.1 is the following approximate version of this theorem for sparse graphs, first proved by Łuczak [2000]. Let us say that a graph  $G$  is  $t$ -close to bipartite if there exists a bipartite subgraph  $G' \subset G$  with  $e(G') \geq e(G) - t$ .

**Theorem 2.4.** *For every  $\alpha > 0$ , there exists  $C > 0$  such that the following holds. If  $m \geq Cn^{3/2}$ , then almost all triangle-free graphs with  $n$  vertices and  $m$  edges are  $\alpha m$ -close to bipartite.*

We will again (cf. the previous subsection) prove this theorem under the marginally stronger assumption that  $m \gg n^{3/2} \log n$ . To do so, we will need a finer characterisation of graphs with  $o(n^3)$  triangles that takes into account whether or not a graph is close

to bipartite. Proving such a result is less straightforward than [Lemma 2.3](#); for example, one natural proof combines the triangle removal lemma of [Ruzsa and Szemerédi \[1978\]](#) with the classical stability theorem of [Erdős \[1967\]](#) and [Simonovits \[1968\]](#). However, an extremely simple, beautiful, and elementary proof was given recently by [Füredi \[2015\]](#).

**Lemma 2.5** (Robust stability for triangles). *For every  $\delta > 0$ , there exists  $\varepsilon > 0$  such that the following holds. If  $G$  is a graph on  $n$  vertices with*

$$e(G) \geq \left(\frac{1}{2} - \varepsilon\right) \binom{n}{2},$$

*then either  $G$  is  $\delta n^2$ -close to bipartite or  $G$  contains at least  $\varepsilon n^3$  triangles.*

Applying [Lemma 2.5](#) with  $\delta = \delta(\alpha) > 0$  sufficiently small and [Theorem 2.1](#) with  $\varepsilon = \varepsilon(\delta)$  given by the lemma, we obtain a family of containers  $\mathcal{G}$  such that every  $G \in \mathcal{G}$  is either  $\delta n^2$ -close to bipartite or

$$(3) \quad e(G) \leq \left(\frac{1}{2} - \varepsilon\right) \binom{n}{2}.$$

Let us count those triangle-free graphs  $H$  with  $n$  vertices and  $m$  edges that are not  $\alpha m$ -close to bipartite; note that each such graph is a subgraph of some container  $G \in \mathcal{G}$ .

Suppose first that  $G$  satisfies (3); in this case we simply use the trivial bound

$$\binom{e(G)}{m} \leq \binom{\left(\frac{1}{2} - \varepsilon\right) \binom{n}{2}}{m} \leq (1 - \varepsilon)^m \binom{n^2/4}{m}$$

for the number of choices for  $H \subset G$ . On the other hand, if  $G$  is  $\delta n^2$ -close to bipartite, then there is some bipartite  $G' \subset G$  with  $e(G') \geq e(G) - \delta n^2$ . Since  $e(H \cap G') \leq (1 - \alpha)m$  by our assumption on  $H$ , we bound the number of choices for  $H$  by

$$\binom{e(G) - e(G')}{\alpha m} \binom{e(G)}{(1 - \alpha)m} \leq \binom{\delta n^2}{\alpha m} \binom{\binom{n}{2}}{(1 - \alpha)m} \leq 2^{-m} \binom{n^2/4}{m},$$

provided that  $\delta = \delta(\alpha)$  is sufficiently small. Summing over all choices of  $G \in \mathcal{G}$  and using (1), it follows that if  $m \gg n^{3/2} \log n$ , then there are at most

$$n^{O(n^{3/2})} \cdot (1 - \varepsilon)^m \binom{n^2/4}{m} \ll \binom{\lfloor n^2/4 \rfloor}{m}$$

triangle-free graphs  $H$  with  $n$  vertices and  $m$  edges that are not  $\alpha m$ -close to bipartite. However, there are clearly at least  $\binom{\lfloor n^2/4 \rfloor}{m}$  triangle-free graphs  $H$  with  $n$  vertices and  $m$

edges, since every bipartite graph is triangle-free, so the conclusion of [Theorem 2.4](#) holds when  $m \gg n^{3/2} \log n$ . We again postpone a discussion of how to remove the unwanted factor of  $\log n$  to [Section 3](#).

**2.3 Ramsey properties of sparse random graphs.** A folklore fact that is presented in each introduction to Ramsey theory states that every 2-colouring of the edges of  $K_6$  contains a monochromatic triangle. With the aim of constructing a small  $K_4$ -free graph that has the same property, [Frankl and Rödl \[1986\]](#) proved that if  $p \gg 1/\sqrt{n}$ , then a.a.s. every 2-colouring of the edges of  $G(n, p)$  contains a monochromatic triangle. Ramsey properties of random graphs were later thoroughly investigated by [Rödl and Ruciński \[1993, 1994, 1995\]](#); for example, they proved the following theorem.

**Theorem 2.6.** *For every  $r \in \mathbb{N}$ , there exists  $C > 0$  such that the following holds. If  $p \gg C/\sqrt{n}$ , then a.a.s. every  $r$ -colouring of the edges of  $G(n, p)$  contains a monochromatic triangle.*

We will present a simple proof of this theorem that was discovered recently by [Nenadov and Steger \[2016\]](#). For the sake of simplicity, we will again use the marginally stronger assumption that  $p \gg \log n/\sqrt{n}$ . The proof exploits the following property of  $n$ -vertex graphs with  $o(n^3)$  triangles: the union of any bounded number of such graphs cannot cover a  $(1 - o(1))$ -proportion of the edges of  $K_n$ . This property is a straightforward corollary of the following lemma, which can be proved by applying Ramsey's theorem to the colourings induced by all subsets of  $V(K_n)$  of size  $O(1)$ .

**Lemma 2.7.** *For every  $r \in \mathbb{N}$ , there exist  $n_0$  and  $\varepsilon > 0$  such that for all  $n \geq n_0$ , every  $(r+1)$ -colouring of the edges of  $K_n$  contains at least  $(r+1)\varepsilon n^3$  monochromatic triangles.*

Applying [Theorem 2.1](#) with  $\varepsilon = \varepsilon(r)$  given by the lemma, we obtain a family of containers  $\mathcal{G}$  such that every  $G \in \mathcal{G}$  has fewer than  $\varepsilon n^3$  triangles. If  $G(n, p)$  does not have the desired Ramsey property, then there are triangle-free graphs  $H_1, \dots, H_r$  such that  $H_1 \cup \dots \cup H_r = G(n, p)$ . It follows that  $G(n, p) \subset G_1 \cup \dots \cup G_r$ , where each  $G_i \in \mathcal{G}$  is a container for  $H_i$ . Since each  $G_i$  has fewer than  $\varepsilon n^3$  triangles, then [Lemma 2.7](#) implies that  $K_n \setminus (G_1 \cup \dots \cup G_r)$  contains at least  $\varepsilon n^3$  triangles.<sup>4</sup> Since each edge of  $K_n$  belongs to fewer than  $n$  triangles, we must have  $e(K_n \setminus (G_1 \cup \dots \cup G_r)) \geq \varepsilon n^2$ . Consequently, for each fixed  $G_1, \dots, G_r \in \mathcal{G}$ ,

$$\mathbb{P}(G(n, p) \subset G_1 \cup \dots \cup G_r) = (1 - p)^{e(K_n \setminus (G_1 \cup \dots \cup G_r))} \leq (1 - p)^{\varepsilon n^2} \leq e^{-\varepsilon p n^2}.$$

<sup>4</sup>To see this, consider an  $(r+1)$ -colouring of the edges of  $K_n$  that assigns to each edge  $e \in G_1 \cup \dots \cup G_r$  some colour  $i$  such that  $e \in G_i$  and assigns colour  $r+1$  to all edges of  $K_n \setminus (G_1 \cup \dots \cup G_r)$ .

Taking a union bound over all  $r$ -tuples of containers, we conclude that

$$\mathbb{P}(G(n, p) \text{ admits a 'bad' } r\text{-colouring}) \leq n^{O(n^{3/2})} \cdot e^{-\varepsilon p n^2} \rightarrow 0$$

as  $n \rightarrow \infty$ , provided that  $p \gg \log n / \sqrt{n}$ . As before, the unwanted factor of  $\log n$  can be removed with a somewhat more careful analysis that we shall discuss in [Section 3](#).

**2.4 An application in discrete geometry.** In order to give some idea of the flexibility of the container method, we will next present a somewhat more elaborate application of the container lemma for 3-uniform hypergraphs, which was discovered recently by [Balogh and Solymosi \[n.d.\]](#), to the following question posed by [Erdős \[1988\]](#). Given  $n$  points in the Euclidean plane  $\mathbb{R}^2$ , with at most three on any line, how large a subset are we guaranteed to find in general position (i.e., with at most two on any line)? [Füredi \[1991\]](#) proved that one can always find such a subset of size  $\Omega(\sqrt{n \log n})$  and gave a construction (which relied on the density Hales–Jewett theorem of [Furstenberg and Katznelson \[1991\]](#)) in which the largest such set has size  $o(n)$ . Using the method of hypergraph containers, Balogh and Solymosi obtained the following stronger upper bound.

**Theorem 2.8.** *There exists a set  $S \subset \mathbb{R}^2$  of size  $n$ , containing no four points on a line, such that every subset of  $S$  of size  $n^{5/6+o(1)}$  contains three points on a line.*

The key idea of Balogh and Solymosi was to first construct a set  $P$  of points that contains ‘few’ collinear quadruples, but such that every ‘large’ subset of  $P$  contains ‘many’ collinear triples. Then a random subset  $R$  of  $P$  of a carefully chosen density will typically contain only  $o(|R|)$  collinear quadruples, since the density is not too large and there are few collinear quadruples. On the other hand, every subset of  $R$  with more than  $|R|^{5/6+o(1)}$  elements will still contain a collinear triple; this follows from the hypergraph container lemma, as large sets contain many collinear triples and the density is not too small. Removing one element from each collinear quadruple in  $R$  gives the desired set  $A$ .

Formally, we first define the following 3-uniform hypergraph  $\mathcal{H}$ . We let  $V(\mathcal{H}) = [m]^3$  (so the vertices are lattice points in  $\mathbb{R}^3$ ) and let  $E(\mathcal{H})$  be the collection of triples of points that lie on a common line. Thus, a subset of  $V(\mathcal{H})$  is in general position if and only if it is an independent set of  $\mathcal{H}$ . The following lemma was proved in [Balogh and Solymosi \[n.d.\]](#).

**Lemma 2.9** (Supersaturation for collinear triples). *For every  $0 < \gamma < 1/2$  and every  $S \subset [m]^3$  of size at least  $m^{3-\gamma}$ , there exist at least  $m^{6-4\gamma-o(1)}$  collinear triples of points in  $S$ .*

We now repeatedly apply the hypergraph container lemma for 3-uniform hypergraphs to subhypergraphs of  $\mathcal{H}$ . Suppose that  $s \geq m^{8/3+o(1)}$  and let  $S \subset [m]^3$  be an arbitrary



$s$ -element set. [Lemma 2.9](#) gives

$$e(\mathcal{H}[S]) \geq s^4/m^{6+o(1)} \quad \text{and} \quad \Delta_2(\mathcal{H}[S]) \leq \Delta_2(\mathcal{H}) \leq m.$$

Moreover, it is not difficult to deduce that there exists a subhypergraph  $\mathcal{H}' \subset \mathcal{H}[S]$  with

$$v(\mathcal{H}') = |S| = s, \quad e(\mathcal{H}') = s^4/m^{6+o(1)}, \quad \text{and} \quad \Delta_1(\mathcal{H}') = O(e(\mathcal{H}')/v(\mathcal{H}')).$$

We may therefore apply the container lemma for 3-uniform hypergraphs to  $\mathcal{H}'$  to obtain a collection  $\mathcal{C}$  of at most  $\exp(m^{3+o(1)}/\sqrt{s})$  subsets of  $S$  with the following properties:

- (a) Every set of points of  $S$  in general position is contained in some  $C \in \mathcal{C}$ ,
- (b) Each  $C \in \mathcal{C}$  has size at most  $(1 - \delta)|S|$ .

Starting with  $S = [m]^3$  and iterating this process for  $O(\log m)$  steps, we obtain the following container theorem for sets of points in general position.

**Theorem 2.10.** *For each  $m \in \mathbb{N}$ , there exists a collection  $\mathcal{C}$  of subsets of  $[m]^3$  with*

$$(4) \quad |\mathcal{C}| \leq \exp(m^{5/3+o(1)})$$

such that

- (a)  $|C| \leq m^{8/3+o(1)}$  for each  $C \in \mathcal{C}$ ;
- (b) each set of points of  $[m]^3$  in general position is contained in some  $C \in \mathcal{C}$ .

Now, let  $p = m^{-1+o(1)}$  and consider a  $p$ -random subset  $R \subset [m]^3$ , that is, each element of  $[m]^3$  is included in  $R$  independently at random with probability  $p$ . Since  $[m]^3$  contains  $m^{6+o(1)}$  sets of four collinear points<sup>5</sup>, it follows that, with high probability,  $|R| = pm^{3+o(1)} = m^{2+o(1)}$  and  $R$  contains  $p^4 m^{6+o(1)} = o(|R|)$  collinear 4-tuples. Moreover, since  $|C| \leq m^{8/3+o(1)}$  for each  $C \in \mathcal{C}$ , it follows from (4) and standard estimates on the tail of the binomial distribution that with high probability we have  $|R \cap C| \leq m^{5/3+o(1)}$  for every  $C \in \mathcal{C}$ . In particular, removing one element from each collinear 4-tuple in  $R$  yields a set  $A \subset [m]^3$  of size  $m^{2+o(1)}$  with no collinear 4-tuple and containing no set of points in general position of size larger than  $m^{5/3+o(1)}$ . Finally, project the points of  $A$  to the plane in such a way that collinear triples remain collinear, and no new collinear triple is created. In this way, we obtain a set of  $n = m^{2+o(1)}$  points in the plane, no four of them on a line, such that no set of size greater than  $n^{5/6+o(1)} = m^{5/3+o(1)}$  is in general position, as required.

<sup>5</sup>This is because there are  $O(m^6/t^4)$  lines in  $\mathbb{R}^3$  that contain more than  $t$  points of  $[m]^3$ .

### 3 The key container lemma

In this section, we state a container lemma for hypergraphs of arbitrary uniformity. The version of the lemma stated below, which comes from [Morris, Samotij, and Saxton \[n.d.\]](#), differs from the statement originally proved by the authors of this survey [Balogh, Morris, and Samotij \[2015, Proposition 3.1\]](#) only in that the dependencies between the various constants have been made more explicit here; a careful analysis of the proof of [Balogh, Morris, and Samotij \[ibid., Proposition 3.1\]](#) will yield this slightly sharper statement.<sup>6</sup> Let us recall that for a hypergraph  $\mathcal{H}$  and an integer  $\ell$ , we write  $\Delta_\ell(\mathcal{H})$  for the maximum degree of a set of  $\ell$  vertices of  $\mathcal{H}$ , that is,

$$\Delta_\ell(\mathcal{H}) = \max \{d_{\mathcal{H}}(A) : A \subset V(\mathcal{H}), |A| = \ell\},$$

where  $d_{\mathcal{H}}(A) = |\{B \in E(\mathcal{H}) : A \subset B\}|$ , and  $\mathfrak{I}(\mathcal{H})$  for the collection of independent sets of  $\mathcal{H}$ . The lemma states, roughly speaking, that each independent set  $I$  in a uniform hypergraph  $\mathcal{H}$  can be assigned a *fingerprint*  $S \subset I$  in such a way that all sets with the same fingerprint are contained in a single set  $C = f(S)$ , called a *container*, whose size is bounded away from  $v(\mathcal{H})$ . More importantly, the sizes of these fingerprints (and hence also the number of containers) can be bounded from above (in an optimal way!) by basic parameters of  $\mathcal{H}$ .

**The hypergraph container lemma.** *Let  $k \in \mathbb{N}$  and set  $\delta = 2^{-k(k+1)}$ . Let  $\mathcal{H}$  be a  $k$ -uniform hypergraph and suppose that*

$$(5) \quad \Delta_\ell(\mathcal{H}) \leq \left( \frac{b}{v(\mathcal{H})} \right)^{\ell-1} \frac{e(\mathcal{H})}{r}$$

*for some  $b, r \in \mathbb{N}$  and every  $\ell \in \{1, \dots, k\}$ . Then there exists a collection  $\mathcal{C}$  of subsets of  $V(\mathcal{H})$  and a function  $f : \mathcal{P}(V(\mathcal{H})) \rightarrow \mathcal{C}$  such that:*

- (a) *for every  $I \in \mathfrak{I}(\mathcal{H})$ , there exists  $S \subset I$  with  $|S| \leq (k-1)b$  and  $I \subset f(S)$ ;*
- (b)  *$|C| \leq v(\mathcal{H}) - \delta r$  for every  $C \in \mathcal{C}$ .*

The original statement of the container lemma [Balogh, Morris, and Samotij \[ibid., Proposition 3.1\]](#) had  $r = v(\mathcal{H})/c$  for some constant  $c$ , since this choice of parameters is required in most standard applications. In particular, the simple container lemma for 3-uniform hypergraphs presented in [Section 2](#) is easily derived from the above statement by letting  $b = v(\mathcal{H})/(2\sqrt{d})$  and  $r = v(\mathcal{H})/(6c)$ , where  $d = 3e(\mathcal{H})/v(\mathcal{H})$  is the average

---

<sup>6</sup>A complete proof of the version of the container lemma stated here can be found in [Morris, Samotij, and Saxton \[n.d.\]](#).

degree of  $\mathcal{H}$ . There are, however, arguments that benefit from setting  $r = o(v(\mathcal{H}))$ ; we present one of them in [Section 5](#).

Even though the property  $|C| \leq v(\mathcal{H}) - \delta r$  that is guaranteed for all containers  $C \in \mathbb{C}$  seems rather weak at first sight, it can be easily strengthened with repeated applications of the lemma. In particular, if for some hypergraph  $\mathcal{H}$ , condition (5) holds (for all  $\ell$ ) with some  $b = o(v(\mathcal{H}))$  and  $r = \Omega(v(\mathcal{H}))$ , then recursively applying the lemma to subhypergraphs of  $\mathcal{H}$  induced by all the containers  $C$  for which  $e(\mathcal{H}[C]) \geq \varepsilon e(\mathcal{H})$  eventually produces a collection  $\mathbb{C}$  of containers indexed by sets of size  $O(b)$  such that  $e(\mathcal{H}[C]) < \varepsilon e(\mathcal{H})$  for every  $C \in \mathbb{C}$ . This is precisely how (in [Section 2](#)) we derived [Theorem 2.1](#) from the container lemma for 3-uniform hypergraphs. For a formal argument showing how such a family of ‘tight’ containers may be constructed, we refer the reader to [Balogh, Morris, and Samotij \[2015\]](#).

One may thus informally say that the hypergraph container lemma provides a covering of the family of all independent sets of a uniform hypergraph with ‘few’ sets that are ‘almost independent’. In many natural settings, these almost independent sets closely resemble truly independent sets. In some cases, this is a straightforward consequence of corresponding removal lemmas. A more fundamental reason is that many sequences of hypergraphs  $\mathcal{H}_n$  of interest possess the following self-similarity property: For all (or many) pairs  $m$  and  $n$  with  $m < n$ , the hypergraph  $\mathcal{H}_n$  admits a very uniform covering by copies of  $\mathcal{H}_m$ . For example, this is the case when  $\mathcal{H}_n$  is the hypergraph encoding triangles in  $K_n$ , simply because every  $m$ -element set of vertices of  $K_n$  induces  $K_m$ . Such self-similarity enables one to use elementary averaging arguments to characterise almost independent sets; for example, the standard proof of [Lemma 2.3](#) uses such an argument.

The fact that the fingerprint  $S$  of each independent set  $I \in \mathcal{I}(\mathcal{H})$  is a subset of  $I$  is not merely a by-product of the proof of the hypergraph container lemma. On the contrary, it is an important property of the family of containers that can be often exploited to make union bound arguments tighter. This is because each  $I \in \mathcal{I}(\mathcal{H})$  is sandwiched between  $S$  and  $f(S)$  and consequently when enumerating independent sets one may use a union bound over all fingerprints  $S$  and enumerate only over the sets  $I \setminus S$  (which are contained in  $f(S)$ ). In particular, such finer arguments can be used to remove the superfluous logarithmic factor from the assumptions of the proofs outlined in [Section 2](#). For example, in the proof of [Theorem 2.2](#) presented in [Section 2.1](#), the fingerprints of triangle-free subgraphs of  $K_n$  form a family  $\mathcal{S}$  of  $n$ -vertex graphs, each with at most  $C_\varepsilon n^{3/2}$  edges. Setting  $m = (\frac{1}{2} + \alpha)p\binom{n}{2}$ , this allows us to replace (2) with the following estimate:

$$(6) \quad \mathbb{P}\left(\text{ex}(G(n, p), K_3) \geq m\right) \leq \sum_{S \in \mathcal{S}} \mathbb{P}\left(S \subset G(n, p) \text{ and } e((f(S) \setminus S) \cap G(n, p)) \geq m - |S|\right).$$

Since the two events in the right-hand side of (6) concern the intersections of  $G(n, p)$  with two disjoint sets of edges of  $K_n$ , they are independent. If  $p \gg n^{-1/2}$ , then  $|S| \ll p \binom{n}{2}$  and consequently, recalling that  $e(f(S)) \leq \binom{1+\alpha}{2} \binom{n}{2}$ , we may bound the right-hand side of (6) from above by

$$\sum_{S \in \mathcal{S}} p^{|S|} e^{-\beta p n^2} \leq \sum_{s \leq C_\varepsilon n^{3/2}} \binom{\binom{n}{2}}{s} \cdot p^s e^{-\beta p n^2} \leq \sum_{s \leq C_\varepsilon n^{3/2}} \left( \frac{\binom{n}{2} p}{s} \right)^s e^{-\beta p n^2} \leq e^{-\beta p n^2/2}$$

for some  $\beta = \beta(\alpha) > 0$ .

Finally, what is the intuition behind condition (5)? A natural way to define  $f(S)$  for a given (independent) set  $S$  is to let  $f(S) = V(\mathcal{H}) \setminus X(S)$ , where  $X(S)$  comprises all vertices  $v$  such that  $A \subset S \cup \{v\}$  for some  $A \in E(\mathcal{H})$ . Indeed, every independent set  $I$  that contains  $S$  must be disjoint from  $X(S)$ . (In reality, the definition of  $X(S)$  is – and has to be – more complicated than this, and some vertices are placed in  $X(S)$  simply because they do not belong to  $S$ .) Suppose, for the sake of argument, that  $S$  is a random set of  $b$  vertices of  $\mathcal{H}$ . Letting  $\tau = b/v(\mathcal{H})$ , we have

$$(7) \quad \mathbb{E}[|X(S)|] \leq \sum_{A \in E(\mathcal{H})} \mathbb{P}(|A \cap S| = k-1) \leq k \cdot \tau^{k-1} \cdot e(\mathcal{H}).$$

Since we want  $X(S)$  to have at least  $\delta r$  elements for every fingerprint  $S$ , it seems reasonable to require that

$$\Delta_k(\mathcal{H}) = 1 \leq \frac{k}{\delta} \cdot \tau^{k-1} \cdot \frac{e(\mathcal{H})}{r},$$

which is, up to a constant factor, condition (5) with  $\ell = k$ . For some hypergraphs  $\mathcal{H}$  however, the first inequality in (7) can be very wasteful, since some  $v \in X(S)$  may have many  $A \in E(\mathcal{H})$  such that  $A \subset S \cup \{v\}$ . This can happen if for some  $\ell \in \{1, \dots, k-1\}$ , there is an  $\ell$ -uniform hypergraph  $\mathcal{G}$  such that each edge of  $\mathcal{H}$  contains an edge of  $\mathcal{G}$ ; note that  $e(\mathcal{G})$  can be as small as  $e(\mathcal{H})/\Delta_\ell(\mathcal{H})$ . Our assumption implies that  $\mathfrak{L}(\mathcal{G}) \subset \mathfrak{L}(\mathcal{H})$  and thus, letting  $Y(S)$  be the set of all vertices  $w$  such that  $B \subset S \cup \{w\}$  for some  $B \in E(\mathcal{G})$ , we have  $X(S) \subset Y(S)$ . In particular, we want  $Y(S)$  to have at least  $\delta r$  elements for every fingerprint  $S$  of an independent set  $I \in \mathfrak{L}(\mathcal{G})$ . Repeating (7) with  $X$  replaced by  $Y$ ,  $\mathcal{H}$  replaced by  $\mathcal{G}$ , and  $k$  replaced by  $\ell$ , we arrive at the inequality

$$\delta r \leq \ell \cdot \tau^{\ell-1} \cdot e(\mathcal{G}) = \ell \cdot \tau^{\ell-1} \cdot \frac{e(\mathcal{H})}{\Delta_\ell(\mathcal{H})},$$

which is, up to a constant factor, condition (5).

One may further develop the above argument to show that condition (5) is asymptotically optimal, at least when  $r = \Omega(v(\mathcal{H}))$ . Roughly speaking, one can construct  $k$ -uniform

hypergraphs that have  $\binom{(1-o(1))v(\mathcal{H})}{m}$  independent  $m$ -sets for every  $m = o(b)$ , where  $b$  is minimal so that condition (5) holds, whereas the existence of containers of size at most  $(1-\delta)v(\mathcal{H})$  indexed by fingerprints of size  $o(b)$  would imply that the number of such sets is at most  $\binom{(1-\varepsilon)v(\mathcal{H})}{m}$  for some constant  $\varepsilon > 0$ .

## 4 Counting $H$ -free graphs

How many graphs are there on  $n$  vertices that do not contain a copy of  $H$ ? An obvious lower bound is  $2^{\text{ex}(n,H)}$ , since each subgraph of an  $H$ -free graph is also  $H$ -free. For non-bipartite graphs, this is not far from the truth. Writing  $\mathfrak{F}_n(H)$  for the family of  $H$ -free graphs on  $n$  vertices, if  $\chi(H) \geq 3$ , then

$$(8) \quad |\mathfrak{F}_n(H)| = 2^{(1+o(1))\text{ex}(n,H)}$$

as  $n \rightarrow \infty$ , as was first shown by Erdős, Kleitman, and Rothschild [1976] (when  $H$  is a complete graph) and then by Erdős, Frankl, and Rödl [1986]. For bipartite graphs, on the other hand, the problem is much more difficult. In particular, the following conjecture (first stated in print in Kleitman and Winston [1982]), which played a major role in the development of the container method, remains open.

**Conjecture 4.1.** *For every bipartite graph  $H$  that contains a cycle, there exists  $C > 0$  such that*

$$|\mathfrak{F}_n(H)| \leq 2^{C\text{ex}(n,H)}$$

for every  $n \in \mathbb{N}$ .

The first significant progress on Conjecture 4.1 was made by Kleitman and Winston [ibid.]. Their proof of the case  $H = C_4$  of the conjecture introduced (implicitly) the container method for graphs. Nevertheless, it took almost thirty years<sup>7</sup> until their theorem was generalized to the case  $H = K_{s,t}$ , by Balogh and Samotij [2011a,b], and then (a few years later) to the case  $H = C_{2k}$ , by Morris and Saxton [2016]. More precisely, it was proved in Balogh and Samotij [2011b] and Morris and Saxton [2016] that

$$|\mathfrak{F}_n(K_{s,t})| = 2^{O(n^{2-1/s})} \quad \text{and} \quad |\mathfrak{F}_n(C_{2k})| = 2^{O(n^{1+1/k})}$$

for every  $2 \leq s \leq t$  and every  $k \geq 2$ , which implies Conjecture 4.1 when  $t > (s-1)!$  and  $k \in \{2, 3, 5\}$ , since in these cases it is known that  $\text{ex}(n, K_{s,t}) = \Theta(n^{2-1/s})$  and  $\text{ex}(n, C_{2k}) = \Theta(n^{1+1/k})$ .

Very recently, Ferber, McKinley, and Samotij [n.d.], inspired by a similar result of Balogh, Liu, and Sharifzadeh [2017] on sets of integers with no  $k$ -term arithmetic progression, found a very simple proof of the following much more general theorem.

<sup>7</sup>An unpublished manuscript of Kleitman and Wilson from 1996 proves that  $|\mathfrak{F}_n(C_6)| = 2^{O(\text{ex}(n, C_6))}$ .

**Theorem 4.2.** *Suppose that  $H$  contains a cycle. If  $\text{ex}(n, H) = O(n^\alpha)$  for some constant  $\alpha$ , then*

$$|\mathcal{F}_n(H)| = 2^{O(n^\alpha)}.$$

Note that [Theorem 4.2](#) resolves [Conjecture 4.1](#) for every  $H$  such that  $\text{ex}(n, H) = \Theta(n^\alpha)$  for some constant  $\alpha$ . Ferber, McKinley, and Samotij showed moreover that the weaker assumption that  $\text{ex}(n, H) \gg n^{2-1/m_2(H)+\varepsilon}$  for some  $\varepsilon > 0$  already implies that the assertion of [Conjecture 4.1](#) holds for infinitely many  $n$ ; we refer the interested reader to their paper for the details. Let us also note here that, while it is natural to suspect that in fact the stronger bound (8) holds for all graphs  $H$  that contain a cycle, this is false for  $H = C_6$ , as was shown by [Morris and Saxton \[2016\]](#). However, it may still hold for  $H = C_4$  and it would be very interesting to determine whether or not this is indeed the case.

The proof of [Theorem 4.2](#) for general  $H$  is somewhat technical, so let us instead sketch the proof in the case  $H = C_4$ . In this case, the proof combines the hypergraph container lemma stated in the previous section with the following supersaturation lemma.

**Lemma 4.3.** *There exist constants  $\beta > 0$  and  $k_0 \in \mathbb{N}$  such that the following holds for every  $k \geq k_0$  and every  $n \in \mathbb{N}$ . Given a graph  $G$  with  $n$  vertices and  $k \cdot \text{ex}(n, C_4)$  edges, there exists a collection  $\mathcal{H}$  of at least  $\beta k^5 \cdot \text{ex}(n, C_4)$  copies of  $C_4$  in  $G$  that satisfies:*

- (a) *Each edge belongs to at most  $k^4$  members of  $\mathcal{H}$ .*
- (b) *Each pair of edges is contained in at most  $k^2$  members of  $\mathcal{H}$ .*

The proof of [Lemma 4.3](#) employs several simple but important ideas that can be used in a variety of other settings, so let us sketch the details. The first key idea, which was first used in [Morris and Saxton \[ibid.\]](#), is to build the required family  $\mathcal{H}$  one  $C_4$  at a time. Let us say that a collection  $\mathcal{H}$  of copies of  $C_4$  is *legal* if it satisfies conditions (a) and (b) and suppose that we have already found a legal collection  $\mathcal{H}_m$  of  $m$  copies of  $C_4$  in  $G$ . Note that we are done if  $m \geq \beta k^5 \cdot \text{ex}(n, C_4)$ , so let us assume that the reverse inequality holds and construct a legal collection  $\mathcal{H}_{m+1} \supset \mathcal{H}_m$  of  $m + 1$  copies of  $C_4$  in  $G$ .

We claim that there exists a collection  $\mathcal{Q}_m$  of  $\beta k^5 \cdot \text{ex}(n, C_4)$  copies of  $C_4$  in  $G$ , any of which can be added to  $\mathcal{H}_m$  without violating conditions (a) and (b), that is, such that  $\mathcal{H}_m \cup \{C\}$  is legal for any  $C \in \mathcal{Q}_m$ . (Let us call these *good* copies of  $C_4$ .) Since  $m < \beta k^5 \cdot \text{ex}(n, C_4)$ , then at least one element of  $\mathcal{Q}_m$  is not already in  $\mathcal{H}_m$ , so this will be sufficient to prove the lemma.

To find  $\mathcal{Q}_m$ , observe first that (by simple double-counting) at most  $4\beta k \cdot \text{ex}(n, C_4)$  edges of  $G$  lie in exactly  $k^4$  members of  $\mathcal{H}_m$  and similarly at most  $6\beta k^3 \cdot \text{ex}(n, C_4)$  pairs of edges of  $G$  lie in exactly  $k^2$  members of  $\mathcal{H}_m$ . Now, consider a random subset  $A \subset V(G)$  of size  $pn$ , where  $p = D/k^2$  for some large constant  $D$ . Typically  $G[A]$  contains about

$p^2k \cdot \text{ex}(n, C_4)$  edges. After removing from  $G[A]$  all *saturated* edges (i.e., those belonging to  $k^4$  members of  $\mathcal{H}_m$ ) and one edge from each *saturated* pair (i.e., pair of edges that is contained in  $k^2$  members of  $\mathcal{H}_m$ ), we expect to end up with at least

$$p^2k \cdot \text{ex}(n, C_4) - 4\beta p^2k \cdot \text{ex}(n, C_4) - 6\beta p^3k^3 \cdot \text{ex}(n, C_4) \geq \frac{p^2k \cdot \text{ex}(n, C_4)}{2} \geq 2 \cdot \text{ex}(pn, C_4)$$

edges, where the first inequality follows since  $p = D/k^2$  and  $\beta$  is sufficiently small, and the second holds because  $\text{ex}(n, C_4) = \Theta(n^{3/2})$  and  $D$  is sufficiently large. Finally, observe that any graph on  $pn$  vertices with at least  $2 \cdot \text{ex}(pn, C_4)$  edges contains at least

$$\text{ex}(pn, C_4) = \Omega(p^{3/2} \cdot \text{ex}(n, C_4))$$

copies of  $C_4$ . But each copy of  $C_4$  in  $G$  was included in the random subgraph  $G[A]$  with probability at most  $p^4$  and hence (with a little care) one can show that there must exist at least  $\Omega(p^{-5/2} \cdot \text{ex}(n, C_4))$  copies of  $C_4$  in  $G$  that avoid all saturated edges and pairs of edges. Since  $p^{-5/2} = k^5/D^{5/2}$  and  $\beta$  is sufficiently small, we have found  $\beta k^5 \cdot \text{ex}(n, C_4)$  good copies of  $C_4$  in  $G$ , as required.

We now show how one may combine [Lemma 4.3](#) and the hypergraph container lemma to construct families of containers for  $C_4$ -free graphs. Let  $\beta$  and  $k_0$  be the constants from the statement of [Lemma 4.3](#) and assume that  $G$  is an  $n$ -vertex graph with at least  $k \cdot \text{ex}(n, C_4)$  and at most  $2k \cdot \text{ex}(n, C_4)$  edges, where  $k \geq k_0$ . Denote by  $\mathcal{H}_G$  the 4-uniform hypergraph with vertex set  $E(G)$ , whose edges are the copies of  $C_4$  in  $G$  given by [Lemma 4.3](#). Since

$$v(\mathcal{H}_G) = e(G), \quad e(\mathcal{H}_G) \geq \beta k^5 \cdot \text{ex}(n, C_4), \quad \Delta_1(\mathcal{H}_G) \leq k^4, \quad \Delta_2(\mathcal{H}_G) \leq k^2,$$

and  $\Delta_3(\mathcal{H}_G) = \Delta_4(\mathcal{H}_G) = 1$ , the hypergraph  $\mathcal{H}_G$  satisfies the assumptions of the container lemma with  $r = \beta k \cdot \text{ex}(n, C_4)$  and  $b = 2k^{-1/3} \cdot \text{ex}(n, C_4)$ . Consequently, there exist an absolute constant  $\delta$  and a collection  $\mathcal{C}$  of subgraphs of  $G$  with the following properties:

- (a) every  $C_4$ -free subgraph of  $G$  is contained in some  $C \in \mathcal{C}$ ,
- (b) each  $C \in \mathcal{C}$  has at most  $(1 - \delta)e(G)$  edges,

and moreover

$$|\mathcal{C}| \leq \sum_{s=0}^{3b} \binom{e(G)}{s} \leq \left( \frac{e(G)}{b} \right)^{3b} \leq k^{4b} \leq \exp\left(8k^{-1/3} \log k \cdot \text{ex}(n, C_4)\right).$$

Note that we have just replaced a single container for the family of  $C_4$ -free subgraphs of  $G$  (namely  $G$  itself) with a small collection of containers for this family, each of which is

somewhat smaller than  $G$ . Since every  $C_4$ -free graph with  $n$  vertices is contained in  $K_n$ , by repeatedly applying this ‘breaking down’ process, we obtain the following container theorem for  $C_4$ -free graphs.

**Theorem 4.4.** *There exist constants  $k_0 > 0$  and  $C > 0$  such that the following holds for all  $n \in \mathbb{N}$  and  $k \geq k_0$ . There exists a collection  $\mathcal{G}(n, k)$  of at most*

$$\exp\left(\frac{C \log k}{k^{1/3}} \cdot \text{ex}(n, C_4)\right)$$

*graphs on  $n$  vertices such that*

$$e(G) \leq k \cdot \text{ex}(n, C_4)$$

*for every  $G \in \mathcal{G}(n, k)$  and every  $C_4$ -free graph on  $n$  vertices is a subgraph of some  $G \in \mathcal{G}(n, k)$ .*

To obtain the claimed upper bound on  $|\mathcal{G}(n, k)|$ , note that if  $k \cdot \text{ex}(n, C_4) \geq \binom{n}{2}$  then we may take  $\mathcal{G}(n, k) = \{K_n\}$ , and otherwise the argument presented above yields

$$|\mathcal{G}(n, k)| \leq |\mathcal{G}(n, k/(1-\delta))| \cdot \exp\left(8k^{-1/3} \log k \cdot \text{ex}(n, C_4)\right).$$

In particular, applying [Theorem 4.4](#) with  $k = k_0$ , we obtain a collection of  $2^{O(\text{ex}(n, C_4))}$  containers for  $C_4$ -free graphs on  $n$  vertices, each with  $O(\text{ex}(n, C_4))$  edges. This immediately implies that [Conjecture 4.1](#) holds for  $H = C_4$ . The proof for a general graph  $H$  (under the assumption that  $\text{ex}(n, H) = \Theta(n^\alpha)$  for some  $\alpha \in (1, 2)$ ) is similar, though the details are rather technical.

**4.1 Turán’s problem in random graphs.** Given that the problem of estimating  $|\mathcal{T}_n(H)|$  for bipartite graphs  $H$  is notoriously difficult, it should not come as a surprise that determining the typical value of the Turán number  $\text{ex}(G(n, p), C_4)$  for bipartite  $H$  also poses considerable challenges. Compared to the non-bipartite case, which was essentially solved by [Conlon and Gowers \[2016\]](#) and [Schacht \[2016\]](#), see [Theorem 1.1](#), the typical behaviour of  $\text{ex}(G(n, p), H)$  for bipartite graphs  $H$  is much more subtle.

For simplicity, let us restrict our attention to the case  $H = C_4$ . Recall from [Theorem 1.1](#) that the typical value of  $\text{ex}(G(n, p), C_4)$  changes from  $(1 + o(1))p\binom{n}{2}$  to  $o(pn^2)$  when  $p = \Theta(n^{-2/3})$ , as was first proved by [Haxell, Kohayakawa, and Łuczak \[1995\]](#). However, already several years earlier [Füredi \[1991\]](#) used the method of [Kleitman and Winston \[1982\]](#) to prove<sup>8</sup> the following much finer estimates of this extremal number for  $p$  somewhat above the threshold.

---

<sup>8</sup>To be precise, Füredi proved that, if  $m \geq 2n^{4/3}(\log n)^2$ , then there are at most  $(4n^3/m^2)^m$   $C_4$ -free graphs with  $n$  vertices and  $m$  edges, which implies the upper bounds in [Theorem 4.5](#). For the lower bounds, see [Kohayakawa, Kreuter, and Steger \[1998\]](#) and [Morris and Saxton \[2016\]](#).



**Theorem 4.5.** *Asymptotically almost surely,*

$$\text{ex}(G(n, p), C_4) = \begin{cases} (1 + o(1))p \binom{n}{2} & \text{if } n^{-1} \ll p \ll n^{-2/3}, \\ n^{4/3}(\log n)^{O(1)} & \text{if } n^{-2/3} \leq p \leq n^{-1/3}(\log n)^4, \\ \Theta(\sqrt{p} \cdot n^{3/2}) & \text{if } p \geq n^{-1/3}(\log n)^4. \end{cases}$$

We would like to draw the reader's attention to the (somewhat surprising) fact that in the middle range  $n^{-2/3+o(1)} \leq p \leq n^{-1/3+o(1)}$ , the typical value of  $\text{ex}(G(n, p), C_4)$  stays essentially constant. A similar phenomenon has been observed in random Turán problems for other forbidden bipartite graphs (even cycles and complete bipartite graphs, see Kohayakawa, Kreuter, and Steger [1998] and Morris and Saxton [2016]) as well as Turán-type problems in additive combinatorics, see Dellamonica, Kohayakawa, Lee, Rödl, and Samotij [2016b, n.d.]. It would be very interesting to determine whether or not a similar 'long flat segment' appears in the graph of  $p \mapsto \text{ex}(G(n, p), H)$  for every bipartite graph  $H$ . We remark that the lower bound in the middle range is given (very roughly speaking) by taking a random subgraph of  $G(n, p)$  with density  $n^{-2/3+o(1)}$  and then finding<sup>9</sup> a large  $C_4$ -free subgraph of this random graph; the lower bound in the top range is given by intersecting  $G(n, p)$  with a suitable blow-up of an extremal  $C_4$ -free graph and destroying any  $C_4$ s that occur; see Kohayakawa, Kreuter, and Steger [1998] and Morris and Saxton [2016] for details.

Even though Theorem 4.4 immediately implies that  $\text{ex}(G(n, p), C_4) = o(pn^2)$  if  $p \gg n^{-2/3} \log n$ , it is not strong enough to prove Theorem 4.5. A stronger container theorem for  $C_{2\ell}$ -free graphs (based on a supersaturation lemma that is sharper than Lemma 4.3) was obtained in Morris and Saxton [2016]. In the case  $\ell = 2$ , the statement is as follows.

**Theorem 4.6.** *There exist constants  $k_0 > 0$  and  $C > 0$  such that the following holds for all  $n \in \mathbb{N}$  and  $k_0 \leq k \leq n^{1/6} / \log n$ . There exists a collection  $\mathcal{G}(n, k)$  of at most*

$$\exp\left(\frac{C \log k}{k} \cdot \text{ex}(n, C_4)\right)$$

*graphs on  $n$  vertices such that*

$$e(G) \leq k \cdot \text{ex}(n, C_4)$$

*for every  $G \in \mathcal{G}(n, k)$  and every  $C_4$ -free graph on  $n$  vertices is a subgraph of some  $G \in \mathcal{G}(n, k)$ .*

<sup>9</sup>One easy way to do this is simply to remove one edge from each copy of  $C_4$ . A more efficient method, used by Kohayakawa, Kreuter, and Steger [1998] to improve the lower bound by a polylogarithmic factor, utilizes a version of the general result of Ajtai, Komlós, Pintz, Spencer, and Szemerédi [1982] on independent sets in hypergraphs obtained in Duke, Lefmann, and Rödl [1995]; see also Ferber, McKinley, and Samotij [n.d.].

Choosing  $k$  to be a suitable function of  $p$ , it is straightforward to use [Theorem 4.6](#) to prove a slightly weaker version of [Theorem 4.5](#), with an extra factor of  $\log n$  in the upper bound on  $\text{ex}(G(n, p), C_4)$ . As usual, this logarithmic factor can be removed via a more careful application of the container method, using the fact that the fingerprint of an independent set is contained in it, cf. the discussion in [Section 3](#); see [Morris and Saxton \[ibid.\]](#) for the details. However, we are not able to determine the correct power of  $\log n$  in  $\text{ex}(G(n, p), C_4)$  in the middle range  $n^{-2/3+o(1)} \ll p \ll n^{-1/3+o(1)}$  and we consider this to be an important open problem. It would also be very interesting to prove similarly sharp container theorems for other bipartite graphs  $H$ .

## 5 Containers for multicoloured structures

All of the problems that we have discussed so far, and many others, are naturally expressed as questions about independent sets in various hypergraphs. There are, however, questions of a very similar flavour that are not easily described in this way but are still amenable to the container method. As an example, consider the problem of enumerating large graphs with no *induced* copy of a given graph  $H$ . We shall say that a graph  $G$  is *induced- $H$ -free* if no subset of vertices of  $G$  induces a subgraph isomorphic to  $H$ . As it turns out, it is beneficial to think of an  $n$ -vertex graph  $G$  as the characteristic function of its edge set. A function  $g: E(K_n) \rightarrow \{0, 1\}$  is the characteristic function of an induced- $H$ -free graph if and only if for every set  $W$  of  $v(H)$  vertices of  $K_n$ , the restriction of  $g$  to the set of pairs of vertices of  $W$  is not the characteristic function of the edge set of  $H$ . In particular, viewing  $g$  as the set of pairs  $\{(e, g(e)) : e \in E(K_n)\}$ , we see that if  $g$  represents an induced- $H$ -free graph, then it is an independent set in the  $\binom{v(H)}{2}$ -uniform hypergraph  $\mathcal{H}$  with vertex set  $E(K_n) \times \{0, 1\}$  whose edges are the characteristic functions of all copies of  $H$  in  $K_n$ ; formally, for every injection  $\varphi: V(H) \rightarrow V(K_n)$ , the set

$$\{(\varphi(u)\varphi(v), 1) : uv \in E(H)\} \cup \{(\varphi(u)\varphi(v), 0) : uv \notin E(H)\}$$

is an edge of  $\mathcal{H}$ . Even though the converse statement is not true and not every independent set of  $\mathcal{H}$  corresponds to an induced- $H$ -free graph, since we are usually interested in bounding the number of such graphs from above, the above representation can be useful. In particular, [Saxton and Thomason \[2015\]](#) applied the container method to the hypergraph  $\mathcal{H}$  described above to reprove the following analogue of (8), which was originally obtained by [Alekseev \[1992\]](#) and by [Bollobás and Thomason \[1995, 1997\]](#). Letting  $\mathfrak{F}_n^{\text{ind}}(H)$  denote the family of all induced- $H$ -free graphs with vertex set  $\{1, \dots, n\}$ , we have

$$|\mathfrak{F}_n^{\text{ind}}(H)| = 2^{(1-1/\text{col}(H))\binom{n}{2}+o(n^2)},$$

where  $\text{col}(H)$  is the so-called *colouring number*<sup>10</sup> of  $H$ .

This idea of embedding non-monotone properties (such as the family of induced- $H$ -free graphs) into the family of independent sets of an auxiliary hypergraph has been used in several other works. In particular, Kühn, Osthus, Townsend, and Zhao [2017] used it to describe the typical structure of oriented graphs without a transitive tournament of a given order. The recent independent works of Falgas-Ravry, O'Connell, Strömberg, and Uzzell [n.d.] and of Terry [n.d.] have developed a general framework for studying various enumeration problems in the setting of multicoloured graphs Falgas-Ravry, O'Connell, Strömberg, and Uzzell [n.d.] and, more generally, in the very abstract setting of finite (model theoretic) structures Terry [n.d.]. In order to illustrate some of the ideas involved in applications of this kind, we will discuss the problem of counting finite metric spaces with bounded integral distances.

**5.1 Counting metric spaces.** Let  $\mathfrak{M}_n^M$  denote the family of metric spaces on a given set of  $n$  points with distances in the set  $\{1, \dots, M\}$ . Thus  $\mathfrak{M}_n^M$  may be viewed as the set of all functions  $d: E(K_n) \rightarrow \{1, \dots, M\}$  that satisfy the triangle inequality  $d(uv) \leq d(uw) + d(wv)$  for all  $u, v, w$ . Since  $x \leq y + z$  for all  $x, y, z \in \{M/2, \dots, M\}$ , we have

$$(9) \quad |\mathfrak{M}_n^M| \geq \left| \left\{ \left\lceil \frac{M}{2} \right\rceil, \dots, M \right\}^{\binom{n}{2}} \right| = \left\lceil \frac{M+1}{2} \right\rceil^{\binom{n}{2}}.$$

Inspired by a continuous version of the model suggested Benjamini (and first studied by Kozma, Meyerovitch, Peled, and Samotij [n.d.]), Mubayi and Terry [n.d.] proved that for every fixed even  $M$ , the converse of (9) holds asymptotically, that is,  $|\mathfrak{M}_n^M| \leq (1 + o(1)) \left\lceil \frac{M+1}{2} \right\rceil^{\binom{n}{2}}$  as  $n \rightarrow \infty$ . The problem becomes much more difficult, however, when one allows  $M$  to grow with  $n$ . For example, if  $M \gg \sqrt{n}$  then the lower bound

$$|\mathfrak{M}_n^M| \geq \left[ \left( \frac{1}{2} + \frac{c}{\sqrt{n}} \right) M \right]^{\binom{n}{2}}$$

for some absolute constant  $c > 0$ , proved by Kozma, Meyerovitch, Peled, and Samotij [n.d.], is stronger than (9). Balogh and Wagner [2016] proved strong upper bounds on  $|\mathfrak{M}_n^M|$  under the assumption that  $M \ll n^{1/3}/(\log n)^{4/3+o(1)}$ . The following almost optimal estimate was subsequently obtained by Kozma, Meyerovitch, Peled, and Samotij [n.d.].

<sup>10</sup>The *colouring number* of a graph  $H$  is the largest integer  $r$  such that for some pair  $(r_1, r_2)$  satisfying  $r_1 + r_2 = r$ , the vertex set of  $H$  cannot be partitioned into  $r_1$  cliques and  $r_2$  independent sets.

**Theorem 5.1.** *There exists a constant  $C$  such that*

$$(10) \quad |\mathfrak{M}_n^M| \leq \left[ \left( \frac{1}{2} + \frac{2}{M} + \frac{C}{\sqrt{n}} \right) M \right]^{\binom{n}{2}}$$

for all  $n$  and  $M$ .

Here, we present an argument due to Morris and Samotij that derives a mildly weaker estimate using the hypergraph container lemma. Let  $\mathcal{H}$  be the 3-uniform hypergraph with vertex set  $E(K_n) \times \{1, \dots, M\}$  whose edges are all triples  $\{(e_1, d_1), (e_2, d_2), (e_3, d_3)\}$  such that  $e_1, e_2, e_3$  form a triangle in  $K_n$  but  $d_{\sigma(1)} + d_{\sigma(2)} < d_{\sigma(3)}$  for some permutation  $\sigma$  of  $\{1, 2, 3\}$ . The crucial observation, already made in [Balogh and Wagner \[2016\]](#), is that every metric space  $d: E(K_n) \rightarrow \{1, \dots, M\}$ , viewed as the set of pairs  $\{(e, d(e)) : e \in E(K_n)\}$ , is an independent set of  $\mathcal{H}$ . This enables the use of the hypergraph container method for bounding  $|\mathfrak{M}_n^M|$  from above. Define the volume of a set  $A \subset E(K_n) \times \{1, \dots, M\}$ , denoted by  $\text{vol}(A)$ , by

$$\text{vol}(A) = \prod_{e \in E(K_n)} \left| \left\{ d \in \{1, \dots, M\} : (e, d) \in A \right\} \right|$$

and observe that  $A$  contains at most  $\text{vol}(A)$  elements of  $\mathfrak{M}_n^M$ . The following supersaturation lemma was proved by Morris and Samotij.

**Lemma 5.2.** *Let  $n \geq 3$  and  $M \geq 1$  be integers and suppose that  $A \subset E(K_n) \times \{1, \dots, M\}$  satisfies*

$$\text{vol}(A) = \left[ \left( \frac{1}{2} + \varepsilon \right) M \right]^{\binom{n}{2}}$$

for some  $\varepsilon \geq 10/M$ . Then there exist  $m \leq M$  and a set  $A' \subset A$  with  $|A'| \leq mn^2$ , such that the hypergraph  $\mathcal{H}' = \mathcal{H}[A']$  satisfies

$$e(\mathcal{H}') \geq \frac{\varepsilon m^2 M}{50 \log M} \binom{n}{3}, \quad \Delta_1(\mathcal{H}') \leq 4m^2 n, \quad \text{and} \quad \Delta_2(\mathcal{H}') \leq 2m.$$

It is not hard to verify that the hypergraph  $\mathcal{H}'$  given by [Lemma 5.2](#) satisfies the assumptions of the hypergraph container lemma stated in [Section 3](#) with  $r = \varepsilon n^2 M / (2^{11} \log M)$  and  $b = O(n^{3/2})$ . Consequently, there exist an absolute constant  $\delta$  and a collection  $\mathcal{C}$  of subsets of  $A'$ , with

$$|\mathcal{C}| \leq \exp \left( O(n^{3/2} \log(nM)) \right),$$

such that, setting  $A_C = C \cup (A \setminus A') = A \setminus (A' \setminus C)$  for each  $C \in \mathcal{C}$ , the following properties hold:

- (a) every metric space in  $A$ , viewed as a subset of  $E(K_n) \times \{1, \dots, M\}$ , is contained in  $A_C$  for some  $C \in \mathcal{C}$ , and
- (b)  $|C| \leq |A'| - \delta r$  for every  $C \in \mathcal{C}$ .

Observe that

$$\begin{aligned} \text{vol}(A_C) &\leq \left( \frac{M-1}{M} \right)^{|A' \setminus C|} \text{vol}(A) \leq e^{-\delta r/M} \text{vol}(A) \\ &\leq e^{-\delta \varepsilon n^2 / (2^{11} \log M)} \text{vol}(A) \leq \left[ \left( \frac{1}{2} + \left( 1 - \frac{\delta}{2^{12} \log M} \right) \varepsilon \right) M \right]^{\binom{n}{2}}. \end{aligned}$$

Since every metric space in  $\mathfrak{M}_n^M$  is contained in  $E(K_n) \times \{1, \dots, M\}$ , by recursively applying this ‘breaking down’ process to depth  $O(\log M)^2$ , we obtain a family of

$$\exp \left( O(n^{3/2} (\log M)^2 \log(nM)) \right)$$

subsets of  $E(K_n) \times \{1, \dots, M\}$ , each of volume at most  $(M/2 + 10)^{\binom{n}{2}}$ , that cover all of  $\mathfrak{M}_n^M$ . This implies that

$$|\mathfrak{M}_n^M| \leq \left[ \left( \frac{1}{2} + \frac{10}{M} + \frac{C (\log M)^2 \log(nM)}{\sqrt{n}} \right) M \right]^{\binom{n}{2}},$$

which, as promised, is only slightly weaker than (10).

## 6 An asymmetric container lemma

The approach to studying the family of induced- $H$ -free graphs described in the previous section has one (rather subtle) drawback: it embeds  $\mathfrak{T}_n^{\text{ind}}(H)$  into the family of independent sets of a  $\binom{v(H)}{2}$ -uniform hypergraph with  $\Theta(n^2)$  vertices. As a result, the hypergraph container lemma produces fingerprints of the same size as for the family of graphs without a clique on  $v(H)$  vertices. This precludes the study of various threshold phenomena in the context of sparse induced- $H$ -free graphs with the use of the hypergraph container lemma presented in Section 3; this is in sharp contrast with the non-induced case, where the container method proved very useful.

In order to alleviate this shortcoming, Morris, Samotij, and Saxton [n.d.] proved a version of the hypergraph container lemma for 2-coloured structures that takes into account the possible asymmetries between the two colours. We shall not give the precise statement of this new container lemma here (since it is rather technical), but we would like to

emphasize the following key fact: it enables one to construct families of containers for induced- $H$ -free graphs with fingerprints of size  $\Theta(n^{2-1/m_2(H)})$ , as in the non-induced case.

To demonstrate the power of the asymmetric container lemma, the following application was given in [Morris, Samotij, and Saxton \[ibid.\]](#). Let us say that a graph  $G$  is  $\varepsilon$ -close to a split graph if there exists a partition  $V(G) = A \cup B$  such that  $e(G[A]) \geq (1 - \varepsilon) \binom{|A|}{2}$  and  $e(G[B]) \leq \varepsilon e(G)$ .

**Theorem 6.1.** *For every  $\varepsilon > 0$ , there exists a  $\delta > 0$  such that the following holds. Let  $G$  be a uniformly chosen induced- $C_4$ -free graph with vertex set  $\{1, \dots, n\}$  and  $m$  edges.*

- (a) *If  $n \ll m \ll \delta n^{4/3}(\log n)^{1/3}$ , then a.a.s.  $G$  is not  $1/4$ -close to a split graph.*
- (b) *If  $n^{4/3}(\log n)^4 \leq m \leq \delta n^2$ , then a.a.s.  $G$  is  $\varepsilon$ -close to a split graph.*

[Theorem 6.1](#) has the following interesting consequence: it allows one to determine the number of edges in (and, sometimes, also the typical structure of) the binomial random graph  $G(n, p)$  conditioned on the event that it does not contain an induced copy of  $C_4$ . Let us denote by  $G_{n,p}^{\text{ind}}(C_4)$  the random graph chosen according to this conditional distribution.

**Corollary 6.2.** *The following bounds hold asymptotically almost surely as  $n \rightarrow \infty$ :*

$$e(G_{n,p}^{\text{ind}}(C_4)) = \begin{cases} (1 + o(1))p \binom{n}{2} & \text{if } n^{-1} \ll p \ll n^{-2/3}, \\ n^{4/3}(\log n)^{O(1)} & \text{if } n^{-2/3} \leq p \leq n^{-1/3}(\log n)^4, \\ \Theta(p^2 n^2 / \log(1/p)) & \text{if } p \geq n^{-1/3}(\log n)^4. \end{cases}$$

We would like to emphasize the (surprising) similarity between the statements of [Theorem 4.5](#) and [Corollary 6.2](#). In particular, the graph of  $p \mapsto e(G_{n,p}^{\text{ind}}(C_4))$  contains exactly the same ‘long flat segment’ as the graph of  $p \mapsto \text{ex}(G(n, p), C_4)$ , even though the shape of the two graphs above this range is quite different. We do not yet fully understand this phenomenon and it would be interesting to investigate whether or not the function  $p \mapsto e(G_{n,p}^{\text{ind}}(H))$  exhibits similar behaviour for other bipartite graphs  $H$ .

## 7 Hypergraphs of unbounded uniformity

Since the hypergraph container lemma provides explicit dependencies between the various parameters in its statement, it is possible to apply the container method even when the uniformity of the hypergraph considered is a growing function of the number of its vertices. Perhaps the first result of this flavour was obtained by [Mousset, Nenadov, and Steger \[2014\]](#), who proved an upper bound of  $2^{\text{ex}(n, K_r) + o(n^2/r)}$  on the number of  $n$ -vertex

$K_r$ -free graphs for all  $r \leq (\log_2 n)^{1/4}/2$ . Subsequently, [Balogh, Bushaw, Collares, Liu, Morris, and Sharifzadeh \[2017\]](#) strengthened this result by establishing the following precise description of the typical structure of large  $K_r$ -free graphs.

**Theorem 7.1.** *If  $r \leq (\log_2 n)^{1/4}$ , then almost all  $K_r$ -free graphs with  $n$  vertices are  $(r-1)$ -partite.*

Around the same time, the container method applied to hypergraphs with unbounded uniformity was used to analyse Ramsey properties of random graphs and hypergraphs, leading to improved upper bounds on several well-studied functions. In particular, [Rödl, Ruciński, and Schacht \[2017\]](#) gave the following upper bound on the so-called Folkman numbers.

**Theorem 7.2.** *For all integers  $k \geq 3$  and  $r \geq 2$ , there exists a  $K_{k+1}$ -free graph with*

$$\exp(Ck^4 \log k + k^3 r \log r)$$

*vertices, such that every  $r$ -colouring of its edges contains a monochromatic copy of  $K_k$ .*

The previously best known bound was doubly exponential in  $k$ , even in the case  $r = 2$ . Not long afterwards, [Conlon, Dellamonica, La Fleur, Rödl, and Schacht \[2017\]](#) used a similar method to prove the following strong upper bounds on the induced Ramsey numbers of hypergraphs. Define the tower functions  $t_k(x)$  by  $t_1(x) = x$  and  $t_{i+1}(x) = 2^{t_i(x)}$  for each  $i \geq 1$ .

**Theorem 7.3.** *For each  $k \geq 3$  and  $r \geq 2$ , there exists  $c$  such that the following holds. For every  $k$ -uniform hypergraph  $F$  on  $m$  vertices, there exists a  $k$ -uniform hypergraph  $G$  on  $t_k(cm)$  vertices, such that every  $r$ -colouring of  $E(G)$  contains a monochromatic induced copy of  $F$ .*

Finally, let us mention a recent result of [Balogh and Solymosi \[n.d.\]](#), whose proof is similar to that of [Theorem 2.8](#), which we outlined in [Section 2.4](#). Given a family  $\mathcal{F}$  of subsets of an  $n$ -element set  $\Omega$ , an  $\varepsilon$ -net of  $\mathcal{F}$  is a set  $A \subset \Omega$  that intersects every member of  $\mathcal{F}$  with at least  $\varepsilon n$  elements. The concept of an  $\varepsilon$ -net plays an important role in computer science, for example in computational geometry and approximation theory. In a seminal paper, [Haussler and Welzl \[1987\]](#) proved that every set system with VC-dimension<sup>11</sup>  $d$  has an  $\varepsilon$ -net of size  $O((d/\varepsilon) \log(d/\varepsilon))$ . It was believed for more than twenty years that for ‘geometric’ families, the  $\log(d/\varepsilon)$  factor can be removed; however, this was disproved by [Alon \[2012\]](#), who constructed, for each  $C > 0$ , a set of points in the plane such that the smallest  $\varepsilon$ -net for the family of lines (whose VC-dimension is 2) has size at least  $C/\varepsilon$ .

<sup>11</sup>The VC-dimension (VC stands for Vapnik–Chervonenkis) of a family  $\mathcal{F}$  of subsets of  $\Omega$  is the largest size of a set  $X \subset \Omega$  such that the set  $\{A \cap X : A \in \mathcal{F}\}$  has  $2^{|X|}$  elements.

By applying the container method to the hypergraph of collinear  $k$ -tuples in the  $k$ -dimensional  $2^{k^2} \times \dots \times 2^{k^2}$  integer grid, [Balogh and Solymosi \[n.d.\]](#) gave the following stronger lower bound.

**Theorem 7.4.** *For each  $\varepsilon > 0$ , there exists a set  $S \subset \mathbb{R}^2$  such that the following holds. If  $T \subset S$  intersects every line that contains at least  $\varepsilon|S|$  elements of  $S$ , then*

$$|T| \geq \frac{1}{\varepsilon} \left( \log \frac{1}{\varepsilon} \right)^{1/3+o(1)}.$$

It was conjectured by [Alon \[2012\]](#) that there are sets of points in the plane whose small- $\varepsilon$ -nets (for the family of lines) contain  $\Omega(1/\varepsilon \log(1/\varepsilon))$  points.

## 8 Some further applications

There are numerous applications of the method of containers that we do not have space to discuss in detail. Still, we would like to finish this survey by briefly mentioning just a few of them.

**8.1 List colouring.** A hypergraph  $\mathcal{H}$  is said to be  $k$ -choosable if for every assignment of a list  $L_v$  of  $k$  colours to each vertex  $v$  of  $\mathcal{H}$ , it is possible to choose for each  $v$  a colour from the list  $L_v$  in such a way that no edge of  $\mathcal{H}$  has all its vertices of the same colour. The smallest  $k$  for which  $\mathcal{H}$  is  $k$ -choosable is usually called the *list chromatic number* of  $\mathcal{H}$  and denoted by  $\chi_\ell(\mathcal{H})$ . [Alon \[1993, 2000\]](#) showed that for graphs, the list chromatic number grows with the minimum degree, in stark contrast with the usual chromatic number; more precisely,  $\chi_\ell(G) \geq (1/2 + o(1)) \log_2 \delta(G)$  for every graph  $G$ . The following generalisation of this result, which also improves the constant  $1/2$ , was proved by [Saxton and Thomason \[2015\]](#), see also [Saxton and Thomason \[2012, 2016\]](#).

**Theorem 8.1.** *Let  $\mathcal{H}$  be a  $k$ -uniform hypergraph with average degree  $d$  and  $\Delta_2(\mathcal{H}) = 1$ . Then, as  $d \rightarrow \infty$ ,*

$$\chi_\ell(\mathcal{H}) \geq \left( \frac{1}{(k-1)^2} + o(1) \right) \log_k d.$$

Moreover, if  $\mathcal{H}$  is  $d$ -regular, then

$$\chi_\ell(\mathcal{H}) \geq \left( \frac{1}{k-1} + o(1) \right) \log_k d.$$

We remark that proving lower bounds for the list chromatic number of simple hypergraphs was one of the original motivations driving the development of the method of hypergraph containers.



**8.2 Additive combinatorics.** The method of hypergraph containers has been applied to a number of different number-theoretic objects, including sum-free sets [Alon, Balogh, Morris, and Samotij \[2014a,b\]](#) and [Balogh, Liu, Sharifzadeh, and Treglown \[2015, n.d.\]](#), Sidon sets [Saxton and Thomason \[2016\]](#), sets containing no  $k$ -term arithmetic progression [Balogh, Liu, and Sharifzadeh \[2017\]](#) and [Balogh, Morris, and Samotij \[2015\]](#), and general systems of linear equations [Saxton and Thomason \[2016\]](#). (See also [Green \[2004\]](#), [Green and Ruzsa \[2004\]](#), and [Sapozhenko \[2003\]](#) for early applications of the container method to sum-free sets and [Dellamonica, Kohayakawa, Lee, Rödl, and Samotij \[2016a,b, n.d.\]](#) and [Kohayakawa, Lee, Rödl, and Samotij \[2015\]](#) for applications of graph containers to  $B_h$ -sets.) Here we will mention just three of these results.

Let us begin by recalling that a *Sidon set* is a set of integers containing no non-trivial solutions of the equation  $x + y = z + w$ . Results of Chowla, Erdős, Singer, and Turán from the 1940s imply that the maximum size of a Sidon set in  $\{1, \dots, n\}$  is  $(1 + o(1))\sqrt{n}$  and it was conjectured by [Cameron and Erdős \[1990\]](#) that the number of such sets is  $2^{(1+o(1))\sqrt{n}}$ . This conjecture was disproved by [Saxton and Thomason \[2016\]](#), who gave a construction of  $2^{(1+\varepsilon)\sqrt{n}}$  Sidon sets (for some  $\varepsilon > 0$ ), and also used the hypergraph container method to reprove the following theorem, which was originally obtained in [Kohayakawa, Lee, Rödl, and Samotij \[2015\]](#) using the graph container method.

**Theorem 8.2.** *There are  $2^{O(\sqrt{n})}$  Sidon sets in  $\{1, \dots, n\}$ .*

[Dellamonica, Kohayakawa, Lee, Rödl, and Samotij \[n.d.\]](#) later generalized these results to  $B_h$ -sets, that is, set of integers containing no non-trivial solutions of the equation  $x_1 + \dots + x_h = y_1 + \dots + y_h$ .

The second result we would like to state was proved by [Balogh, Liu, and Sharifzadeh \[2017\]](#), and inspired the proof presented in [Section 4](#). Let  $r_k(n)$  be the largest size of a subset of  $\{1, \dots, n\}$  containing no  $k$ -term arithmetic progressions.

**Theorem 8.3.** *For each integer  $k \geq 3$ , there exist a constant  $C$  and infinitely many  $n \in \mathbb{N}$  such that there are at most  $2^{Cr_k(n)}$  subsets of  $\{1, \dots, n\}$  containing no  $k$ -term arithmetic progression.*

We recall (see, e.g., [Gowers \[2013\]](#)) that obtaining good bounds on  $r_k(n)$  is a well-studied and notoriously difficult problem. The proof of [Theorem 8.3](#) avoids these difficulties by exploiting merely the ‘self-similarity’ property of the hypergraph encoding arithmetic progressions in  $\{1, \dots, n\}$ , cf. the discussion in [Section 3](#) and the proof of [Lemma 4.3](#).

The final result we would like to mention was one of the first applications of (and original motivations for the development of) the method of hypergraph containers. Recall that the Cameron–Erdős conjecture, proved by [Green \[2004\]](#) and, independently, by

Sapozhenko [2003], states that there are only  $O(2^{n/2})$  sum-free subsets of  $\{1, \dots, n\}$ . The following sparse analogue of the Cameron–Erdős conjecture was proved by Alon, Balogh, Morris, and Samotij [2014a] using an early version of the hypergraph container lemma for 3-uniform hypergraphs.

**Theorem 8.4.** *There exists a constant  $C$  such that, for every  $n \in \mathbb{N}$  and every  $1 \leq m \leq \lfloor n/2 \rfloor$ , the set  $\{1, \dots, n\}$  contains at most  $2^{Cn/m} \binom{\lfloor n/2 \rfloor}{m}$  sum-free sets of size  $m$ .*

We remark that if  $m \geq \sqrt{n}$ , then Theorem 8.4 is sharp up to the value of  $C$ , since in this case there is a constant  $c > 0$  such that there are at least  $2^{cn/m} \binom{\lfloor n/2 \rfloor}{m}$  sum-free  $m$ -subsets of  $\{1, \dots, n\}$ . For smaller values of  $m$  the answer is different, but the problem in that range is much easier and can be solved using standard techniques. Let us also mention that in the case  $m \gg \sqrt{n \log n}$ , the structure of a typical sum-free  $m$ -subset of  $\{1, \dots, n\}$  was also determined quite precisely in Alon, Balogh, Morris, and Samotij [ibid.].

Finally, we would like to note that, although the statements of Theorems 8.2, 8.3 and 8.4 are somewhat similar, the difficulties encountered during their proofs are completely different.

**8.3 Sharp thresholds for Ramsey properties.** Given an integer  $k \geq 3$ , let us say that a set  $A \subset \mathbb{Z}_n$  has the *van der Waerden property* for  $k$  if every 2-colouring of the elements of  $A$  contains a monochromatic  $k$ -term arithmetic progression; denote this by  $A \rightarrow (k\text{-AP})$ . Rödl and Ruciński [1995] determined the threshold for the van der Waerden property in random subsets of  $\mathbb{Z}_n$  for every  $k \in \mathbb{N}$ . Combining the sharp threshold technology of Friedgut [1999] with the method of hypergraph containers, Friedgut, Hàn, Person, and Schacht [2016] proved that this threshold is sharp. Let us write  $\mathbb{Z}_{n,p}$  to denote a  $p$ -random subset of  $\mathbb{Z}_n$  (i.e., each element is included independently with probability  $p$ ).

**Theorem 8.5.** *For every  $k \geq 3$ , there exist constants  $c_1 > c_0 > 0$  and a function  $p_c: \mathbb{N} \rightarrow [0, 1]$  satisfying  $c_0 n^{-1/(k-1)} < p_c(n) < c_1 n^{-1/(k-1)}$  for every  $n \in \mathbb{N}$ , such that, for every  $\varepsilon > 0$ ,*

$$\mathbb{P}\left(\mathbb{Z}_{n,p} \rightarrow (k\text{-AP})\right) \rightarrow \begin{cases} 0 & \text{if } p \leq (1 - \varepsilon) p_c(n), \\ 1 & \text{if } p \geq (1 + \varepsilon) p_c(n), \end{cases}$$

as  $n \rightarrow \infty$ .

The existence of a sharp threshold in the context of Ramsey’s theorem for the triangle was obtained several years earlier, by Friedgut, Rödl, Ruciński, and Tetali [2006]. Very recently, using similar methods to those in Friedgut, Hàn, Person, and Schacht [2016], Schacht and Schenbourg [2018] gave a simpler proof of this theorem and also generalised it to a large family of graphs, including all odd cycles.

**8.4 Maximal triangle-free graphs and sum-free sets.** In contrast to the large body of work devoted to counting and describing the typical structure of  $H$ -free graphs, relatively little is known about  $H$ -free graphs that are *maximal* (with respect to the subgraph relation). The following construction shows that there are at least  $2^{n^2/8}$  maximal triangle-free graphs with vertex set  $\{1, \dots, n\}$ . Fix a partition  $X \cup Y = \{1, \dots, n\}$  with  $|X|$  even. Define  $G$  by letting  $G[X]$  be a perfect matching, leaving  $G[Y]$  empty, and adding to  $E(G)$  exactly one of  $xy$  or  $x'y$  for every edge  $xx' \in E(G[X])$  and every  $y \in Y$ . It is easy to verify that all such graphs are triangle-free and that almost all of them are maximal.

Using the container theorem for triangle-free graphs (Theorem 2.1), Balogh and Petříčková [2014] proved that the construction above is close to optimal by showing that there are at most  $2^{n^2/8+o(n^2)}$  maximal triangle-free graphs on  $\{1, \dots, n\}$ . Following this breakthrough, Balogh, Liu, Petříčková, and Sharifzadeh [2015] proved the following much stronger theorem, which states that in fact almost all maximal triangle-free graphs can be constructed in this way.

**Theorem 8.6.** *For almost every maximal triangle-free graph  $G$  on  $\{1, \dots, n\}$ , there is a vertex partition  $X \cup Y$  such that  $G[X]$  is a perfect matching and  $Y$  is an independent set.*

A similar result for sum-free sets was obtained by Balogh, Liu, Sharifzadeh, and Treglown [2015, n.d.], who determined the number of maximal sum-free subsets of  $\{1, \dots, n\}$  asymptotically. However, the problem of estimating the number of maximal  $H$ -free graphs for a general graph  $H$  is still wide open. In particular, generalizing the results of Balogh, Liu, Petříčková, and Sharifzadeh [2015] and Balogh and Petříčková [2014] to the family of maximal  $K_k$ -free graphs seems to be a very interesting and difficult open problem.

**8.5 Containers for rooted hypergraphs.** A family  $\mathcal{F}$  of finite sets is *union-free* if  $A \cup B \neq C$  for every three distinct sets  $A, B, C \in \mathcal{F}$ . Kleitman [1976] proved that every union-free family in  $\{1, \dots, n\}$  contains at most  $(1 + o(1))\binom{n}{n/2}$  sets; this is best possible as the family of all  $\lfloor n/2 \rfloor$ -element subsets of  $\{1, \dots, n\}$  is union-free. Balogh and Wagner [2017] proved the following natural counting counterpart of Kleitman's theorem, confirming a conjecture of Burosch, Demetrovics, Katona, Kleitman, and Sapozhenko [1991].

**Theorem 8.7.** *There are  $2^{(1+o(1))\binom{n}{n/2}}$  union-free families in  $\{1, \dots, n\}$ .*

It is natural to attempt to prove this theorem by applying the container method to the 3-uniform hypergraph  $\mathcal{H}$  that encodes triples  $\{A, B, C\}$  with  $A \cup B = C$ . However, there is a problem: for any pair  $(B, C)$ , there exist  $2^{|B|}$  sets  $A$  such that  $A \cup B = C$  and this means that  $\Delta_2(\mathcal{H})$  is too large for a naive application of the hypergraph container lemma.

In order to overcome this difficulty, Balogh and Wagner developed in [Balogh and Wagner \[2017\]](#) a new container theorem for ‘rooted’ hypergraphs (each edge has a designated root vertex) that exploits the asymmetry of the identity  $A \cup B = C$ . In particular, note that while the degree of a pair  $(B, C)$  can be large, the pair  $\{A, B\}$  uniquely determines  $C$ ; it turns out that this is sufficient to prove a suitable container theorem. We refer the reader to [Balogh and Wagner \[ibid.\]](#) for the details.

**8.6 Probabilistic embedding in sparse graphs.** The celebrated regularity lemma of [Szemerédi \[1978\]](#) states that, roughly speaking, the vertex set of every graph can be divided into a bounded number of parts in such a way that most of the bipartite subgraphs induced by pairs of parts are pseudorandom; such a partition is called a *regular partition*. The strength of the regularity lemma stems from the so-called counting and embedding lemmas, which tell us approximately how many copies of a particular subgraph a graph  $G$  contains in terms of basic parameters of the regular partition of  $G$ . While the original statement of the regularity lemma applied only to dense graphs (i.e.,  $n$ -vertex graphs with  $\Omega(n^2)$  edges), the works of [Kohayakawa \[1997\]](#), Rödl (unpublished), and [Scott \[2011\]](#) provide extensions of the lemma that are applicable to sparse graphs. However, these extensions come with a major caveat: the counting and embedding lemmas do not extend to sparse graphs; this unfortunate fact was observed by Łuczak. Nevertheless, it seemed likely that such atypical graphs that fail the counting or embedding lemmas are so rare that they typically do not appear in random graphs. This belief was formalised in a conjecture of [Kohayakawa, Łuczak, and Rödl \[1997\]](#), which can be seen as a ‘probabilistic’ version of the embedding lemma.

The proof of this conjecture, discovered by the authors of this survey [Balogh, Morris, and Samotij \[2015\]](#) and by [Saxton and Thomason \[2015\]](#), was one of the original applications of the hypergraph container lemma. Let us mention here that a closely related result was proved around the same time by [Conlon, Gowers, Samotij, and Schacht \[2014\]](#). A strengthening of the KŁR conjecture, a ‘probabilistic’ version of the counting lemma, proposed by [Gerke, Marciniszyn, and Steger \[2007\]](#), remains open.

**Acknowledgments.** The authors would like to thank Noga Alon, Béla Bollobás, David Conlon, Yoshi Kohayakawa, Gady Kozma, Tom Meyerovitch, Ron Peled, David Saxton, and Andrew Thomason for many stimulating discussions over the course of the last several years. These discussions have had a significant impact on the development of the container method. Last but not least, we would like to again acknowledge David Conlon, Ehud Friedgut, Tim Gowers, Daniel Kleitman, Vojta Rödl, Mathias Schacht, and Kenneth

Wilson, whose work on enumeration of  $C_4$ -free graphs and on extremal and Ramsey properties of random discrete structures inspired and influenced our investigations that led to the hypergraph container theorems.

## References

- M. Ajtai, J. Komlós, J. Pintz, J. Spencer, and E. Szemerédi (1982). “Extremal uncrowded hypergraphs”. *J. Combin. Theory Ser. A* 32, pp. 321–335 (cit. on p. [3096](#)).
- V. E. Alekseev (1992). “Range of values of entropy of hereditary classes of graphs”. *Diskret. Mat.* 4, pp. 148–157 (cit. on p. [3097](#)).
- N. Alon (1993). “Restricted colorings of graphs”. In: *Surveys in Combinatorics, 1993 (Keele)*. Vol. 187. London Math. Soc. Lecture Note Ser. Cambridge Univ. Press, Cambridge, pp. 1–33 (cit. on p. [3103](#)).
- (2000). “Degrees and choice numbers”. *Random Structures Algorithms* 16, pp. 364–368 (cit. on p. [3103](#)).
- (2012). “A non-linear lower bound for planar epsilon-nets”. *Discrete Comput. Geom.* 47, pp. 235–244 (cit. on pp. [3102](#), [3103](#)).
- N. Alon, J. Balogh, R. Morris, and W. Samotij (2014a). “A refinement of the Cameron-Erdős Conjecture”. *Proc. London Math. Soc.* 108, pp. 44–72 (cit. on pp. [3080](#), [3104](#), [3105](#)).
- (2014b). “Counting sum-free sets in abelian groups”. *Israel J. Math.* 199, pp. 309–344 (cit. on pp. [3080](#), [3104](#)).
- J. Balogh, N. Bushaw, M. Collares, H. Liu, R. Morris, and M. Sharifzadeh (2017). “The typical structure of graphs with no large cliques”. *Combinatorica* 37, pp. 617–632 (cit. on p. [3102](#)).
- J. Balogh, H. Liu, Š. Petříčková, and M. Sharifzadeh (2015). “The typical structure of maximal triangle-free graphs”. *Forum Math. Sigma* 3, e20, 19 (cit. on p. [3106](#)).
- J. Balogh, H. Liu, and M. Sharifzadeh (2017). “The number of subsets of integers with no  $k$ -term arithmetic progression”. *Int. Math. Res. Not.* 20, pp. 6168–6186 (cit. on pp. [3092](#), [3104](#)).
- J. Balogh, R. Morris, and W. Samotij (2015). “Independent sets in hypergraphs”. *J. Amer. Math. Soc.* 28, pp. 669–709 (cit. on pp. [3078](#), [3080](#), [3089](#), [3090](#), [3104](#), [3107](#)).
- J. Balogh and Š. Petříčková (2014). “The number of the maximal triangle-free graphs”. *Bull. Lond. Math. Soc.* 46, pp. 1003–1006 (cit. on p. [3106](#)).
- J. Balogh and W. Samotij (2011a). “The number of  $K_{m,m}$ -free graphs”. *Combinatorica* 31, pp. 131–150 (cit. on pp. [3080](#), [3092](#)).
- (2011b). “The number of  $K_{s,t}$ -free graphs”. *J. Lond. Math. Soc.* 83, pp. 368–388 (cit. on pp. [3080](#), [3092](#)).

- J. Balogh and J. Solymosi (n.d.). “On the number of points in general position in the plane”. *Discrete Anal.*, to appear (cit. on pp. [3080](#), [3087](#), [3102](#), [3103](#)).
- J Balogh, H. Liu, M. Sharifzadeh, and A. Treglown (n.d.). “Sharp bound on the number of maximal sum-free subsets of integers”. submitted (cit. on pp. [3104](#), [3106](#)).
- (2015). “The number of maximal sum-free subsets of integers”. *Proc. Amer. Math. Soc.* 143.11, pp. 4713–4721 (cit. on pp. [3104](#), [3106](#)).
- J Balogh and A. Z. Wagner (2016). “Further applications of the container method”. In: *Recent trends in combinatorics*. Vol. 159. IMA Vol. Math. Appl. Springer, Cham, pp. 191–213 (cit. on pp. [3098](#), [3099](#)).
- (2017). “On the number of union-free families”. *Israel J. Math.* 219, pp. 431–448 (cit. on pp. [3106](#), [3107](#)).
- B. Bollobás and A. Thomason (1995). “Projections of bodies and hereditary properties of hypergraphs”. *Bull. London Math. Soc.* 27, pp. 417–424 (cit. on p. [3097](#)).
- (1997). “Hereditary and monotone properties of graphs”. In: *The mathematics of Paul Erdős, II*. Vol. 14. Algorithms Combin. Springer, Berlin, pp. 70–78 (cit. on p. [3097](#)).
- G. Burosch, J. Demetrovics, G. O. H. Katona, D. J. Kleitman, and A. A. Sapozhenko (1991). “On the number of databases and closure operations”. *Theoretical Computer Science* 78, pp. 377–381 (cit. on p. [3106](#)).
- P. Cameron and P. Erdős (1990). *On the number of sets of integers with various properties*. In: *Number Theory* (Mollin, R.A., ed.), 61–79, Walter de Gruyter, Berlin (cit. on p. [3104](#)).
- D. Conlon, D. Dellamonica Jr., S. La Fleur, V. Rödl, and M. Schacht (2017). “A note on induced Ramsey numbers”. In: *A Journey Through Discrete Mathematics: A Tribute to Jiří Matoušek*. Springer, pp. 357–366 (cit. on p. [3102](#)).
- D. Conlon and W. T. Gowers (2016). “Combinatorial theorems in sparse random sets”. *Ann. of Math. (2)* 184, pp. 367–454 (cit. on pp. [3078](#), [3095](#)).
- D. Conlon, W. T. Gowers, W. Samotij, and M. Schacht (2014). “On the KLR conjecture in random graphs”. *Israel J. Math.* 203, pp. 535–580 (cit. on p. [3107](#)).
- D. Dellamonica Jr., Y. Kohayakawa, S. J. Lee, V. Rödl, and W. Samotij (n.d.). “The number of  $B_h$ -sets of a given cardinality”. *Proc. London Math. Soc. (2)*, to appear (cit. on pp. [3096](#), [3104](#)).
- (2016a). “On the number of  $B_h$ -sets”. *Combin. Probab. Comput.* 25, pp. 108–129 (cit. on p. [3104](#)).
  - (2016b). “The number of  $B_3$ -sets of a given cardinality”. *J. Combin. Theory Ser. A* 142, pp. 44–76 (cit. on pp. [3096](#), [3104](#)).
- B. DeMarco and J. Kahn (n.d.). “Turán’s theorem for random graphs”. arXiv: [1501.01340](#) (cit. on p. [3078](#)).
- (2015). “Mantel’s theorem for random graphs”. *Random Structures Algorithms* 47, pp. 59–72 (cit. on p. [3078](#)).

- R. A. Duke, H. Lefmann, and V. Rödl (1995). “On uncrowded hypergraphs”. In: *Proceedings of the Sixth International Seminar on Random Graphs and Probabilistic Methods in Combinatorics and Computer Science*, “Random Graphs ’93” (Poznań, 1993). Vol. 6, pp. 209–212 (cit. on p. [3096](#)).
- P. Erdős (1967). “Some recent results on extremal problems in graph theory. Results”. In: *Theory of Graphs (Internat. Sympos., Rome, 1966)*. Gordon and Breach, New York; Dunod, Paris, 117–123 (English), pp. 124–130 (French) (cit. on p. [3085](#)).
- (1988). “Some old and new problems in combinatorial geometry”. In: *Applications of discrete mathematics (Clemson, SC, 1986)*. SIAM, Philadelphia, PA, pp. 32–37 (cit. on p. [3087](#)).
- P. Erdős, P. Frankl, and V. Rödl (1986). “The asymptotic number of graphs not containing a fixed subgraph and a problem for hypergraphs having no exponent”. *Graphs Combin.* 2, pp. 113–121 (cit. on p. [3092](#)).
- P. Erdős, D. J. Kleitman, and B. L. Rothschild (1976). “Asymptotic enumeration of  $K_n$ -free graphs”. In: *Colloquio Internazionale sulle Teorie Combinatorie (Rome, 1973), Tomo II*. Rome: Accad. Naz. Lincei, 19–27. Atti dei Convegni Lincei, No. 17 (cit. on pp. [3077](#), [3084](#), [3092](#)).
- V. Falgas-Ravry, K. O’Connell, J. Strömberg, and A. Uzzell (n.d.). “Multicolour containers and the entropy of decorated graph limits”. arXiv: [1607.08152](#) (cit. on p. [3098](#)).
- A. Ferber, G. A. McKinley, and W. Samotij (n.d.). “Supersaturated sparse graphs and hypergraphs”. arXiv: [1710.04517](#) (cit. on pp. [3092](#), [3096](#)).
- P. Frankl and V. Rödl (1986). “Large triangle-free subgraphs in graphs without  $K_4$ ”. *Graphs Combin.* 2, pp. 135–144 (cit. on pp. [3078](#), [3083](#), [3086](#)).
- E. Friedgut (1999). “Sharp thresholds of graph properties, and the  $k$ -sat problem”. *J. Amer. Math. Soc.* 12. With an appendix by Jean Bourgain, pp. 1017–1054 (cit. on p. [3105](#)).
- E. Friedgut, H. Hàn, Y. Person, and M. Schacht (2016). “A sharp threshold for van der Waerden’s theorem in random subsets”. *Discrete Anal.* Paper No. 7, 20 (cit. on p. [3105](#)).
- E. Friedgut, V. Rödl, A. Ruciński, and P. Tetali (2006). “A sharp threshold for random graphs with a monochromatic triangle in every edge coloring”. *Mem. Amer. Math. Soc.* 179, pp. vi+66 (cit. on p. [3105](#)).
- E. Friedgut, V. Rödl, and M. Schacht (2010). “Ramsey properties of random discrete structures”. *Random Structures Algorithms* 37, pp. 407–436 (cit. on p. [3078](#)).
- Z. Füredi (1991). “Maximal independent subsets in Steiner systems and in planar sets”. *SIAM J. Discrete Math.* 4, pp. 196–199 (cit. on pp. [3087](#), [3095](#)).
- (2015). “A proof of the stability of extremal graphs, Simonovits’ stability from Szemerédi’s regularity”. *J. Combin. Theory Ser. B* 115, pp. 66–71 (cit. on p. [3085](#)).
- H. Furstenberg and Y. Katznelson (1991). “A density version of the Hales-Jewett theorem”. *J. Anal. Math.* 57, pp. 64–119 (cit. on p. [3087](#)).

- S. Gerke, M. Marciniszyn, and A. Steger (2007). “A probabilistic counting lemma for complete graphs”. *Random Structures Algorithms* 31, pp. 517–534 (cit. on p. 3107).
- W. T. Gowers (2013). *Erdős and Arithmetic Progressions*. Vol. 25. In: Lovász L., Ruzsa I.Z., Sós V.T. (eds.) *Erdős Centennial*, Bolyai Society Mathematical Studies. Springer, Berlin, Heidelberg (cit. on p. 3104).
- B. Green (2004). “The Cameron-Erdős Conjecture”. *Bull. London Math. Soc.* 36, pp. 769–778 (cit. on p. 3104).
- B. Green and I. Z. Ruzsa (2004). “Counting sumsets and sum-free sets modulo a prime”. *Studia Sci. Math. Hungar.* 41, pp. 285–293 (cit. on pp. 3080, 3104).
- D. Haussler and E. Welzl (1987). “ $\epsilon$ -nets and simplex range queries”. *Discrete Comput. Geom.* 2, pp. 127–151 (cit. on p. 3102).
- P. E. Haxell, Y. Kohayakawa, and T. Łuczak (1995). “Turán’s extremal problem in random graphs: forbidding even cycles”. *J. Combin. Theory Ser. B* 64, pp. 273–287 (cit. on pp. 3078, 3095).
- (1996). “Turán’s extremal problem in random graphs: forbidding odd cycles”. *Combinatorica* 16, pp. 107–122 (cit. on p. 3078).
- D. J. Kleitman (1976). “Extremal properties of collections of subsets containing no two sets and their union”. *J. Combin. Theory Ser. A* 20, pp. 390–392 (cit. on p. 3106).
- D. J. Kleitman and K. J. Winston (1980). “The asymptotic number of lattices”. *Ann. Discrete Math.* 6. Combinatorial mathematics, optimal designs and their applications (Proc. Sympos. Combin. Math. and Optimal Design, Colorado State Univ., Fort Collins, Colo., 1978), pp. 243–249 (cit. on p. 3080).
- (1982). “On the number of graphs without 4-cycles”. *Discrete Math.* 41, pp. 167–172 (cit. on pp. 3078, 3080, 3092, 3095).
- Y. Kohayakawa (1997). “Szemerédi’s regularity lemma for sparse graphs”. In: *Foundations of computational mathematics (Rio de Janeiro, 1997)*. Berlin: Springer, pp. 216–230 (cit. on p. 3107).
- Y. Kohayakawa, B. Kreuter, and A. Steger (1998). “An extremal problem for random graphs and the number of graphs with large even-girth”. *Combinatorica* 18, pp. 101–120 (cit. on pp. 3095, 3096).
- Y. Kohayakawa, S. J. Lee, V. Rödl, and W. Samotij (2015). “The number of Sidon sets and the maximum size of Sidon sets contained in a sparse random set of integers”. *Random Structures Algorithms* 46, pp. 1–25 (cit. on p. 3104).
- Y. Kohayakawa, T. Łuczak, and V. Rödl (1997). “On  $K^4$ -free subgraphs of random graphs”. *Combinatorica* 17, pp. 173–213 (cit. on p. 3107).
- G. Kozma, T. Meyerovitch, R. Peled, and W. Samotij (n.d.). “What does a typical metric space look like?” manuscript (cit. on p. 3098).



- D. Kühn, D. Osthus, T. Townsend, and Y. Zhao (2017). “On the structure of oriented graphs and digraphs with forbidden tournaments or cycles”. *J. Combin. Theory Ser. B* 124, pp. 88–127 (cit. on p. 3098).
- T. Łuczak (2000). “On triangle-free random graphs”. *Random Structures Algorithms* 16, pp. 260–276 (cit. on p. 3084).
- W. Mantel (1907). “Problem 28”. *Wiskundige Opgaven* 10, pp. 60–61 (cit. on p. 3083).
- R. Morris, W. Samotij, and D. Saxton (n.d.). “An asymmetric container lemma and the structure of graphs with no induced 4-cycle”. manuscript (cit. on pp. 3089, 3100, 3101).
- R. Morris and D. Saxton (2016). “The number of  $C_{2\ell}$ -free graphs”. *Adv. Math.* 298, pp. 534–580 (cit. on pp. 3092, 3093, 3095–3097).
- F. Mousset, R. Nenadov, and A. Steger (2014). “On the number of graphs without large cliques”. *SIAM J. Discrete Math.* 28, pp. 1980–1986 (cit. on p. 3101).
- D. Mubayi and C. Terry (n.d.). “Discrete metric spaces: structure, enumeration, and 0-1 laws”. *J. Symb. Log.*, to appear (cit. on p. 3098).
- R. Nenadov and A. Steger (2016). “A short proof of the random Ramsey theorem”. *Combin. Probab. Comput.* 25, pp. 130–144 (cit. on p. 3086).
- V. Rödl and A. Ruciński (1993). “Lower bounds on probability thresholds for Ramsey properties”. In: *Combinatorics, Paul Erdős is eighty, Vol. I*. Bolyai Soc. Math. Stud. János Bolyai Math. Soc., Budapest, pp. 317–346 (cit. on p. 3086).
- (1994). “Random graphs with monochromatic triangles in every edge coloring”. *Random Structures Algorithms* 5, pp. 253–270 (cit. on p. 3086).
- (1995). “Threshold functions for Ramsey properties”. *J. Amer. Math. Soc.* 8, pp. 917–942 (cit. on pp. 3086, 3105).
- V. Rödl, A. Ruciński, and M. Schacht (2017). “An exponential-type upper bound for Folkman numbers”. *Combinatorica* 37, pp. 767–784 (cit. on p. 3102).
- V. Rödl and M. Schacht (2013). “Extremal results in random graphs”. In: *Erdős centennial*. Vol. 25. Bolyai Soc. Math. Stud. János Bolyai Math. Soc., Budapest, pp. 535–583 (cit. on p. 3078).
- I. Z. Ruzsa and E. Szemerédi (1978). “Triple systems with no six points carrying three triangles”. In: *Combinatorics (Proc. Fifth Hungarian Colloq., Keszthely, 1976), Vol. II*. Vol. 18. Colloq. Math. Soc. János Bolyai. Amsterdam: North-Holland, pp. 939–945 (cit. on p. 3085).
- W. Samotij (2015). “Counting independent sets in graphs”. *European J. Combin.* 48, pp. 5–18 (cit. on p. 3080).
- A. A. Sapozhenko (2001). “On the number of independent sets in extenders”. *Diskret. Mat.* 13, pp. 56–62 (cit. on pp. 3078, 3080).
- (2003). “The Cameron-Erdős conjecture”. *Dokl. Akad. Nauk* 393, pp. 749–752 (cit. on pp. 3078, 3080, 3104, 3105).

- (2005). “Systems of containers and enumeration problems”. In: *Stochastic Algorithms: Foundations and Applications: Third International Symposium, SAGA 2005*. Vol. 3777. Lecture Notes in Computer Science. Moscow, Russia, pp. 1–13 (cit. on pp. [3078](#), [3080](#)).
- D. Saxton and A. Thomason (2012). “List colourings of regular hypergraphs”. *Combin. Probab. Comput.* 21, pp. 315–322 (cit. on pp. [3080](#), [3103](#)).
- (2015). “Hypergraph containers”. *Invent. Math.* 201, pp. 925–992 (cit. on pp. [3078](#), [3080](#), [3097](#), [3103](#), [3107](#)).
- (2016). “Online containers for hypergraphs, with applications to linear equations”. *J. Combin. Theory Ser. B* 121, pp. 248–283 (cit. on pp. [3103](#), [3104](#)).
- M. Schacht (2016). “Extremal results for random discrete structures”. *Ann. of Math.* (2) 184, pp. 333–365 (cit. on pp. [3078](#), [3095](#)).
- M. Schacht and F. Schulenburg (2018). “Sharp thresholds for Ramsey properties of strictly balanced nearly bipartite graphs”. *Random Structures Algorithms* 52, pp. 3–40 (cit. on p. [3105](#)).
- A. Scott (2011). “Szemerédi’s regularity lemma for matrices and sparse graphs”. *Combin. Probab. Comput.* 20, pp. 455–466 (cit. on p. [3107](#)).
- M. Simonovits (1968). “A method for solving extremal problems in graph theory, stability problems”. In: *Theory of Graphs (Proc. Colloq., Tihany, 1966)*. New York: Academic Press, pp. 279–319 (cit. on p. [3085](#)).
- E. Szemerédi (1975). “On sets of integers containing no  $k$  elements in arithmetic progression”. *Acta Arith.* 27, pp. 199–245 (cit. on p. [3077](#)).
- (1978). “Regular partitions of graphs”. In: *Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976)*. Vol. 260. Colloq. Internat. CNRS. Paris: CNRS, pp. 399–401 (cit. on p. [3107](#)).
- C. Terry (n.d.). “Structure and enumeration theorems for hereditary properties in finite relational languages”. arXiv: [1607.04902](#) (cit. on p. [3098](#)).
- P. Turán (1941). “Eine Extremalaufgabe aus der Graphentheorie”. *Mat. Fiz. Lapok* 48, pp. 436–452 (cit. on p. [3077](#)).

Received 2018-01-24.

JÓZSEF BALOGH

DEPARTMENT OF MATHEMATICS, UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN, URBANA, ILLINOIS 61801, USA  
[jobal@math.uiuc.edu](mailto:jobal@math.uiuc.edu)

ROBERT MORRIS

IMPA, ESTRADA DONA CASTORINA 110, JARDIM BOTÂNICO, RIO DE JANEIRO, 22460-320, BRAZIL  
[rob@impa.br](mailto:rob@impa.br)

WOJCIECH SAMOTIJ

SCHOOL OF MATHEMATICAL SCIENCES, TEL AVIV UNIVERSITY, TEL AVIV 6997801, ISRAEL  
[samotij@post.tau.ac.il](mailto:samotij@post.tau.ac.il)



# COMBINATORIAL APPLICATIONS OF THE HODGE–RIEMANN RELATIONS

JUNE HUH

## Abstract

Why do natural and interesting sequences often turn out to be log-concave? We give one of many possible explanations, from the viewpoint of “standard conjectures”. We illustrate with several examples from combinatorics.

## 1 Log-concave and unimodal sequences

Logarithmic concavity is a property of a sequence of real numbers, occurring throughout algebra, geometry, and combinatorics. A sequence of real numbers  $a_0, \dots, a_d$  is *log-concave* if

$$a_i^2 \geq a_{i-1}a_{i+1} \text{ for all } i.$$

When all the entries are positive, the log-concavity implies unimodality, a property easier to visualize: the sequence is *unimodal* if there is an index  $i$  such that

$$a_0 \leq \dots \leq a_{i-1} \leq a_i \geq a_{i+1} \geq \dots \geq a_d.$$

A rich variety of log-concave and unimodal sequences arising in combinatorics can be found in the surveys [Brenti \[1994\]](#) and [Stanley \[1989, 2000\]](#). For an extensive discussion of log-concavity and its applications in probability and statistics, see [Dharmadhikari and Joag-Dev \[1988\]](#), [Marshall, Olkin, and Arnold \[2011\]](#), and [Saumard and Wellner \[2014\]](#).

Why do natural and interesting sequences often turn out to be log-concave? Below we give one of many possible explanations, from the viewpoint of *standard conjectures*. To illustrate, we discuss three combinatorial sequences appearing in [Stanley \[2000, Problem 25\]](#), in Sections 2.4, 2.5, and 2.8. Another heuristic, based on the physical principle that the entropy of a system should be concave as a function of the energy, can be found in [Okounkov \[2003\]](#).

Let  $X$  be a mathematical object of “dimension”  $d$ . Often it is possible to construct from  $X$  in a natural way a graded vector space over the real numbers

$$A(X) = \bigoplus_{q=0}^d A^q(X),$$

together with a symmetric bilinear map  $P : A(X) \times A(X) \rightarrow \mathbb{R}$  and a graded linear map  $L : A^\bullet(X) \rightarrow A^{\bullet+1}(X)$  that is symmetric with respect to  $P$ . The linear operator  $L$  usually comes in as a member of a family  $K(X)$ , a convex cone in the space of linear operators on  $A(X)$ .<sup>1</sup> For example,  $A(X)$  may be the cohomology of real  $(q, q)$ -forms on a compact Kähler manifold (Gromov [1990]), the ring of algebraic cycles modulo homological equivalence on a smooth projective variety (Grothendieck [1969]), McMullen’s algebra generated by the Minkowski summands of a simple convex polytope (McMullen [1993]), the combinatorial intersection cohomology of a convex polytope (Karu [2004]), the reduced Soergel bimodule of a Coxeter group element (Elias and Williamson [2014]), or the Chow ring of a matroid (Section 2.6).

Often, but not always,  $A(X)$  has the structure of a graded algebra,  $P$  is determined by the multiplicative structure of  $A(X)$  up to a constant multiple, and  $L$  is the multiplication by an element in  $A^1(X)$ . In any case, we expect the following properties to hold for the triple  $(A(X), P(X), K(X))$  for every nonnegative integer  $q \leq \frac{d}{2}$ :

(PD) The bilinear pairing

$$A^q(X) \times A^{d-q}(X) \longrightarrow \mathbb{R}, \quad (\eta, \xi) \longmapsto P(\eta, \xi)$$

is nondegenerate (the Poincaré duality for  $X$ ).

(HL) For any  $L_1, \dots, L_{d-2q} \in K(X)$ , the linear map

$$A^q(X) \longrightarrow A^{d-q}(X), \quad \eta \longmapsto \left( \prod_{i=1}^{d-2q} L_i \right) \eta$$

is bijective (the hard Lefschetz theorem for  $X$ ).

(HR) For any  $L_0, L_1, \dots, L_{d-2q} \in K(X)$ , the bilinear form

$$A^q(X) \times A^q(X) \longrightarrow \mathbb{R}, \quad (\eta_1, \eta_2) \longmapsto (-1)^q P(\eta_1, \left( \prod_{i=1}^{d-2q} L_i \right) \eta_2)$$

---

<sup>1</sup>“P” is for Poincaré, “L” is for Lefschetz, and “K” is for Kähler.

is positive definite on the kernel of the linear map

$$A^q(X) \longrightarrow A^{d-q+1}(X), \quad \eta \longmapsto \left( \prod_{i=0}^{d-2q} L_i \right) \eta$$

(the Hodge-Riemann relation for  $X$ ).

All three properties are known to hold for the objects listed above except one, which is the subject of Grothendieck's standard conjectures on algebraic cycles. The known proofs of the hard Lefschetz theorems and the Hodge-Riemann relations for different objects have certain structural similarities, but there is no known way of deducing one of them from the others.

Hard Lefschetz theorems for various  $X$ 's have found numerous applications to problems of combinatorial nature. An early survey of these applications can be found in [Stanley \[1984\]](#). We highlight the following three:

- (1) Erdős–Moser conjecture ([Erdős \[1965\]](#)), proved by [Stanley \[1980b\]](#): Let  $E$  be a subset of  $\mathbb{R}$  and let  $f(E, k)$  be the number of subsets of  $E$  whose elements sum to  $k$ . If the cardinality of  $E$  is  $2n + 1$ , then

$$f(E, k) \leq f([-n, n] \cap \mathbb{Z}, 0).$$

- (2) McMullen's  $g$ -conjecture ([McMullen \[1971\]](#)), proved by [Billera and Lee \[1980\]](#) and [Stanley \[1980a\]](#): The  $f$ -vector of a  $d$ -dimensional convex polytope  $P$  is the sequence  $f_0(P), \dots, f_d(P)$ , where

$$f_i(P) = \text{the number of } (i-1)\text{-dimensional faces of } P.$$

The  $h$ -vector of  $P$  is the sequence  $h_0(P), \dots, h_d(P)$  defined by the identity

$$\sum_{i=0}^d h_i(P) x^i = \sum_{i=0}^d f_i(P) x^i (1-x)^{d-i}.$$

The  $g$ -conjecture gives a complete numerical characterization of the  $h$ -vectors of simplicial polytopes. In particular, for any  $d$ -dimensional simplicial polytope  $P$ ,

$$h_i(P) = h_{d-i}(P) \text{ and } h_i(P) \leq h_{i+1}(P) \text{ for all } i < d/2.$$

- (3) Dowling-Wilson conjecture ([Dowling and Wilson \[1974, 1975\]](#)), proved by [Huh and Wang \[2017\]](#): Let  $E$  be a finite subset of a vector space, and let  $w_i(E)$  be the number of  $i$ -dimensional subspaces spanned by subsets of  $E$ . If  $E$  spans a  $d$ -dimensional subspace, then

$$w_i(E) \leq w_{d-i}(E) \text{ and } w_i(E) \leq w_{i+1}(E) \text{ for all } i < d/2.$$

All known proofs of the above statements use some version of HL.

When the Poincaré duality for  $X$  is known, the Hodge-Riemann relation for  $X$  is stronger than the hard Lefschetz theorem for  $X$  in the sense that, for every  $q$ ,

$$\text{HR in degrees at most } q \implies \text{HL in degrees at most } q.$$

In the remainder of this survey, we give an overview of applications of the Hodge-Riemann relations to concrete problems. We remark that most known applications only use the following immediate consequence of HR in degrees  $q \leq 1$ : For any  $L_1, \dots, L_{d-2} \in K(X)$ , any matrix representing the symmetric bilinear form

$$A^1(X) \times A^1(X) \longrightarrow \mathbb{R}, \quad (\eta_1, \eta_2) \longmapsto P(\eta_1, \left( \prod_{i=0}^{d-2} L_i \right) \eta_2)$$

has exactly one positive eigenvalue. One notable exception is the implication

Grothendieck standard conjectures on algebraic cycles  $\implies$

Weil conjectures on zeta functions over finite fields,

which was one of the main motivations for formulating the standard conjectures ([Colmez and Serre \[2001\]](#), [Kleiman \[1968\]](#), and [Kleiman \[1994\]](#)). It will be interesting to find applications of HR for  $q > 1$  in other contexts too.

## 2 Applications of the Hodge-Riemann relations

**2.1 Mixed discriminants and permanents.** The notion of mixed discriminant arises when one combines the determinant with the matrix sum. To define the mixed discriminant, let  $\mathbf{A} = (A_1, \dots, A_d)$  be a collection of real symmetric  $d \times d$  matrices, and consider the function

$$\det_{\mathbf{A}} : \mathbb{R}^d \longrightarrow \mathbb{R}, \quad (t_1, \dots, t_d) \longmapsto \det(t_1 A_1 + \dots + t_d A_d),$$

which is a homogeneous polynomial of degree  $d$ . The number

$$D(A_1, \dots, A_d) = \frac{\partial^d}{\partial t_1 \dots \partial t_d} \det_{\mathbf{A}}(t_1, \dots, t_d)$$

is called the *mixed discriminant* of  $\mathbf{A}$ . The mixed discriminant is symmetric in  $\mathbf{A}$ , and it is nonnegative whenever all the matrices in  $\mathbf{A}$  are positive semidefinite.<sup>2</sup>

<sup>2</sup>The latter fact can be viewed as a Hodge-Riemann relation in degree 0.

Let  $\mathbf{P} = (P_1, \dots, P_{d-2})$  be any collection of  $d \times d$  positive semidefinite matrices. Define a symmetric bilinear form  $\text{HR}(\mathbf{P})$  on the space of real symmetric  $d \times d$  matrices by

$$\text{HR}(\mathbf{P}) : \text{Sym}_d \times \text{Sym}_d \longrightarrow \mathbb{R}, \quad (\eta_1, \eta_2) \longmapsto D(\eta_1, \eta_2, P_1, \dots, P_{d-2}).$$

[Aleksandrov \[1938\]](#) proved the following statement and used it in his proof of the Aleksandrov-Fenchel inequality for mixed volumes of convex bodies. To avoid trivialities, we suppose that  $\text{HR}(\mathbf{P})$  is not identically zero.

**Theorem 1.** Any matrix representing  $\text{HR}(\mathbf{P})$  has exactly one positive eigenvalue.

It follows from Cauchy's eigenvalue interlacing theorem that, for any positive semidefinite  $d \times d$  matrices  $A_1, \dots, A_d$ ,

$$\det \begin{pmatrix} D(A_1, A_1, A_3, \dots, A_d) & D(A_1, A_2, A_3, \dots, A_d) \\ D(A_1, A_2, A_3, \dots, A_d) & D(A_2, A_2, A_3, \dots, A_d) \end{pmatrix} \leq 0.$$

[Theorem 1](#) is, in fact, a Hodge-Riemann relation in degree 1. The object  $X$  is the  $d$ -dimensional complex vector space  $\mathbb{C}^d$ , the algebra  $A(X)$  is the ring of real differential forms with constant coefficients on  $\mathbb{C}^d$ , and the cone  $K(X)$  is the spectrahedral cone of all  $d \times d$  positive definite matrices. Elementary proofs of the Hodge-Riemann relation for this  $X$  in any degree can be found in [Gromov \[1990\]](#) and [Timorin \[1998\]](#).

In the important special case when all the matrices are diagonal, the mixed discriminant is a permanent. Precisely, if  $A = (a_{ij})$  is an  $d \times d$  matrix and if  $A_i$  is the diagonal matrix whose  $j$ -th diagonal element is  $a_{ij}$ , then

$$d! D(A_1, \dots, A_d) = \text{per}(A) := \sum_{\sigma} \prod_{i=1}^d a_{i\sigma(i)},$$

where  $\sigma$  runs through all permutations of  $\{1, \dots, d\}$ . Therefore, for any column vectors  $a_1, \dots, a_d$  in  $\mathbb{R}^n$  with nonnegative entries,

$$\text{per}(a_1, a_2, a_3, \dots, a_d)^2 \geq \text{per}(a_1, a_1, a_3, \dots, a_d) \text{per}(a_2, a_2, a_3, \dots, a_d).$$

The above special case of the Hodge-Riemann relations for  $\mathbb{C}^d$  was the main ingredient in Egorychev's and Falikman's proofs of van der Waerden's conjecture that the permanent of any doubly stochastic  $d \times d$  matrix is at least  $d!/d^d$ . See [Knuth \[1981\]](#) and [van Lint \[1982\]](#) for more on van der Waerden's permanent conjecture.



**2.2 Mixed volumes of convex bodies.** The notion of mixed volume arises when one combines the volume with the Minkowski sum. For any collection of convex bodies  $\mathbf{P} = (P_1, \dots, P_d)$  in  $\mathbb{R}^d$ , consider the function

$$\text{vol}_{\mathbf{P}} : \mathbb{R}_{\geq 0}^d \longrightarrow \mathbb{R}_{\geq 0}, \quad (t_1, \dots, t_d) \longmapsto \text{vol}(t_1 P_1 + \dots + t_d P_d).$$

Minkowski noticed that  $\text{vol}_{\mathbf{P}}$  is a homogeneous polynomial of degree  $d$ , and called the number

$$V(P_1, \dots, P_d) = \frac{\partial^d}{\partial t_1 \dots \partial t_d} \text{vol}_{\mathbf{P}}(t_1, \dots, t_d)$$

the *mixed volume* of  $\mathbf{P}$ . The mixed volume is symmetric in  $\mathbf{P}$ , and it is nonnegative for any  $\mathbf{P}$ .<sup>3</sup>

Now let  $\eta_1, \dots, \eta_n$  be another collection of convex bodies in  $\mathbb{R}^d$ , and define an  $n \times n$  matrix  $\text{AF} = (\text{AF}_{ij})$  by

$$\text{AF}_{ij} = V(\eta_i, \eta_j, P_1, \dots, P_{d-2}).$$

If  $\text{AF} \neq 0$ , then the mixed volume analog of [Theorem 1](#) holds.

**Theorem 2.** The matrix  $\text{AF}$  has exactly one positive eigenvalue.

It follows that the mixed volume satisfies the *Aleksandrov-Fenchel inequality*

$$\det \begin{pmatrix} V(P_1, P_1, P_3, \dots, P_d) & V(P_1, P_2, P_3, \dots, P_d) \\ V(P_1, P_2, P_3, \dots, P_d) & V(P_2, P_2, P_3, \dots, P_d) \end{pmatrix} \leq 0.$$

In particular, the sequence of mixed volumes of two convex bodies is log-concave:

$$V(\underbrace{P_1, \dots, P_1}_i, \underbrace{P_2, \dots, P_2}_{d-i})^2 \geq V(\underbrace{P_1, \dots, P_1}_{i-1}, \underbrace{P_2, \dots, P_2}_{d-i+1}) V(\underbrace{P_1, \dots, P_1}_{i+1}, \underbrace{P_2, \dots, P_2}_{d-i-1}).$$

Aleksandrov reduced [Theorem 2](#) to the case when the Minkowski sum of all the relevant convex bodies, say  $\Delta$ , is a simple convex polytope. Under this hypothesis, [Theorem 2](#) is a Hodge-Riemann relation in degree 1 ([Gromov \[1990\]](#), [McMullen \[1993\]](#), and [Teissier \[1979\]](#)). The object  $X$  is the convex polytope  $\Delta$ , the algebra  $A(X)$  is McMullen's polytope algebra generated by the Minkowski summands of  $\Delta$ , and the cone  $K(X)$  is the cone of convex polytopes that share the normal fan with  $\Delta$ . Elementary proofs of the Hodge-Riemann relation for this  $X$  in any degree can be found in [Fleming and Karu \[2010\]](#), [McMullen \[1993\]](#), and [Timorin \[1999\]](#).

<sup>3</sup>The latter fact can be viewed as a Hodge-Riemann relation in degree 0.

The Alexandrov-Fenchel inequality has been used to understand linear extensions of partially ordered sets. For example, [Chung, Fishburn, and Graham \[1980\]](#) conjectured that, for any finite poset  $Q$ ,

$$\Pr_i(x)^2 \geq \Pr_{i-1}(x)\Pr_{i+1}(x) \text{ for all } i \text{ and all } x \in Q,$$

where  $\Pr_i(x)$  is the fraction of linear extensions of  $Q$  in which  $x$  is the  $i$ -th largest element. [Stanley \[1981\]](#) proved the conjecture by constructing suitable convex polytopes from  $x \in Q$  and using the Alexandrov-Fenchel inequality. For another example, write  $\Pr(x_1 < x_2)$  for the fraction of linear extensions of  $Q$  in which  $x_1$  is smaller than  $x_2$ . [Kahn and Saks \[1984\]](#) employed Stanley's method to deduce the following remarkable fact from the Alexandrov-Fenchel inequality:

If  $Q$  is not a chain, then there are elements  $x_1, x_2 \in Q$  such that

$$3/11 < \Pr(x_1 < x_2) < 8/11.$$

This confirmed a conjecture of [Fredman \[1975/76\]](#) and [Linial \[1984\]](#) that the information theoretic lower bound for the general sorting problem is tight up to a multiplicative constant.

**2.3 The correlation constant of a field.** Let  $G$  be a finite connected graph, let  $i, j$  be distinct edges, and let  $T$  be a random spanning tree of  $G$ . Kirchhoff's effective resistance formula can be used to show that the probability that  $i$  is in  $T$  can only decrease by assuming that  $j$  is in  $T$ :

$$\Pr(i \in T) \geq \Pr(i \in T \mid j \in T).$$

In other words, the number  $b_-$  of spanning trees containing given edges satisfies

$$\frac{b_i}{b} \geq \frac{b_{ij}}{b_j}.$$

Now let  $M$  be a finite spanning subset of a vector space  $V$ , let  $i, j$  be distinct nonzero vectors in  $M$ , and write  $b_-$  for the number of bases in  $M$  containing given vectors. Do we still have the negative correlation

$$\frac{b_i}{b} \geq \frac{b_{ij}}{b_j}?$$

The previous statement on graphs is the special case when  $M$  is the vertex-edge incidence matrix over the field with two elements.

[Seymour and Welsh \[1975\]](#) gave the first example of  $M$  over a field of characteristic 2 with  $\frac{b_i b_{ij}}{b_j b} = \frac{36}{35}$  for some  $i$  and  $j$ . How large can the ratio be?

**Definition 3.** The *correlation constant* of a field  $k$  is the supremum of  $\frac{b_{ij}}{b_i b_j}$  over all pairs of distinct vectors  $i$  and  $j$  in finite vector configurations in vector spaces over  $k$ .

The correlation constant may be an interesting invariant of a field, although it is not immediately clear that the constant is finite. In fact, the finiteness of the correlation constant is one of the consequences of the Hodge-Riemann relations for vector configurations. Let  $n$  be the number of vectors in  $M$ , and let  $\text{HR}(M)$  be the symmetric  $n \times n$  matrix

$$\text{HR}(M)_{ij} = \begin{cases} 0 & \text{if } i = j, \\ b_{ij} & \text{if } i \neq j. \end{cases}$$

To avoid the trivial case  $\text{HR}(M) = 0$ , we suppose that the dimension of  $V$  is at least 2. For example, if  $K_4$  is the set of six column vectors of the matrix

$$\begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ -1 & 0 & 0 & 1 & 1 & 0 \\ 0 & -1 & 0 & -1 & 0 & 1 \\ 0 & 0 & -1 & 0 & -1 & -1 \end{pmatrix},$$

then  $\text{HR}(K_4)$  is the  $6 \times 6$  symmetric matrix

$$\begin{pmatrix} 0 & 3 & 3 & 3 & 3 & 4 \\ 3 & 0 & 3 & 3 & 4 & 3 \\ 3 & 3 & 0 & 4 & 3 & 3 \\ 3 & 3 & 4 & 0 & 3 & 3 \\ 3 & 4 & 3 & 3 & 0 & 3 \\ 4 & 3 & 3 & 3 & 3 & 0 \end{pmatrix}.$$

In [Huh and Wang \[2017\]](#), the following statement was deduced from [Theorem 12](#).

**Theorem 4.** The matrix  $\text{HR}(M)$  has exactly one positive eigenvalue.

In fact, the same statement holds more generally for any matroid  $M$  ([Huh and Wang \[ibid., Remark 15\]](#)). To deduce a bound on the correlation constant, consider the restriction of  $\text{HR}(M)$  to the three-dimensional subspace of  $\mathbb{R}^n$  spanned by  $\mathbf{e}_i$ ,  $\mathbf{e}_j$ , and  $(1, \dots, 1)$ . Cauchy's eigenvalue interlacing theorem shows that the resulting  $3 \times 3$  symmetric matrix also has exactly one positive eigenvalue. Expressing the  $3 \times 3$  determinant, which should be nonnegative, we get the inequality

$$\frac{b_{ij}}{b_i b_j} \leq 2 - 2(\dim V)^{-1}.$$

Thus the correlation constant of any field is at most 2. What is the correlation constant of, say,  $\mathbb{Z}/2\mathbb{Z}$ ? Does the correlation constant depend on the field?

**2.4 The chromatic polynomial of a graph.** Generalizing earlier work of Birkhoff, [Whitney \[1932\]](#) introduced the *chromatic polynomial* of a connected graph  $G$  as the function on  $\mathbb{N}$  defined by

$$\chi_G(q) = \text{the number of proper } q\text{-colorings of } G.$$

In other words,  $\chi_G(q)$  is the number of ways to color the vertices of  $G$  using  $q$  colors so that the endpoints of every edge have different colors. Whitney noticed that the chromatic polynomial is indeed a polynomial. In fact, we can write

$$\chi_G(q)/q = a_0(G)q^d - a_1(G)q^{d-1} + \cdots + (-1)^d a_d(G)$$

for some positive integers  $a_0(G), \dots, a_d(G)$ , where  $d$  is one less than the number of vertices.

*Example 5.* The cycle  $C_4$  with 4 vertices and 4 edges has the chromatic polynomial

$$\chi_{C_4}(q) = 1q^4 - 4q^3 + 6q^2 - 3q.$$

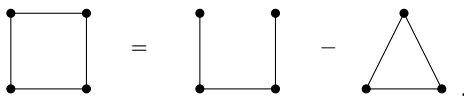
The chromatic polynomial was originally devised as a tool for attacking the Four Color Problem, but soon it attracted attention in its own right. [Read \[1968\]](#) conjectured that the coefficients of the chromatic polynomial form a unimodal sequence for any graph. A few years later, [Hoggar \[1974\]](#) conjectured more generally that the coefficients form a log-concave sequence:

$$a_i(G)^2 \geq a_{i-1}(G)a_{i+1}(G) \text{ for any } i \text{ and } G.$$

Notice that the chromatic polynomial can be computed using the *deletion-contraction relation*: if  $G \setminus e$  is the deletion of an edge  $e$  from  $G$  and  $G/e$  is the contraction of the same edge, then

$$\chi_G(q) = \chi_{G \setminus e}(q) - \chi_{G/e}(q).$$

The first term counts the proper colorings of  $G$ , the second term counts the otherwise-proper colorings of  $G$  where the endpoints of  $e$  are permitted to have the same color, and the third term counts the otherwise-proper colorings of  $G$  where the endpoints of  $e$  are mandated to have the same color. For example, to compute the chromatic polynomial of the cycle  $C_4$  in [Example 5](#), we write

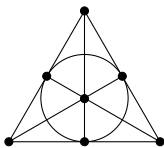


and use that the chromatic polynomials of the two smaller graphs are  $q(q-1)^3$  and  $q(q-1)(q-2)$ , respectively. Note that, in general, the sum of log-concave sequences need not be log-concave.

The log-concavity conjecture for chromatic polynomials was proved in [Huh \[2012\]](#) by showing that the absolute values of the coefficients of  $\chi_G(q)/(q-1)$  are mixed multiplicities of certain homogeneous ideals constructed from  $G$ . The notion of mixed multiplicities is a commutative algebraic analog of the notion of mixed volumes, and it can be shown that mixed multiplicities of homogeneous ideals satisfy a version of the Aleksandrov-Fenchel inequality. To formulate the underlying Hodge-Riemann relation in purely combinatorial terms was the primary motivation for [Adiprasito, Huh, and Katz \[2015\]](#). The main result of [Adiprasito, Huh, and Katz \[ibid.\]](#) will be reviewed in [Section 2.6](#) below.

**2.5 Counting independent sets.** How many linearly independent collection of  $i$  vectors are there in a given configuration of vectors? Let's write  $M$  for a finite subset of a vector space and  $f_i(M)$  for the number of independent subsets of  $M$  of size  $i$ .

*Example 6.* Let  $F$  be the set of all nonzero vectors in the three-dimensional vector space over the field with two elements. Nontrivial dependencies between elements of  $F$  can be read off from the picture of the Fano plane shown below.



The nonempty independent subsets of  $F$  correspond to the seven points in  $F$ , the twenty-one pairs of points in  $F$ , and the twenty-eight triple of points in  $F$  not in one of the seven lines:

$$f_0(F) = 1, \quad f_1(F) = 7, \quad f_2(F) = 21, \quad f_3(F) = 28.$$

[Welsh \[1971\]](#) conjectured that the sequence  $f_i(M)$  is unimodal for any  $M$ . Shortly after, [Mason \[1972\]](#) conjectured more generally that the sequence is log-concave:

$$f_i(M)^2 \geq f_{i-1}(M)f_{i+1}(M) \text{ for any } i \text{ and } M.$$

In any small specific case, the conjecture can be verified by computing the  $f_i(M)$ 's by the *deletion-contraction relation*: if  $M \setminus v$  is the deletion of a nonzero vector  $v$  from  $M$  and  $M/v$  is the projection of  $M$  in the direction of  $v$ , then

$$f_i(M) = f_i(M \setminus v) + f_{i-1}(M/v).$$

The first term counts the number of independent subsets of size  $i$ , the second term counts the independent subsets of size  $i$  not containing  $v$ , and the third term counts the independent subsets of size  $i$  containing  $v$ . As in the case of graphs, we notice the apparent conflict between the log-concavity conjecture and the additive nature of  $f_i(M)$ .

The log-concavity conjecture for  $f_i(M)$  was proved in Lenz [2013] by combining a geometric construction of Huh and Katz [2012] and a matroid-theoretic construction of Brylawski [1977]. Given a spanning subset  $M$  of a  $d$ -dimensional vector space over a field  $k$ , one can construct a  $d$ -dimensional smooth projective variety  $X(M)$  over  $k$  and globally generated line bundles  $L_1, L_2$  on  $X(M)$  so that

$$f_i(M) = \int_{X(M)} L_1^{d-i} L_2^i.$$

The Hodge-Riemann relation for smooth projective varieties is known to hold in degrees  $q \leq 1$  (Grothendieck [1958] and Segre [1937]), and this implies the log-concavity of  $f_i(M)$  as in Sections 2.1, 2.2. To express and verify the general Hodge-Riemann relation for  $X(M)$  in purely combinatorial terms was another motivation for Adiprasito, Huh, and Katz [2015].

**2.6 The Hodge-Riemann relations for matroids.** In the 1930s, Hassler Whitney observed that several notions in graph theory and linear algebra fit together in a common framework, that of *matroids* (Whitney [1935]). This observation started a new subject with applications to a wide range of topics like characteristic classes, optimization, and moduli spaces.

**Definition 7.** A *matroid*  $M$  on a finite set  $E$  is a collection of subsets of  $E$ , called *flats* of  $M$ , satisfying the following axioms:

- (1) The ground set  $E$  is a flat.
- (2) If  $F_1$  and  $F_2$  are flats, then  $F_1 \cap F_2$  is a flat.
- (3) If  $F$  is a flat, then any element not in  $F$  is contained in exactly one flat covering  $F$ .

Here, a flat is said to *cover* another flat  $F$  if it is minimal among the flats properly containing  $F$ .

For our purposes, we may and will suppose that  $M$  is *loopless*:

- (4) The empty subset of  $E$  is a flat.

Every maximal chain of flats in  $F$  has the same length, and this common length is called the *rank* of the flat  $F$ . The rank of the flat  $E$  is called the rank of the matroid  $M$ . Matroids are determined by their *independent sets* (the idea of “*general position*”), and can be alternatively defined in terms of independent sets (Oxley [2011, Chapter 1]).

*Example 8.* Let  $E$  be the set of edges of a finite graph  $G$ . Call a subset  $F$  of  $E$  a flat when there is no edge in  $E \setminus F$  whose endpoints are connected by a path in  $F$ . This defines a *graphic matroid* on  $E$ .

*Example 9.* A *projective space*  $\mathbb{P}$  is a set with distinguished subsets, called *lines*, satisfying:

- (1) Every line contains more than two points.
- (2) If  $x, y$  are distinct points, then there is exactly one line  $xy$  containing  $x$  and  $y$ .
- (3) If  $x, y, z, w$  are distinct points, no three collinear, then

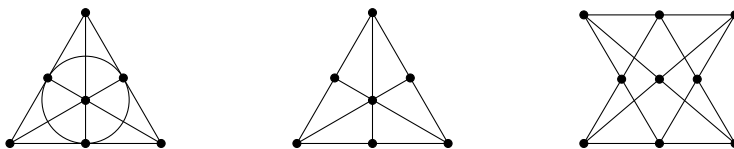
the line  $xy$  intersects the line  $zw \implies$  the line  $xz$  intersects the line  $yw$ .

A subspace of  $\mathbb{P}$  is a subset  $\mathbb{S}$  of  $\mathbb{P}$  such that

$x$  and  $y$  are distinct points in  $\mathbb{S} \implies$  the line  $xy$  is in  $\mathbb{S}$ .

For any finite subset  $E$  of  $\mathbb{P}$ , the collection of sets of the form  $E \cap \mathbb{S}$  has the structure of a matroid. Matroids arising from subsets of projective spaces over a field  $k$  are said to be *realizable* over  $k$  (the idea of “coordinates”).

Not surprisingly, the notion of realizability is sensitive to the field  $k$ . A matroid may arise from a vector configuration over one field while no such vector configuration exists over another field.



Among the rank 3 matroids pictured above, where rank 1 flats are represented by points and rank 2 flats containing more than 2 points are represented by lines, the first is realizable over  $k$  if and only if the characteristic of  $k$  is 2, the second is realizable over  $k$  if and only if the characteristic of  $k$  is not 2, and the third is not realizable over any field. Recently, [Nelson \[2016\]](#) showed that almost all matroids are not realizable over any field [Nelson \[ibid.\]](#).

**Definition 10.** We introduce variables  $x_F$ , one for each nonempty proper flat  $F$  of  $M$ , and consider the polynomial ring

$$S(M) = \mathbb{R}[x_F]_{F \neq \emptyset, F \neq E}.$$

The *Chow ring*  $A(M)$  of  $M$  is the quotient of  $S(M)$  by the ideal generated by the linear forms

$$\sum_{i_1 \in F} x_F - \sum_{i_2 \in F} x_F,$$

one for each pair of distinct elements  $i_1$  and  $i_2$  of  $E$ , and the quadratic monomials

$$x_{F_1} x_{F_2},$$

one for each pair of incomparable nonempty proper flats  $F_1$  and  $F_2$  of  $M$ . We have

$$A(M) = \bigoplus_q A^q(M),$$

where  $A^q(M)$  is the span of degree  $q$  monomials in  $A(M)$ .

Feichtner and Yuzvinsky introduced the Chow ring of  $M$  ([Feichtner and Yuzvinsky \[2004\]](#)). When  $M$  is realizable over a field  $k$ , it is the Chow ring of the “wonderful” compactification of the complement of a hyperplane arrangement defined over  $k$  studied by [De Concini and Procesi \[1995\]](#).

To formulate the hard Lefschetz theorem and the Hodge-Riemann relations for  $A(M)$ , we define a matroid analog of the Kähler cone in complex geometry.

**Definition 11.** A real-valued function  $c$  on  $2^E$  is said to be *strictly submodular* if

$$c_\emptyset = 0, \quad c_E = 0,$$

and, for any two incomparable subsets  $I_1, I_2 \subseteq E$ ,

$$c_{I_1} + c_{I_2} > c_{I_1 \cap I_2} + c_{I_1 \cup I_2}.$$

We note that strictly submodular functions exist. For example,

$$I \mapsto |I||E \setminus I|$$

is a strictly submodular function. A strictly submodular function  $c$  defines an element

$$L_c = \sum_F c_F x_F \in A^1(M).$$

The Kähler cone  $K(M)$  is defined to be the set of all such elements in  $A^1(M)$ .

Now let  $d + 1$  be the rank of  $M$ , and write “deg” for the unique linear isomorphism

$$\deg: A^d(M) \longrightarrow \mathbb{R}$$



which maps  $x_{F_1} \cdots x_{F_d}$  to 1 for every maximal chain  $F_1 \subsetneq \cdots \subsetneq F_d$  of nonempty proper flats (Adiprasito, Huh, and Katz [2015, Proposition 5.10]). We are ready to state the hard Lefschetz theorem and the Hodge-Riemann relation for  $M$  (Adiprasito, Huh, and Katz [ibid., Theorem 8.9]).

**Theorem 12.** Let  $q$  be a nonnegative integer  $\leq \frac{d}{2}$ , and let  $L_0, L_1, \dots, L_{d-2q} \in K(M)$ .

(PD) The product in  $A(M)$  defines a nondegenerate bilinear pairing

$$A^q(M) \times A^{d-q}(M) \longrightarrow \mathbb{R}, \quad (\eta, \xi) \longmapsto \deg(\eta \xi).$$

(HL) The multiplication by  $L_1, \dots, L_{d-2q}$  defines a linear bijection

$$A^q(M) \longrightarrow A^{d-q}(M), \quad \eta \longmapsto \left( \prod_{i=1}^{d-2q} L_i \right) \eta.$$

(HR) The symmetric bilinear form

$$A^q(M) \times A^q(M) \longrightarrow \mathbb{R}, \quad (\eta_1, \eta_2) \longmapsto (-1)^q \deg\left(\left( \prod_{i=1}^{d-2q} L_i \right) \eta_1 \eta_2\right)$$

is positive definite on the kernel of the multiplication map

$$A^q(M) \longrightarrow A^{d-q+1}(M), \quad \eta \longmapsto \left( \prod_{i=0}^{d-2q} L_i \right) \eta.$$

We highlight the following consequence of HR in degrees  $\leq 1$ : For any  $\xi_1, \xi_2 \in K(M)$ ,

$$\begin{pmatrix} \deg\left(\left(\prod_{i=1}^{d-2} L_i\right) \xi_1 \xi_1\right) & \deg\left(\left(\prod_{i=1}^{d-2} L_i\right) \xi_1 \xi_2\right) \\ \deg\left(\left(\prod_{i=1}^{d-2} L_i\right) \xi_1 \xi_2\right) & \deg\left(\left(\prod_{i=1}^{d-2} L_i\right) \xi_2 \xi_2\right) \end{pmatrix}$$

has exactly one positive eigenvalue. Taking the determinant, we get an analog of the Alexandrov-Fenchel inequality

$$\deg\left(\left(\prod_{i=1}^{d-2} L_i\right) \xi_1 \xi_2\right)^2 \geq \deg\left(\left(\prod_{i=1}^{d-2} L_i\right) \xi_1 \xi_1\right) \deg\left(\left(\prod_{i=1}^{d-2} L_i\right) \xi_2 \xi_2\right).$$

We apply the inequality to the *characteristic polynomial*  $\chi_M(q)$ , a generalization of the chromatic polynomial  $\chi_G(q)$  to a matroid  $M$  that is not necessarily graphic (Welsh [1976,

Chapter 15]). For this, we consider two distinguished elements of  $A^1(M)$ . For fixed  $j \in E$ , the elements are

$$\alpha = \sum_{j \in F} x_F, \quad \beta = \sum_{j \notin F} x_F.$$

The two elements do not depend on the choice of  $j$ , and they are limits of elements of the form  $\ell_c$  for a strictly submodular function  $c$ . A bijective counting argument in [Adiprasito, Huh, and Katz \[2015\]](#) shows that

$$e_i(M) = \deg(\alpha^i \beta^{d-i}) \text{ for every } i,$$

where  $e_i(M)$  is the sequence of integers satisfying the identity

$$\chi_M(q)/(q-1) = e_0(M)q^d - e_1(M)q^{d-1} + \cdots + (-1)^d e_d(M).$$

Thus the sequence  $e_i(M)$  is log-concave, which implies the following conjecture of [Heron \[1972\]](#), [Rota \[1971\]](#), and [Welsh \[1976\]](#):

The coefficients of  $\chi_M(q)$  form a log-concave sequence for any matroid  $M$ .

The above implies the log-concavity of the sequence  $a_i(G)$  in [Section 2.4](#) and the log-concavity of the sequence  $f_i(M)$  in [Section 2.5](#). See [Oxley \[2011, Chapter 15\]](#) and [White \[1987, Chapter 8\]](#) for overviews and historical accounts.

**2.7 The reliability polynomial of a network.** Let  $G$  be a finite connected graph with  $v$  vertices and  $n$  edges. The *reliability* of  $G$  is the probability that any two vertices remain connected when each edge is independently removed with the same probability  $q$ . Let's write  $o_i(G)$  for the number of  $i$ -edge operational states. For example,  $o_{v-1}(G)$  is the number of spanning trees and  $o_{n-1}(G)$  is the number of non-bridges. Thus the reliability of  $G$  is

$$R_G(q) = \sum_i o_i(G)(1-q)^i q^{n-i}.$$

We define a sequence of integers  $h_0(G), \dots, h_d(G)$  by the identity

$$R_G(q)/(1-q)^{v-1} = h_d(G)q^d + h_{d-1}(G)q^{d-1} + \cdots + h_0(G),$$

where  $d$  is one more than the difference  $n - v$ .

*Example 13.* The complete graph on 4 vertices has the reliability polynomial

$$\begin{aligned} R_{K_4}(q) &= 16q^3(1-q)^3 + 15q^2(1-q)^4 + 6q(1-q)^5 + 1(1-q)^6 \\ &= (1-q)^3(6q^3 + 6q^2 + 3q + 1). \end{aligned}$$

The numbers  $h_i$  are closely related to the numbers  $f_i$  of independent sets in [Section 2.5](#). Writing  $M$  for the dual of the graphic matroid of  $G$ , we have

$$\sum_{i=0}^d h_i(G)x^i = \sum_{i=0}^d f_i(M)x^i(1-x)^{d-i} = \sum_{i=0}^d h_i(M)x^i.$$

[Dawson \[1984\]](#) conjectured that the sequence  $h_i(M)$  defined by the second equality is log-concave for any matroid  $M$ :

$$h_i(M)^2 \geq h_{i-1}(M)h_{i+1}(M) \text{ for any } i \text{ and } M.$$

[Colbourn \[1987\]](#) independently conjectured the same in the context of reliability polynomials.

When  $M$  is the dual of a graphic matroid, or more generally when  $M$  is realizable over the complex numbers, the log-concavity conjecture for  $h_i(M)$  was proved in [Huh \[2015\]](#) by applying an algebraic analog of the Alexandrov-Fenchel inequality to the variety of critical points of the master function of a realization of  $M$  studied by [Denham, Garrousian, and Schulze \[2012\]](#). The underlying combinatorial Hodge-Riemann relation is yet to be formulated, and Dawson's conjecture for general matroids remains open. The argument in the complex realizable case is tightly connected to the geometry of characteristic cycles ([Huh \[2013\]](#)), suggesting that the combinatorial Hodge-Riemann relation in this context will be strictly stronger than that of [Section 2.6](#).

**2.8 Unsolved problems.** The log-concavity of a sequence is not only important because of its applications but because it hints the existence of a structure that satisfies PD, HL, and HR. We close by listing some of the most interesting sequences that are conjectured to be log-concave.

- (1) Rota's unimodality conjecture ([Rota \[1971\]](#)): If  $w_k(M)$  is the number of rank  $k$  flats of a rank  $d$  matroid  $M$ , then the sequence  $w_0(M), \dots, w_d(M)$  is unimodal. [Welsh \[1976\]](#) conjectured more generally that the sequence is log-concave.
- (2) Fox's trapezoidal conjecture ([Fox \[1962\]](#)): The sequence of absolute values of the coefficients of the Alexander polynomial of an alternating knot strictly increases, possibly plateaus, then strictly decreases. [Stoimenow \[2005\]](#) conjectured more generally that the sequence is log-concave.
- (3) Kazhdan-Lusztig polynomials of matroids ([Elias, Proudfoot, and Wakefield \[2016\]](#)): For any matroid  $M$ , the coefficients of the Kazhdan-Lusztig polynomial of  $M$  form a nonnegative log-concave sequence.

## References

- Karim Adiprasito, June Huh, and Eric Katz (Nov. 2015). “Hodge Theory for Combinatorial Geometries”. arXiv: [1511.02888](#) (cit. on pp. [3120](#), [3121](#), [3124](#), [3125](#)).
- Alexander D Aleksandrov (1938). “Zur Theorie der gemischten Volumina von konvexen Körpern, IV. Die gemischten Diskriminanten und die gemischten Volumina”. *Mat. Sbornik* 3.45, pp. 227–251 (cit. on p. [3115](#)).
- Louis J. Billera and Carl W. Lee (1980). “Sufficiency of McMullen’s conditions for  $f$ -vectors of simplicial polytopes”. *Bull. Amer. Math. Soc. (N.S.)* 2.1, pp. 181–185. MR: [551759](#) (cit. on p. [3113](#)).
- Francesco Brenti (1994). “Log-concave and unimodal sequences in algebra, combinatorics, and geometry: an update”. In: *Jerusalem combinatorics ’93*. Vol. 178. Contemp. Math. Amer. Math. Soc., Providence, RI, pp. 71–89. MR: [1310575](#) (cit. on p. [3111](#)).
- Tom Brylawski (1977). “The broken-circuit complex”. *Trans. Amer. Math. Soc.* 234.2, pp. 417–433. MR: [468931](#) (cit. on p. [3121](#)).
- F. R. K. Chung, P. C. Fishburn, and R. L. Graham (1980). “On unimodality for linear extensions of partial orders”. *SIAM J. Algebraic Discrete Methods* 1.4, pp. 405–410. MR: [593850](#) (cit. on p. [3117](#)).
- Charles J. Colbourn (1987). *The combinatorics of network reliability*. International Series of Monographs on Computer Science. The Clarendon Press, Oxford University Press, New York, pp. xii+160. MR: [902584](#) (cit. on p. [3126](#)).
- Pierre Colmez and Jean-Pierre Serre, eds. (2001). *Correspondance Grothendieck-Serre*. Vol. 2. Documents Mathématiques (Paris) [Mathematical Documents (Paris)]. Société Mathématique de France, Paris, pp. xii+288. MR: [1942134](#) (cit. on p. [3114](#)).
- Jeremy E. Dawson (1984). “A collection of sets related to the Tutte polynomial of a matroid”. In: *Graph theory, Singapore 1983*. Vol. 1073. Lecture Notes in Math. Springer, Berlin, pp. 193–204. MR: [761018](#) (cit. on p. [3126](#)).
- C. De Concini and C. Procesi (1995). “Wonderful models of subspace arrangements”. *Selecta Math. (N.S.)* 1.3, pp. 459–494. MR: [1366622](#) (cit. on p. [3123](#)).
- Graham Denham, Mehdi Garrousian, and Mathias Schulze (2012). “A geometric deletion-restriction formula”. *Adv. Math.* 230.4-6, pp. 1979–1994. MR: [2927361](#) (cit. on p. [3126](#)).
- Sudhakar Dharmadhikari and Kumar Joag-Dev (1988). *Unimodality, convexity, and applications*. Probability and Mathematical Statistics. Academic Press, Inc., Boston, MA, pp. xiv+278. MR: [954608](#) (cit. on p. [3111](#)).
- Thomas A. Dowling and Richard M. Wilson (1974). “The slimmest geometric lattices”. *Trans. Amer. Math. Soc.* 196, pp. 203–215. MR: [0345849](#) (cit. on p. [3113](#)).
- (1975). “Whitney number inequalities for geometric lattices”. *Proc. Amer. Math. Soc.* 47, pp. 504–512. MR: [0354422](#) (cit. on p. [3113](#)).

- Ben Elias, Nicholas Proudfoot, and Max Wakefield (2016). “[The Kazhdan-Lusztig polynomial of a matroid](#)”. *Adv. Math.* 299, pp. 36–70. MR: [3519463](#) (cit. on p. [3126](#)).
- Ben Elias and Geordie Williamson (2014). “[The Hodge theory of Soergel bimodules](#)”. *Ann. of Math.* (2) 180.3, pp. 1089–1136. MR: [3245013](#) (cit. on p. [3112](#)).
- P. Erdős (1965). “Extremal problems in number theory”. In: *Proc. Sympos. Pure Math., Vol. VIII*. Amer. Math. Soc., Providence, R.I., pp. 181–189. MR: [0174539](#) (cit. on p. [3113](#)).
- Eva Maria Feichtner and Sergey Yuzvinsky (2004). “[Chow rings of toric varieties defined by atomic lattices](#)”. *Invent. Math.* 155.3, pp. 515–536. MR: [2038195](#) (cit. on p. [3123](#)).
- Balin Fleming and Kalle Karu (2010). “[Hard Lefschetz theorem for simple polytopes](#)”. *J. Algebraic Combin.* 32.2, pp. 227–239. MR: [2661416](#) (cit. on p. [3116](#)).
- R. H. Fox (1962). “Some problems in knot theory”. In: *Topology of 3-manifolds and related topics (Proc. The Univ. of Georgia Institute, 1961)*. Prentice-Hall, Englewood Cliffs, N.J., pp. 168–176. MR: [0140100](#) (cit. on p. [3126](#)).
- Michael L. Fredman (1975/76). “[How good is the information theory bound in sorting?](#)” *Theoret. Comput. Sci.* 1.4, pp. 355–361. MR: [0416100](#) (cit. on p. [3117](#)).
- M. Gromov (1990). “Convex sets and Kähler manifolds”. In: *Advances in differential geometry and topology*. World Sci. Publ., Teaneck, NJ, pp. 1–38. MR: [1095529](#) (cit. on pp. [3112](#), [3115](#), [3116](#)).
- A. Grothendieck (1958). “[Sur une note de Mattuck-Tate](#)”. *J. Reine Angew. Math.* 200, pp. 208–215. MR: [0136607](#) (cit. on p. [3121](#)).
- (1969). “Standard conjectures on algebraic cycles”. In: *Algebraic Geometry (Internat. Colloq., Tata Inst. Fund. Res., Bombay, 1968)*. Oxford Univ. Press, London, pp. 193–199. MR: [0268189](#) (cit. on p. [3112](#)).
- A. P. Heron (1972). “Matroid polynomials”, pp. 164–202. MR: [0340058](#) (cit. on p. [3125](#)).
- S. G. Hoggar (1974). “Chromatic polynomials and logarithmic concavity”. *J. Combinatorial Theory Ser. B* 16, pp. 248–254. MR: [0342424](#) (cit. on p. [3119](#)).
- June Huh (2012). “[Milnor numbers of projective hypersurfaces and the chromatic polynomial of graphs](#)”. *J. Amer. Math. Soc.* 25.3, pp. 907–927. MR: [2904577](#) (cit. on p. [3120](#)).
- (2013). “[The maximum likelihood degree of a very affine variety](#)”. *Compos. Math.* 149.8, pp. 1245–1266. MR: [3103064](#) (cit. on p. [3126](#)).
- (2015). “[h-vectors of matroids and logarithmic concavity](#)”. *Adv. Math.* 270, pp. 49–59. MR: [3286530](#) (cit. on p. [3126](#)).
- June Huh and Eric Katz (2012). “[Log-concavity of characteristic polynomials and the Bergman fan of matroids](#)”. *Math. Ann.* 354.3, pp. 1103–1116. MR: [2983081](#) (cit. on p. [3121](#)).
- June Huh and Botong Wang (2017). “[Enumeration of points, lines, planes, etc](#)”. *Acta Math.* 218.2, pp. 297–317. MR: [3733101](#) (cit. on pp. [3113](#), [3118](#)).
- Jeff Kahn and Michael Saks (1984). “[Balancing poset extensions](#)”. *Order* 1.2, pp. 113–126. MR: [764319](#) (cit. on p. [3117](#)).

- Kalle Karu (2004). “[Hard Lefschetz theorem for nonrational polytopes](#)”. *Invent. Math.* 157.2, pp. 419–447. MR: [2076929](#) (cit. on p. [3112](#)).
- S. L. Kleiman (1968). “Algebraic cycles and the Weil conjectures”. In: *Dix exposés sur la cohomologie des schémas*. Vol. 3. Adv. Stud. Pure Math. North-Holland, Amsterdam, pp. 359–386. MR: [292838](#) (cit. on p. [3114](#)).
- Steven L. Kleiman (1994). “[The standard conjectures](#)”. In: *Motives (Seattle, WA, 1991)*. Vol. 55. Proc. Sympos. Pure Math. Amer. Math. Soc., Providence, RI, pp. 3–20. MR: [1265519](#) (cit. on p. [3114](#)).
- Donald E. Knuth (1981). “[A permanent inequality](#)”. *Amer. Math. Monthly* 88.10, pp. 731–740, 798. MR: [668399](#) (cit. on p. [3115](#)).
- Matthias Lenz (2013). “[The  \$f\$ -vector of a representable-matroid complex is log-concave](#)”. *Adv. in Appl. Math.* 51.5, pp. 543–545. MR: [3118543](#) (cit. on p. [3121](#)).
- Nathan Linial (1984). “[The information-theoretic bound is good for merging](#)”. *SIAM J. Comput.* 13.4, pp. 795–801. MR: [764179](#) (cit. on p. [3117](#)).
- J. H. van Lint (1982). “[The van der Waerden conjecture: two proofs in one year](#)”. *Math. Intelligencer* 4.2, pp. 72–77. MR: [672919](#) (cit. on p. [3115](#)).
- Albert W. Marshall, Ingram Olkin, and Barry C. Arnold (2011). *[Inequalities: theory of majorization and its applications](#)*. Second. Springer Series in Statistics. Springer, New York, pp. xxviii+909. MR: [2759813](#) (cit. on p. [3111](#)).
- J. H. Mason (1972). “Matroids: unimodal conjectures and Motzkin’s theorem”, pp. 207–220. MR: [0349445](#) (cit. on p. [3120](#)).
- P. McMullen (1971). “[The numbers of faces of simplicial polytopes](#)”. *Israel J. Math.* 9, pp. 559–570. MR: [0278183](#) (cit. on p. [3113](#)).
- Peter McMullen (1993). “[On simple polytopes](#)”. *Invent. Math.* 113.2, pp. 419–444. MR: [1228132](#) (cit. on pp. [3112](#), [3116](#)).
- Peter Nelson (May 2016). “[Almost all matroids are non-representable](#)”. arXiv: [1605.04288](#) (cit. on p. [3122](#)).
- Andrei Okounkov (2003). “Why would multiplicities be log-concave?” In: *The orbit method in geometry and physics (Marseille, 2000)*. Vol. 213. Progr. Math. Birkhäuser Boston, Boston, MA, pp. 329–347. MR: [1995384](#) (cit. on p. [3111](#)).
- James Oxley (2011). *[Matroid theory](#)*. Second. Vol. 21. Oxford Graduate Texts in Mathematics. Oxford University Press, Oxford, pp. xiv+684. MR: [2849819](#) (cit. on pp. [3121](#), [3125](#)).
- Ronald C. Read (1968). “An introduction to chromatic polynomials”. *J. Combinatorial Theory* 4, pp. 52–71. MR: [0224505](#) (cit. on p. [3119](#)).
- Gian-Carlo Rota (1971). “Combinatorial theory, old and new”, pp. 229–233. MR: [0505646](#) (cit. on pp. [3125](#), [3126](#)).
- Adrien Saumard and Jon A. Wellner (2014). “[Log-concavity and strong log-concavity: a review](#)”. *Stat. Surv.* 8, pp. 45–114. MR: [3290441](#) (cit. on p. [3111](#)).

- Beniamino Segre (1937). “Intorno ad un teorema di Hodge sulla teoria della base per le curve di una superficie algebrica”. *Ann. Mat. Pura Appl.* 16.1, pp. 157–163. MR: [1553294](#) (cit. on p. [3121](#)).
- P. D. Seymour and D. J. A. Welsh (1975). “Combinatorial applications of an inequality from statistical mechanics”. *Math. Proc. Cambridge Philos. Soc.* 77, pp. 485–495. MR: [0376378](#) (cit. on p. [3117](#)).
- Richard P. Stanley (1980a). “The number of faces of a simplicial convex polytope”. *Adv. in Math.* 35.3, pp. 236–238. MR: [563925](#) (cit. on p. [3113](#)).
- (1980b). “Weyl groups, the hard Lefschetz theorem, and the Sperner property”. *SIAM J. Algebraic Discrete Methods* 1.2, pp. 168–184. MR: [578321](#) (cit. on p. [3113](#)).
  - (1981). “Two combinatorial applications of the Aleksandrov-Fenchel inequalities”. *J. Combin. Theory Ser. A* 31.1, pp. 56–65. MR: [626441](#) (cit. on p. [3117](#)).
  - (1984). “Combinatorial applications of the hard Lefschetz theorem”. In: *Proceedings of the International Congress of Mathematicians, Vol. 1, 2 (Warsaw, 1983)*. PWN, Warsaw, pp. 447–453. MR: [804700](#) (cit. on p. [3113](#)).
  - (1989). “Log-concave and unimodal sequences in algebra, combinatorics, and geometry”. In: *Graph theory and its applications: East and West (Jinan, 1986)*. Vol. 576. Ann. New York Acad. Sci. New York Acad. Sci., New York, pp. 500–535. MR: [1110850](#) (cit. on p. [3111](#)).
  - (2000). “Positivity problems and conjectures in algebraic combinatorics”. In: *Mathematics: frontiers and perspectives*. Amer. Math. Soc., Providence, RI, pp. 295–319. MR: [1754784](#) (cit. on p. [3111](#)).
- Alexander Stoimenow (2005). “Newton-like polynomials of links”. *Enseign. Math. (2)* 51.3-4, pp. 211–230. MR: [2214886](#) (cit. on p. [3126](#)).
- Bernard Teissier (1979). “Du théorème de l’index de Hodge aux inégalités isopérimétriques”. *C. R. Acad. Sci. Paris Sér. A-B* 288.4, A287–A289. MR: [524795](#) (cit. on p. [3116](#)).
- V. A. Timorin (1998). “Mixed Hodge-Riemann bilinear relations in a linear context”. *Funktsional. Anal. i Prilozhen.* 32.4, pp. 63–68, 96. MR: [1678857](#) (cit. on p. [3115](#)).
- (1999). “An analogue of the Hodge-Riemann relations for simple convex polyhedra”. *Uspekhi Mat. Nauk* 54.2(326), pp. 113–162. MR: [1711255](#) (cit. on p. [3116](#)).
- D. J. A. Welsh (1971). “Combinatorial problems in matroid theory”. In: *Combinatorial Mathematics and its Applications (Proc. Conf., Oxford, 1969)*. Academic Press, London, pp. 291–306. MR: [0278975](#) (cit. on p. [3120](#)).
- (1976). *Matroid theory*. L. M. S. Monographs, No. 8. Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York, pp. xi+433. MR: [0427112](#) (cit. on pp. [3124](#)–[3126](#)).
- Neil White, ed. (1987). *Combinatorial geometries*. Vol. 29. Encyclopedia of Mathematics and its Applications. Cambridge University Press, Cambridge, pp. xii+212. MR: [921064](#) (cit. on p. [3125](#)).

- Hassler Whitney (1932). “[A logical expansion in mathematics](#)”. *Bull. Amer. Math. Soc.* 38.8, pp. 572–579. MR: [1562461](#) (cit. on p. [3119](#)).
- (1935). “[On the Abstract Properties of Linear Dependence](#)”. *Amer. J. Math.* 57.3, pp. 509–533. MR: [1507091](#) (cit. on p. [3121](#)).

Received 2017-11-29.

JUNE HUH  
INSTITUTE FOR ADVANCED STUDY  
FULD HALL  
1 EINSTEIN DRIVE  
PRINCETON, NJ  
USA  
[huh@princeton.edu](mailto:huh@princeton.edu)  
[junehuh@ias.edu](mailto:junehuh@ias.edu)





# HYPERGRAPH MATCHINGS AND DESIGNS

PETER KEEVASH

## Abstract

We survey some aspects of the perfect matching problem in hypergraphs, with particular emphasis on structural characterisation of the existence problem in dense hypergraphs and the existence of designs.

## 1 Introduction

Matching theory is a rich and rapidly developing subject that touches on many areas of Mathematics and its applications. Its roots are in the work of [Steinitz \[1894\]](#), [Egerváry \[1931\]](#), [Hall \[1935\]](#) and [König \[1931\]](#) on conditions for matchings in bipartite graphs. After the rise of academic interest in efficient algorithms during the mid 20th century, three cornerstones of matching theory were Kuhn’s ‘Hungarian’ algorithm ([Kuhn \[1955\]](#)) for the Assignment Problem, Edmonds’ algorithm ([Edmonds \[1965\]](#)) for finding a maximum matching in general (not necessarily bipartite) graphs, and the [Gale and Shapley \[1962\]](#) algorithm for Stable Marriages. For an introduction to matching theory in graphs we refer to [Lovász and Plummer \[2009\]](#), and for algorithmic aspects to parts II and III of [Schrijver \[2003\]](#).

There is also a very large literature on matchings in hypergraphs. This article will be mostly concerned with one general direction in this subject, namely to determine conditions under which the necessary ‘geometric’ conditions of ‘space’ and ‘divisibility’ are sufficient for the existence of a perfect matching. We will explain these terms and discuss some aspects of this question in the next two sections, but first, for the remainder of this introduction, we will provide some brief pointers to the literature in some other directions.

We do not expect a simple general characterisation of the perfect matching problem in hypergraphs, as by contrast with the graph case, it is known to be NP-complete even for 3-graphs (i.e. when all edges have size 3), indeed, this was one of Karp’s original 21 NP-complete problems [Karp \[1972\]](#). Thus for algorithmic questions related to hypergraph

matching, we do not expect optimal solutions, and may instead consider Approximation Algorithms (see e.g. [Williamson and Shmoys \[2011\]](#), [Asadpour, Feige, and Saberi \[2008\]](#), and [Lau, Ravi, and Singh \[2011\]](#)).

Another natural direction is to seek nice sufficient conditions for perfect matchings. There is a large literature in Extremal Combinatorics on results under minimum degree assumptions, known as ‘Dirac-type’ theorems, after the classical result of [Dirac \[1952\]](#) that any graph on  $n \geq 3$  vertices with minimum degree at least  $n/2$  has a Hamiltonian cycle. It is easy to see that  $n/2$  is also the minimum degree threshold for a graph on  $n$  vertices (with  $n$  even) to have a perfect matching, and this exemplifies the considerable similarities between the perfect matching and Hamiltonian problems (but there are also substantial differences). A landmark in the efforts to obtain hypergraph generalisations of Dirac’s theorem was the result of [Rödl, Ruciński, and Szemerédi \[2009\]](#) that determined the codegree threshold for perfect matchings in uniform hypergraphs; this paper was significant for its proof method as well as the result, as it introduced the Absorbing Method (see [Section 5](#)), which is now a very important tool for proving the existence of spanning structures. There is such a large body of work in this direction that it needs several surveys to describe, and indeed these surveys already exist [Rödl and Ruciński \[2010\]](#), [Kühn and Osthus \[2009\]](#), [Kühn and Osthus \[2014\]](#), [Zhao \[2016\]](#), and [Yuster \[2007\]](#). The most fundamental open problem in this area is the Erdős Matching Conjecture [Erdős \[1965\]](#), namely that the maximum number of edges in an  $r$ -graph<sup>1</sup> on  $n$  vertices with no matching of size  $t$  is either achieved by a clique of size  $tr - 1$  or the set of all edges hitting some fixed set of size  $t - 1$  (see [Frankl and Tokushige \[2016, Section 3\]](#) for discussion and a summary of progress).

The duality between matching and covers in hypergraphs is of fundamental importance in Combinatorics (see [Füredi \[1988\]](#)) and Combinatorial Optimisation (see [Cornuéjols \[2001\]](#)). A defining problem for this direction of research within Combinatorics is ‘Ryser’s Conjecture’ (published independently by [Henderson \[1971\]](#) and [Lovász \[1975\]](#)) that in any  $r$ -partite  $r$ -graph the ratio of the covering and matching numbers is at most  $r - 1$ . For  $r = 2$  this is König’s Theorem. The only other known case is  $r = 3$ , due to [Aharoni \[2001\]](#), using a hypergraph analogue of Hall’s theorem due to [Aharoni and Haxell \[2000\]](#), which has a topological proof. There are now many applications of topology to hypergraph matching, and more generally ‘independent transversals’ (see the survey [Haxell \[2016\]](#)). In the other direction, the hypergraph matching complex is now a fundamental object of Combinatorial Topology, with applications to Quillen complexes in Group Theory, Vassiliev knot invariants and Computational Geometry (see the survey [Wachs \[2003\]](#)).

From the probabilistic viewpoint, there are (at least) two natural questions:

- (i) does a random hypergraph have a perfect matching with high probability (whp)?

---

<sup>1</sup> An  $r$ -graph is a hypergraph in which every edge contains  $r$  vertices.

(ii) what does a random matching from a given (hyper)graph look like?

The first question for the usual (binomial) random hypergraph was a longstanding open problem, perhaps first stated by Erdős [1981] (who attributed it to Shamir), finally solved by Johansson, Kahn, and Vu [2008]; roughly speaking, the threshold is ‘where it should be’, namely around the edge probability at which with high probability every vertex is in at least one edge. Another such result due to Cooper, Frieze, Molloy, and Reed [1996] is that random regular hypergraphs (of fixed degree and edge size) whp have perfect matchings.

The properties of random matchings in lattices have been extensively studied under the umbrella of the ‘dimer model’ (see Kenyon [2010]) in Statistical Physics. However, rather little is known regarding the typical structure of random matchings in general graphs, let alone hypergraphs. Substantial steps in this direction have been taken by results of Kahn [2000] characterising when the size of a random matching has an approximate normal distribution, and Kahn and Kayll [1997] establishing long-range decay of correlations of edges in random matchings in graphs; the final section of Kahn [2000] contains many open problems, including conjectural extensions to simple hypergraphs.

Prerequisite to the understanding of random matchings are the closely related questions of Sampling and Approximate Counting (as established in the Markov Chain Monte Carlo framework of Jerrum and Sinclair, see Jerrum [2003]). An approximate counting result for hypergraph matchings with respect to balanced weight functions was obtained by Barvinok and Samorodnitsky [2011]. Extremal problems also arise naturally in this context, for the number of matchings, and more generally for other models in Statistical Physics, such as the hardcore model for independent sets. Much of the recent progress here appears in the survey Zhao [2017], except for the very recent solution of (almost all cases of) the Upper Matching Conjecture of Friedland, Krop, and Markström [2008] by Davies, Jenssen, Perkins, and Roberts [2017].

## 2 Space and divisibility

In this section we discuss a result (joint work with Mycroft 2015) that characterises the obstructions to perfect matching in dense hypergraphs (under certain conditions to be discussed below). The obstructions are geometric in nature and are of two types: Space Barriers (metric obstructions) and Divisibility Barriers (arithmetic obstructions).

The simplest illustration of these two phenomena is seen by considering extremal examples for the simple observation mentioned earlier that a graph on  $n$  vertices ( $n$  even) with minimum degree at least  $n/2$  has a perfect matching. One example of a graph with minimum degree  $n/2 - 1$  and no perfect matching is obtained by fixing a set  $S$  of  $n/2 - 1$  vertices and taking all edges that intersect  $S$ . Then in any matching  $M$ , each edge of  $M$

uses at least one vertex of  $S$ , so  $|M| \leq |S| < n/2$ ; there is no ‘space’ for a perfect matching. For another example, suppose  $n = 2 \bmod 4$  and consider the graph that is the disjoint union of two cliques each of size  $n/2$  (which is odd). As edges have size 2, which is even, there is an arithmetic (parity) obstruction to a perfect matching.

There is an analogous parity obstruction to matching in general  $r$ -graphs, namely an  $r$ -graph  $G$  with vertices partitioned as  $(A, B)$ , so that  $|A|$  is odd and  $|e \cap A|$  is even for each edge  $e$  of  $G$ ; this is one of the extremal examples for the codegree threshold of perfect matchings (see Rödl, Ruciński, and Szemerédi [2009]).

In general, space barriers are constructions for each  $1 \leq i \leq r$ , obtained by fixing a set  $S$  of size less than  $in/r$  and taking the  $r$ -graph of all edges  $e$  with  $|e \cap S| \geq i$ . Then for any matching  $M$  we have  $|M| \leq |S|/i < n/r$ , so  $M$  is not perfect.

General divisibility barriers are obtained by fixing a lattice (additive subgroup)  $L$  in  $\mathbb{Z}^d$  for some  $d$ , fixing a vertex set partitioned as  $(V_1, \dots, V_d)$ , with  $(|V_1|, \dots, |V_d|) \notin L$ , and taking the  $r$ -graph of all edges  $e$  such that  $(|e \cap V_1|, \dots, |e \cap V_d|) \in L$ . For example, the parity obstruction corresponds to the lattice  $\{(2x, y) : x, y \in \mathbb{Z}\}$ .

To state the result of Keevash and Mycroft [2015a] that is most conveniently applicable we introduce the setting of simplicial complexes and degree sequences. We consider a simplicial complex  $J$  on  $[n] = \{1, \dots, n\}$ , write  $J_i = \{e \in J : |e| = i\}$  and look for a perfect matching in the  $r$ -graph  $J_r$ . We define the degree sequence  $(\delta_0(J), \dots, \delta_{r-1}(J))$  so that each  $\delta_i(J)$  is the least  $m$  such that each  $e \in J_i$  is contained in at least  $m$  edges of  $J_{i+1}$ . We define the critical degree sequence  $\delta^c = (\delta_0^c, \dots, \delta_{r-1}^c)$  by  $\delta_i^c = (1 - i/r)n$ . The space barrier constructions show that for each  $i$  there is a complex with  $\delta_i(J)$  slightly less than  $\delta_i^c$  but no perfect matching. An informal statement of Keevash and Mycroft [ibid., Theorem 2.9] is that if  $J$  is an  $r$ -complex on  $[n]$  (where  $r \mid n$ ) with all  $\delta_i(J) \geq \delta_i^c - o(n)$  such that  $J_r$  has no perfect matching then  $J$  is close (in edit distance) to a space barrier or a divisibility barrier.

One application of this result (also given in Keevash and Mycroft [ibid.]) is to determine the exact codegree threshold for packing tetrahedra in 3-graphs; it was surprising that it was possible to obtain such a result given that the simpler-sounding problems of determining the thresholds (edge or codegree) for the existence of just one tetrahedron are open, even asymptotically (the edge threshold is a famous conjecture of Turán; for more on Turán problems for hypergraphs see the survey Keevash [2011b]). Other applications are a multipartite version of the Hajnal-Szemerédi theorem (see Keevash and Mycroft [2015b]) and determining the ‘hardness threshold’ for perfect matchings in dense hypergraphs (see Keevash, Knox, and Mycroft [2015] and Han [2017]).

We will describe the hardness threshold in more detail, as it illustrates some important features of space and divisibility, and the distinction between perfect matchings and almost perfect matchings. For graphs there is no significant difference in the thresholds

for these problems, whereas for general  $r$ -graphs there is a remarkable contrast: the codegree threshold for perfect matchings Rödl, Ruciński, and Szemerédi [2009] is about  $n/2$ , whereas Han [2015], proving a conjecture from Rödl, Ruciński, and Szemerédi [2009], showed that a minimum codegree of only  $n/r$  guarantees a matching of size  $n/r - 1$ , i.e. one less than perfect. The explanation for this contrast is that the divisibility barrier is no obstacle to almost perfect matching, whereas the space barrier is more robust, and can be continuously ‘tuned’ to exclude a matching of specified size.

To illustrate this, we consider a 3-graph  $G_0$  on  $[n]$  where the edges are all triples that intersect some fixed set  $S$  of size  $(1/3 - c)n$ , for some small  $c > 0$ . Then the minimum codegree and maximum matching size in  $G_0$  are both equal to  $|S|$ . Furthermore, if we consider  $G = G_0 \cup G_1$  where all edges of  $G_1$  lie within some  $S'$  disjoint from  $S$  with  $|S'| = 3cn$  then  $G$  has a perfect matching if and only if  $G_1$  has a perfect matching, which is NP-complete to decide for arbitrary  $G_1$ . Thus the robustness of the space barrier provides a reduction showing that the codegree threshold for the existence of an algorithm for the perfect matching is at least the threshold for an approximate perfect matching.

Now consider the decision problem for perfect matchings in 3-graphs on  $[n]$  (where  $3 \mid n$ ) with minimum codegree at least  $\delta n$ . For  $\delta < 1/3$  the problem is NP-complete, and for  $\delta > 1/2$  it is trivial (there is a perfect matching by Rödl, Ruciński, and Szemerédi [ibid.]). For intermediate  $\delta$  there is a polynomial-time algorithm, and this is in essence a structural stability result: the main ingredient of the algorithm is a result of Keevash, Knox, and Mycroft [2015] that any such 3-graph with no perfect matching is contained in a divisibility barrier. (For general  $r$  the structural characterisation is more complicated.)

### 3 Fractional matchings

The key idea of the Absorbing Method of Rödl, Ruciński, and Szemerédi [2009] mentioned earlier is that the task of finding perfect matchings can often be broken into two subproblems: (i) finding almost perfect matchings, (ii) absorbing uncovered vertices into an almost perfect matching until it becomes perfect. We have already seen that the almost perfect matching problem appears naturally as a relaxation of the perfect matching problem in which we eliminate divisibility obstacles but retain space obstacles. This turns out to fit into a more general framework of fractional matchings, in which the relaxed problem is a question of convex geometry, and space barriers correspond to separating hyperplanes.

The fractional (linear programming) relaxation of the perfect matching problem in a hypergraph is to assign non-negative weights to the edges so that for any vertex  $v$ , there is a total weight of 1 on all edges incident to  $v$ . A perfect matching corresponds to a  $\{0, 1\}$ -valued solution, so the existence of a fractional perfect matching is necessary for the existence of a perfect matching. We can adopt a similar point of view regarding divisibility

conditions. Indeed, we can similarly define the integer relaxation of the perfect matching problem in which we now require the weights to be integers (not necessarily non-negative); then the existence of an integral perfect matching is necessary for the existence of a perfect matching.

The fractional matching problem appears naturally in Combinatorial Optimisation (see Cornuéjols [2001] and Schrijver [2003]) because it brings in polyhedral methods and duality to bear on the matching problem. It has also been studied as a problem in its own right from the perspective of random thresholds (e.g. Devlin and Kahn [2017] and Krivelevich [1996]), and it appears naturally in combinatorial existence problems, as in dense hypergraphs almost perfect matchings and fractional matchings tend to appear at the same threshold. Indeed, for many open problems, such as the Erdős Matching Conjecture Erdős [1965] or the Nash-Williams Triangle Decomposition Conjecture Nash-Williams [1970], any progress on the fractional problem translates directly into progress on the original problem (see Barber, Kühn, Lo, and Osthus [2016]).

This therefore makes the threshold problem for fractional matchings and decompositions a natural problem in its own right. For example, an asymptotic solution of the Nash-Williams Conjecture would follow from the following conjecture: any graph on  $n$  vertices with minimum degree at least  $3n/4$  has a fractional triangle decomposition, i.e. an assignment of non-negative weights to its triangles so that for any edge  $e$  there is total weight 1 on the triangles containing  $e$ . An extremal example  $G$  for this question can be obtained by taking a balanced complete bipartite graph  $H$  and adding a  $(n/4 - 1)$ -regular graph inside each part; indeed, this is a space barrier to a fractional triangle decomposition, as any triangle uses at least one edge not in  $H$ , but  $|H| > 2|G \setminus H|$ . The best known upper bound is  $0.913n$  by Dross [2016]. More generally, Barber, Kühn, Lo, Montgomery, and Osthus [2017] give the current best known bounds on the thresholds for fractional clique decompositions (in graphs and hypergraphs), but these seem to be far from optimal.

There are (at least) two ways to think about the relationship between almost perfect matchings and fractional matchings. The first goes back to the ‘nibble’ (semi-random) method of Rödl [1985], introduced to solve the Erdős and Hanani [1963] conjecture on approximate Steiner systems (see the next section), which has since had a great impact on Combinatorics (e.g. Alon, J. H. Kim, and Spencer [1997], Bennett and Bohman [2012], Bohman [2009], Bohman, Frieze, and Lubetzky [2015], Bohman and Keevash [2010, 2013], Pontiveros, Griffiths, and Morris [2013], Frankl and Rödl [1985], Grable [1999], Kahn [1996b,a], J. H. Kim [2001], Kostochka and Rödl [1998], Kuzjurin [1995], Pippenger and Spencer [1989], Spencer [1995], and Vu [2000]). A special case of a theorem of Kahn [1996a] is that if there is a fractional perfect matching on the edges of an  $r$ -graph  $G$  on  $[n]$  such that for any pair of vertices  $x, y$  the total weight on edges containing  $\{x, y\}$  is  $o(1)$  then  $G$  has a matching covering all but  $o(n)$  vertices. In this viewpoint, it is natural to interpret the weights of a fractional matching as probabilities, and an almost perfect

matching as a random rounding; in fact, this random rounding is obtained iteratively, so there are some parallels with the development of iterative rounding algorithms (see [Lau, Ravi, and Singh \[2011\]](#)).

Another way to establish the connection between almost perfect matchings and fractional matchings is via the theory of Regularity, developed by [Szemerédi \[1978\]](#) for graphs and extended to hypergraphs independently by [Gowers \[2007\]](#) and [Nagle, Rödl, and Schacht \[2006\]](#), [Rödl and Schacht \[2007b,a\]](#), and [Rödl and Skokan \[2004\]](#). (The connection was first established by [Haxell and Rödl \[2001\]](#) for graphs and [Rödl, Schacht, Siggers, and Tokushige \[2007\]](#) for hypergraphs.) To apply Regularity to obtain spanning structures (such as perfect matchings) requires an accompanying result known as a blowup lemma, after the original such result for graphs obtained by [Komlós, Sárközy, and Szemerédi \[1997\]](#); we proved the hypergraph version in [Keevash \[2011a\]](#). More recent developments (for graphs) along these lines include the Sparse Blowup Lemmas of [Allen, Böttcher, Hàn, Kohayakawa, and Person \[2016\]](#) and a blowup-up lemma suitable for decompositions (as in the next section) obtained by [J. Kim, Kühn, Osthus, and Tyomkyn \[2016\]](#) (it would be interesting and valuable to obtain hypergraph versions of these results). The technical difficulties of the Hypergraph Regularity method are a considerable barrier to its widespread application, and preclude us giving here a precise statement of [Keevash and Mycroft \[2015a, Theorem 9.1\]](#), which informally speaking characterises the perfect matching problem in dense hypergraphs with certain extendability conditions in terms of space and divisibility.

## 4 Designs and decompositions

A *Steiner system* with parameters  $(n, q, r)$  is a  $q$ -graph  $G$  on  $[n]$  such that any  $r$ -set of vertices is contained in exactly one edge. For example, a Steiner Triple System on  $n$  points has parameters  $(n, 3, 2)$ . The question of whether there is a Steiner system with given parameters is one of the oldest problems in combinatorics, dating back to work of Plücker (1835), Kirkman (1846) and Steiner (1853); see [R. Wilson \[2003\]](#) for a historical account.

Note that a Steiner system with parameters  $(n, q, r)$  is equivalent to a  $K_q^r$ -decomposition of  $K_n^r$  (the complete  $r$ -graph on  $[n]$ ). It is also equivalent to a perfect matching in the auxiliary  $\binom{q}{r}$ -graph on  $\binom{[n]}{r}$  (the  $r$ -subsets of  $[n] := \{1, \dots, n\}$ ) with edge set  $\{\binom{Q}{r} : Q \in \binom{[n]}{q}\}$ .

More generally, we say that a set  $S$  of  $q$ -subsets of an  $n$ -set  $X$  is a *design* with parameters  $(n, q, r, \lambda)$  if every  $r$ -subset of  $X$  belongs to exactly  $\lambda$  elements of  $S$ . (This is often called an ‘ $r$ -design’ in the literature.) There are some obvious necessary ‘divisibility conditions’ for the existence of such  $S$ , namely that  $\binom{q-i}{r-i}$  divides  $\lambda \binom{n-i}{r-i}$  for every



$0 \leq i \leq r-1$  (fix any  $i$ -subset  $I$  of  $X$  and consider the sets in  $S$  that contain  $I$ ). It is not known who first advanced the ‘Existence Conjecture’ that the divisibility conditions are also sufficient, apart from a finite number of exceptional  $n$  given fixed  $q$ ,  $r$  and  $\lambda$ .

The case  $r = 2$  has received particular attention due to its connections to statistics, under the name of ‘balanced incomplete block designs’. We refer the reader to [Colbourn and Dinitz \[2007\]](#) for a summary of the large literature and applications of this field. The Existence Conjecture for  $r = 2$  was a long-standing open problem, eventually resolved by [R. M. Wilson \[1972a,b, 1975\]](#) in a series of papers that revolutionised Design Theory. The next significant progress on the general conjecture was in the solution of the two relaxations (fractional and integer) discussed in the previous section (both of which are interesting in their own right and useful for the original problem). We have already mentioned Rödl’s solution of the Erdős–Hanani Conjecture on approximate Steiner systems. The integer relaxation was solved independently by [Graver and Jurkat \[1973\]](#) and [R. M. Wilson \[1973\]](#), who showed that the divisibility conditions suffice for the existence of integral designs (this is used in [R. M. Wilson \[ibid.\]](#) to show the existence for large  $\lambda$  of integral designs with non-negative coefficients). [R. M. Wilson \[1999\]](#) also characterised the existence of integral  $H$ -decompositions for any  $r$ -graph  $H$ .

The existence of designs with  $r \geq 7$  and any ‘non-trivial’  $\lambda$  was open before the breakthrough result of [Teirlinck \[1987\]](#) confirming this. An improved bound on  $\lambda$  and a probabilistic method (a local limit theorem for certain random walks in high dimensions) for constructing many other rigid combinatorial structures was recently given by [Kuperberg, Lovett, and Peled \[2017\]](#). [Ferber, Hod, Krivelevich, and Sudakov \[2014\]](#) gave a construction of ‘almost Steiner systems’, in which every  $r$ -subset is covered by either one or two  $q$ -subsets.

In [Keevash \[2014\]](#) we proved the Existence Conjecture in general, via a new method of Randomised Algebraic Constructions. Moreover, in [Keevash \[2015\]](#) we obtained the following estimate for the number  $D(n, q, r, \lambda)$  of designs with parameters  $(n, q, r, \lambda)$  satisfying the necessary divisibility conditions: writing  $Q = \binom{q}{r}$  and  $N = \binom{n-r}{q-r}$ , we have

$$D(n, q, r, \lambda) = \lambda!^{-\binom{n}{r}} ((\lambda/e)^{Q-1} N + o(N))^{\lambda Q^{-1} \binom{n}{r}}.$$

Our counting result is complementary to that in [Kuperberg, Lovett, and Peled \[2017\]](#), as it applies (e.g.) to Steiner Systems, whereas theirs is only applicable to large multiplicities (but also allows the parameters  $q$  and  $r$  to grow with  $n$ , and gives an asymptotic formula when applicable).

The upper bound on the number of designs follows from the entropy method pioneered by [Radhakrishnan \[1997\]](#); more generally, [Luria \[2017\]](#) has recently established a similar upper bound on the number of perfect matchings in any regular uniform hypergraph with small codegreess. The lower bound essentially matches the number of choices in the

Random Greedy Hypergraph Matching process (see [Bennett and Bohman \[2012\]](#)) in the auxiliary  $Q$ -graph defined above, so the key to the proof is showing that this process can be stopped so that whp it is possible to complete the partial matching thus obtained to a perfect matching. In other words, instead of a design, which can be viewed as a  $K_q^r$ -decomposition of the  $r$ -multigraph  $\lambda K_n^r$ , we require a  $K_q^r$ -decomposition of some sparse submultigraph, that satisfies the necessary divisibility conditions, and has certain pseudo-randomness properties (guaranteed whp by the random process).

The main result of [Keevash \[2014\]](#) achieved this, and indeed (in the second version of the paper) we obtained a more general result in the same spirit as [Keevash and Mycroft \[2015a\]](#), namely that we can find a clique decomposition of any  $r$ -multigraph with a certain ‘extendability’ property that satisfies the divisibility conditions and has a ‘suitably robust’ fractional clique decomposition.

[Glock, Kühn, Lo, and Osthus \[2016, 2017\]](#) have recently given a new proof of the existence of designs, as well as some generalisations, including the existence of  $H$ -decompositions for any hypergraph  $H$  (a question from [Keevash \[2014\]](#)), relaxing the quasirandomness condition from [Keevash \[ibid.\]](#) (version 1) to an extendability condition in the same spirit as [Keevash \[ibid.\]](#) (version 2), and a more effective bound than that in [Keevash \[ibid.\]](#) on the minimum codegree decomposition threshold; the main difference in our approaches lies in the treatment of absorption (see the next section).

## 5 Absorption

Over the next three sections we will sketch some approaches to what is often the most difficult part of a hypergraph matching or decomposition problem, namely converting an approximate solution into an exact solution. We start by illustrating the Absorbing Method in its original form, namely the determination in [Rödl, Ruciński, and Szemerédi \[2009\]](#) of the codegree threshold for perfect matchings in  $r$ -graphs; for simplicity we consider  $r = 3$ .

We start by solving the almost perfect matching problem. Let  $G$  be a 3-graph on  $[n]$  with  $3 \mid n$  and minimum codegree  $\delta(G) = n/3$ , i.e. every pair of vertices is in at least  $n/3$  edges. We show that  $G$  has a matching of size  $n/3 - 1$  (i.e. one less than perfect). To see this, consider a maximum size matching  $M$ , let  $V_0 = V(G) \setminus V(M)$ , and suppose  $|V_0| > 3$ . Then  $|V_0| \geq 6$ , so we can fix disjoint pairs  $a_1b_1, a_2b_2, a_3b_3$  in  $V_0$ . For each  $i$  there are at least  $n/3$  choices of  $c$  such that  $a_ib_ic \in E(G)$ , and by maximality of  $M$  any such  $c$  lies in  $V(M)$ . We define the weight  $w_e$  of each  $e \in M$  as the number of edges of  $G$  of the form  $a_ib_ic$  with  $c \in e$ . Then  $\sum_{e \in M} w_e \geq n$ , and  $|M| < n/3$ , so there is  $e \in M$  with  $w_e \geq 4$ . Then there must be distinct  $c, c'$  in  $e$  and distinct  $i, i'$  in  $[3]$  such

that  $a_i b_i c$  and  $a_i' b_i' c'$  are edges. However, deleting  $e$  and adding these edges contradicts maximality of  $M$ .

Now suppose  $\delta(G) = n/2 + cn$ , where  $c > 0$  and  $n > n_0(c)$  is large. Our plan for finding a perfect matching is to first put aside an ‘absorber’  $A$ , which will be a matching in  $G$  with the property that for any triple  $T$  in  $V(G)$  there is some edge  $e \in A$  such that  $T \cup e$  can be expressed as the disjoint union of two edges in  $G$  (then we say that  $e$  absorbs  $T$ ). Suppose that we can find such  $A$ , say with  $|A| < n/20$ . Deleting the vertices of  $A$  leaves a 3-graph  $G'$  on  $n' = n - |A|$  vertices with  $\delta(G') \geq \delta(G) - 3|A| > n'/3$ . As shown above,  $G'$  has a matching  $M'$  with  $|M'| = n'/3 - 1$ . Let  $T = V(G') \setminus V(M')$ . By choice of  $A$  there is  $e \in A$  such that  $T \cup e = e_1 \cup e_2$  for some disjoint edges  $e_1, e_2$  in  $G$ . Then  $M' \cup (A \setminus \{e\}) \cup \{e_1, e_2\}$  is a perfect matching in  $G$ .

It remains to find  $A$ . The key idea is that for any triple  $T$  there are many edges in  $G$  that absorb  $T$ , and so if  $A$  is random then whp many of them will be present. We can bound the number of absorbers for any triple  $T = xyz$  by choosing vertices sequentially. Say we want to choose an edge  $e = x'y'z'$  so that  $x'yz$  and  $xy'z'$  are also edges. There are at least  $n/2 + cn$  choices for  $x'$  so that  $x'yz$  is an edge. Then for each of the  $n - 4$  choices of  $y' \in V(G) \setminus \{x, y, z, x'\}$  there are at least  $2cn - 1$  choices for  $z' \neq z$  so that  $x'y'z'$  and  $xy'z'$  are edges. Multiplying the choices we see that  $T$  has at least  $cn^3$  absorbers.

Now suppose that we construct  $A$  by choosing each edge of  $G$  independently with probability  $c/(4n^2)$  and deleting any pair that intersect. Let  $X$  be the number of deleted edges. There are fewer than  $n^5$  pairs of edges that intersect, so  $\mathbb{E}X < c^2n/16$ , so  $\mathbb{P}(X < c^2n/8) \geq 1/2$ . Also, the number of chosen absorbers  $N_T$  for any triple  $T$  is binomial with mean at least  $c^2n/4$ , so whp all  $N_T > c^2n/8$ . Thus there is a choice of  $A$  such that every  $T$  has an absorber in  $A$ . This completes the proof of the approximate version of Rödl, Ruciński, and Szemerédi [2009], i.e. that minimum codegree  $n/2 + cn$  guarantees a perfect matching.

The idea for the exact result is to consider an attempt to construct absorbers as above under the weaker assumption  $\delta(G) \geq n/2 - o(n)$ . It is not hard to see that absorbers exist unless  $G$  is close to one of the extremal examples. The remainder of the proof (which we omit) is then a stability analysis to show that the extremal examples are locally optimal, and so optimal.

In the following two sections we will illustrate two approaches to absorption for designs and hypergraph decompositions, in the special case of triangle decompositions of graphs, which is considerably simpler, and so allows us to briefly illustrate some (but not all) ideas needed for the general case. First we will conclude this section by indicating why the basic method described above does not suffice.

Suppose we seek a triangle decomposition of a graph  $G$  on  $[n]$  with  $e(G) = \Omega(n^2)$  in which there is no space or divisibility obstruction: we assume that  $G$  is ‘tridivisible’ (meaning that  $3 \mid e(G)$  and all degrees are even) and ‘triangle-regular’ (meaning that

there is a set  $T$  of triangles in  $G$  such that every edge is in  $(1 + o(1))tn$  triangles of  $T$ , where  $t > 0$  and  $n > n_0(t)$ . This is equivalent to a perfect matching in the auxiliary 3-graph  $H$  with  $V(H) = E(G)$  and  $E(H) = \{\{ab, bc, ca\} : abc \in T\}$ . Note that  $H$  is ‘sparse’: we have  $e(H) = O(v(H)^{3/2})$ . Triangle regularity implies that Pippenger’s generalisation (see [Pippenger and Spencer \[1989\]](#)) of the Rödl nibble can be applied to give an almost perfect matching in  $H$ , so the outstanding question is whether there is an absorber.

Let us consider a potential random construction of an absorber  $A$  in  $H$ . It will contain at most  $O(n^2)$  triangles, so the probability of any triangle (assuming no heavy bias) will be  $O(n^{-1})$ . On the other hand, to absorb some fixed (tridivisible)  $S \subseteq E(G)$ , we need  $A$  to contain a set  $A_S$  of  $a$  edge-disjoint triangles (for some constant  $a$ ) such that  $S \cup A_S$  has a triangle decomposition  $B_S$ , so we need  $\Omega(n^a)$  such  $A_S$  in  $G$ . To see that this is impossible, we imagine selecting the triangles of  $A_S$  one at a time and keeping track of the number  $E_S$  of edges that belong to a unique triangle of  $S \cup A_S$ . If a triangle uses a vertex that has not been used previously then it increases  $E_S$ , and otherwise it decreases  $E_S$  by at most 3. We can assume that no triangle is used in both  $A_S$  and  $B_S$ , so we terminate with  $E_S = 0$ . Thus there can be at most  $3a/4$  steps in which  $E_S$  increases, so there are only  $O(n^{3a/4})$  such  $A_S$  in  $G$ .

The two ideas discussed below to overcoming this obstacle can be briefly summarised as follows. In Randomised Algebraic Construction (introduced in [Keevash \[2014\]](#)), instead of choosing independent random triangles for an absorber, they are randomly chosen according to a superimposed algebraic structure that has ‘built-in’ absorbers. In Iterative Absorption (used for designs and decompositions in [Glock, Kühn, Lo, and Osthus \[2016, 2017\]](#)), instead of a single absorption step, there is a sequence of absorptions, each of which replaces the current subgraph of uncovered edges by an ‘easier’ subgraph, until we obtain  $S$  that is so simple that it can be absorbed by an ‘exclusive’ absorber put aside at the beginning of the proof for the eventuality that we end up with  $S$ . This is a powerful idea with many other applications (see the survey [Kühn and Osthus \[2014\]](#)).

## 6 Iterative Absorption

Here we will sketch an application of iterative absorption to finding a triangle decomposition of a graph  $G$  with no space or divisibility obstruction as in the previous subsection. (We also make certain ‘extendability’ assumptions that we will describe later when they are needed.) Our sketch will be loosely based on a mixture of the methods used in [Barber, Kühn, Lo, and Osthus \[2016\]](#) and [Glock, Kühn, Lo, and Osthus \[2016\]](#), thus illustrating some ideas of the general case but omitting most of the technicalities.

The plan for the decomposition is to push the graph down a ‘vortex’, which consists of a nested sequence  $V(G) = V_0 \supseteq V_1 \supseteq \dots \supseteq V_\tau$ , where  $|V_i| = \theta|V_{i-1}|$  for each  $i \in [\tau]$  with  $n^{-1} \ll \theta \ll t$ , and  $|V_\tau| = O(1)$  (so  $\tau$  is logarithmic in  $n = v(G)$ ). Suppose  $G$  has a set  $T$  of triangles such that every edge is in  $(1 \pm c)tn$  triangles of  $T$ , where  $n^{-1} \ll \theta \ll c, t$ . By choosing the  $V_i$  randomly we can ensure that each edge of  $G[V_i]$  is in  $(1 \pm 2c)t|V_i|$  triangles of  $T_i = \{f \in T : f \subseteq V_i\}$ . At step  $i$  with  $0 \leq i \leq t$  we will have covered all edges of  $G$  not contained in  $V_i$  by edge-disjoint triangles, and also some edges within  $V_i$ , in a suitably controlled manner, so that we still have good triangle regularity in  $G[V_i]$ .

At the end of the process, the uncovered subgraph  $L$  will be contained in  $V_\tau$ , so there are only constantly many possibilities for  $L$ . Before starting the process, for each tridivisible subgraph  $S$  of the complete graph on  $V_\tau$  we put aside edge-disjoint ‘exclusive absorbers’  $A_S$ , i.e. sets of edge-disjoint triangles in  $G$  such that  $S \cup A_S$  has a triangle decomposition  $B_S$  (we omit here the details of this construction). Then  $L$  will be equal to one of these  $S$ , so replacing  $A_S$  by  $B_S$  completes the triangle decomposition of  $G$ .

Let us now consider the process of pushing  $G$  down the vortex; for simplicity of notation we describe the first step of covering all edges not within  $V_1$ . The plan is to cover almost all of these edges by a nibble, and then the remainder by a random greedy algorithm (which will also use some edges within  $V_1$ ). At first sight this idea sounds suspicious, as one would think that the triangle regularity parameter  $c$  must increase substantially at each step, and so the process could not be iterated logarithmically many times before the parameters blow up.

However, quite suprisingly, if we make the additional extendability assumption that every edge is in at least  $c'n^3$  copies of  $K_5$  (where  $c'$  is large compared with  $c$  and  $t \geq c'$ ), then we can pass to a different set of triangles which dramatically ‘boost’ the regularity. The idea (see [Glock, Kühn, Lo, and Osthus \[2016, Lemma 6.3\]](#)) is that a relatively weak triangle regularity assumption implies the existence of a perfect fractional triangle decomposition, which can be interpreted as selection probabilities (in the same spirit as [Kahn \[1996a\]](#)) for a new set of triangles that is much more regular. A similar idea appears in the Rödl-Schacht proof of the hypergraph regularity lemma via regular approximation (see [Rödl and Schacht \[2007b\]](#)). It may also be viewed as a ‘guided version’ of the self-correction that appears naturally in random greedy algorithms (see [Bohman and Keevash \[2013\]](#) and [Pontiveros, Griffiths, and Morris \[2013\]](#)).

Let us then consider  $G^* = G \setminus G[V_1] \setminus H$ , where  $H$  contains each edge of  $G$  crossing between  $V_1$  and  $V^* := V(G) \setminus V_1$  independently with some small probability  $p \ll c, \theta$ . (We reserve  $H$  to help with the covering step.) Then whp every edge of  $G^*$  is in  $(1 \pm c)tn \pm |V_1| \pm 3pn$  triangles of  $T$  within  $G^*$ . By boosting, we can find a set  $T^*$  of triangles in  $G^*$  such that every edge of  $G^*$  is in  $(1 \pm c_0)tn/2$  triangles of  $T^*$ , where  $c_0 \ll p$ . By the nibble, we can choose a set of edge-disjoint triangles in  $T^*$  covering

all of  $G^*$  except for some ‘leave’  $L$  of maximum degree  $c_1 n$ , where we introduce new constants  $c_0 \ll c_1 \ll c_2 \ll p$ .

Now we cover  $L$  by two random greedy algorithms, the first to cover all remaining edges in  $V^*$  and the second to cover all remaining cross edges. The analysis of these algorithms is not as difficult as that of the nibble, as we have ‘plenty of space’, in that we only have to cover a sparse graph within a much denser graph, whereas the nibble seeks to cover almost all of a graph. In particular, the behaviour of these algorithms is well-approximated by a ‘binomial heuristic’ in which we imagine choosing random triangles to cover the uncovered edges without worrying about whether these triangles are edge-disjoint (so we make independent choices for each edge). In the actual algorithm we have to exclude any triangle that uses an edge covered by a previous step of the algorithm, but if we are covering a sparse graph one can show that whp at most half (say) of the choices are forbidden at each step, so any whp estimate in the binomial process will hold in the actual process up to a factor of two. (This idea gives a much simpler proof of the result of [Ferber, Hod, Krivelevich, and Sudakov \[2014\]](#).)

For the first greedy algorithm we consider each remaining edge in  $V^*$  in some arbitrary order, and when we consider  $e$  we choose a triangle on  $e$  whose two other edges are in  $H$ , and have not been previously covered. In general we would make this choice uniformly at random, although the triangle case is sufficiently simple that an arbitrary choice suffices; indeed, there are whp at least  $p^2 \theta n / 2$  such triangles in  $H$ , of which at most  $2c_1 n$  are forbidden due to using a previously covered edge (by the maximum degree of  $L$ ). Thus the algorithm can be completed with arbitrary choices.

The second greedy algorithm for covering the cross edges is more interesting (the analogous part of the proof for general designs is the most difficult part of [Glock, Kühn, Lo, and Osthus \[2016\]](#)). Let  $H'$  denote the subgraph of cross edges that are still uncovered. We consider each  $x \in V^*$  sequentially and cover all edges of  $H'$  incident to  $x$  by the set of triangles obtained by adding  $x$  to each edge of a perfect matching  $M_x$  in  $G[H'(x)]$ , i.e. the restriction of  $G$  to the  $H'$ -neighbourhood of  $x$ . We must choose  $M_x$  edge-disjoint from  $M_{x'}$  for all  $x'$  preceding  $x$ , so an arbitrary choice will not work; indeed, whp the degree of each vertex  $y$  in  $G[H'(x)]$  is  $(1 \pm c_2) p \theta n$ , but our upper bound on the degree of  $y$  in  $H'$  may be no better than  $pn$ , so previous choices of  $M_{x'}$  could isolate  $y$  in  $G[H'(x)]$ .

To circumvent this issue we choose random perfect matchings. A uniformly random choice would work, but it is easier to analyse the process where we fix many edge-disjoint matchings in  $G[H'(x)]$  and then choose one uniformly at random to be  $M_x$ . We need some additional assumption to guarantee that  $G[H'(x)]$  has even one perfect matching (the approximate regularity only guarantees an almost perfect matching).

One way to achieve this is to make the additional mild extendability assumption that every pair of vertices have at least  $c'n$  common neighbours in  $G[H'(x)]$ , i.e. any adjacent pair of edges  $xy, xy'$  in  $G$  have at least  $c'n$  choices of  $z$  such that  $xz, yz$  and  $y'z$  are

edges. It is then not hard to see that a random balanced bipartite subgraph of  $G[H'(x)]$  whp satisfies Hall's condition for a perfect matching. Moreover, we can repeatedly delete  $p^{3/2}\theta c'n$  perfect matchings in  $G[H'(x)]$ , as this maintains all degrees  $(1 \pm 2\sqrt{p})p\theta tn$  and codegrees at least  $p\theta c'n/2$ .

The punchline is that for any edge  $e$  in  $G[H'(x)]$  there are whp at most  $2p^2n$  earlier choices of  $x'$  with  $e$  in  $G[H'(x')]$ , and the random choice of  $M_{x'}$  covers  $e$  with probability at most  $(p^{3/2}\theta c'n)^{-1}$ , so  $e$  is covered with probability at most  $2p^2n(p^{3/2}\theta c'n)^{-1} < p^{1/3}$ , say. Thus whp  $G[H'(x)]$  still has sufficient degree and codegree properties to find the perfect matchings described above, and the algorithm can be completed. Moreover, any edge of  $G[V_1]$  is covered with probability at most  $p^{1/3}$ , so whp we maintain good triangle regularity in  $G[V_1]$  and can proceed down the vortex.

## 7 Randomised Algebraic Construction

Here we sketch an alternative proof (via our method of Randomised Algebraic Construction from Keevash [2014]) of the same result as in the previous subsection, i.e. finding a triangle decomposition of a graph  $G$  with certain extendability properties and no space or divisibility obstruction. Our approach will be quite similar to that in Keevash [2015], except that we will illustrate the ‘cascade’ approach to absorption which is more useful for general designs.

As discussed above, we circumvent the difficulties in the basic method for absorption by introducing an algebraic structure with built-in absorbers. Let  $\pi : V(G) \rightarrow \mathbb{F}_{2^a} \setminus \{0\}$  be a uniformly random injection, where  $2^{a-2} < n \leq 2^{a-1}$ . Our absorber (which in this context we call the ‘template’) is defined as the set  $T$  of all triangles in  $G$  such that  $\pi(x) + \pi(y) + \pi(z) = 0$ . Clearly  $T$  consists of edge-disjoint triangles. We let  $G^* = \bigcup T$  be the underlying graph of the template and suppress  $\pi$ , imagining  $V(G)$  as a subset of  $\mathbb{F}_{2^a}$ .

Standard concentration arguments show that whp  $G \setminus G^*$  has the necessary properties to apply the nibble, so we can find a set  $N$  of edge-disjoint triangles in  $G \setminus G^*$  with leave  $L := (G \setminus G^*) \setminus \bigcup N$  of maximum degree  $c_1n$  (we use a similar hierarchy of very small parameters  $c_i$  as before). To absorb  $L$ , it is convenient to first ‘move the problem’ into the template: we apply a random greedy algorithm to cover  $L$  by a set  $M^c$  of edge-disjoint triangles, each of which has one edge in  $L$  and the other two edges in  $G^*$ . Thus some subgraph  $S$  of  $G^*$ , which we call the ‘spill’ has now been covered twice. The binomial heuristic discussed in the previous subsection applies to show that whp this algorithm is successful, and moreover  $S$  is suitably bounded. (To be precise, we also ensure that each edge of  $S$  belongs to a different triangle of  $T$ , and that the union  $S^*$  of all such triangles is  $c_2$ -bounded.)

The remaining task of the proof is to modify the current set of triangles to eliminate the problem with the spill. The overall plan is to find a ‘hole’ in the template that exactly matches the spill. This will consist of two sets of edge-disjoint triangles, namely  $M^o$  (outer set) and  $M^i$  (inner set), such that  $M^o \subseteq T$  and  $\bigcup M^o$  is the disjoint union of  $S$  and  $\bigcup M^i$ . Then replacing  $M^o$  by  $M^i$  will fix the problem: formally, our final triangle decomposition of  $G$  is  $M := N \cup M^c \cup (T \setminus M^o) \cup M^i$ .

We break down the task of finding the hole into several steps. The first is a refined form of the integral decomposition theorem of Graver and Jurkat [1973] and R. M. Wilson [1973], i.e. that there is an assignment of integers to triangles so that total weight of triangles on any edge  $e$  is 1 if  $e \in S$  or 0 otherwise. Our final hole can be viewed as such an assignment, in which a triangle  $f$  has weight 1 if  $f \in M^o$ ,  $-1$  if  $f \in M^i$ , or 0 otherwise. We intend to start from some assignment and repeatedly modify it by random greedy algorithms until it has the properties required for the hole. As discussed above, the success of such random greedy algorithms requires control on the maximum degree, so our refined version of Graver and Jurkat [1973] and R. M. Wilson [1973] is that we can choose the weights  $w_T$  on triangles with  $\sum_{T:v \in T} |w_T| < c_3 n$  for every vertex  $v$ . (The proof is fairly simple, but the analogous statement for general hypergraphs seems to be much harder to prove.) Note that in this step we allow the use of any triangle in  $K_n$  (the complete graph on  $V(G)$ ), without considering whether they belong to  $G^*$ : ‘illegal’ triangles will be eliminated later.

Let us now consider how to modify assignments of weights to triangles so as to obtain a hole. Our first step is to ignore the requirement  $M^o \subseteq T$ , which makes our task much easier, as  $T$  is a special set of only  $O(n^2)$  triangles. Thus we seek a signed decomposition of  $S$  within  $G^*$ , i.e. an assignment from  $\{-1, 0, 1\}$  to each triangle of  $G^*$  so that the total weight on any  $e$  is 1 if  $e \in S$  or 0 otherwise, and every edge appears in at most one triangle of each sign.

To achieve this, we start from the simple observation that the graph of the octahedron has 8 triangles, which can be split into two groups of 4, each forming a triangle decomposition. For any copy of the octahedron in  $K_n$  we can add 1 to the triangles of one decomposition and subtract 1 from the triangles of the other without affecting the total weight of triangles on any edge. We can use this construction to repeatedly eliminate ‘cancelling pairs’, consisting of two triangles on a common edge with opposite sign. (There is a preprocessing step to ensure that each triangle to be eliminated can be assigned to a unique such pair.) In particular, as edges not in  $G^*$  have weight 0, this will eliminate all illegal triangles. The boundedness condition facilitates a random greedy algorithm for choosing edge-disjoint octahedra for these eliminations, which constructs the desired signed decomposition of  $S$ .

Now we remember that we wanted the outer triangle decomposition  $M^o$  to be contained in the template  $T$ . Finally, the algebraic structure will come into play, in absorbing the set



$M^+$  of positive triangles in the signed decomposition. To see how this can be achieved, consider any positive triangle  $xyz$ , recall that vertices are labelled by elements of  $\mathbb{F}_{2^d} \setminus \{0\}$ , and suppose first for simplicity that  $xyz$  is ‘octahedral’, meaning that  $G^*$  contains the ‘associated octahedron’ of  $xyz$ , defined as the complete 3-partite graph  $O$  with parts  $\{x, y + z\}$ ,  $\{y, z + x\}$ ,  $\{z, x + y\}$ . Then  $xyz$  is a triangle of  $O$ , and we note that  $O$  has a triangle decomposition consisting entirely of template triangles, namely  $\{x, y, x + y\}$ ,  $\{y + z, y, z\}$ ,  $\{x, z + x, z\}$  and  $\{y + z, z + x, x + y\}$ . Thus we can ‘flip’  $O$  (i.e. add and subtract the two triangle decompositions as before) to eliminate  $xyz$  while only introducing positive triangles that are in  $T$ .

The approach taken in [Keevash \[2015\]](#) was to ensure in the signed decomposition that every positive triangle is octahedral, with edge-disjoint associated octahedra, so that all positive triangles can be absorbed as indicated above without interfering with each other. For general designs, it is more convenient to define a wider class of triangles (in general hypergraph cliques) that can be absorbed by the following two step process, which we call a ‘cascade’. Suppose that we want to absorb some positive triangle  $xyz$ . We look for some octahedron  $O$  with parts  $\{x, x'\}$ ,  $\{y, y'\}$ ,  $\{z, z'\}$  such that each of the 4 triangles of the decomposition not using  $xyz$  is octahedral. We can flip the associated octahedra of these triangles so as to include them in the template, and now  $O$  is decomposed by template triangles, so can play the role of an associated octahedron for  $xyz$ : we can flip it to absorb  $xyz$ . The advantage of this approach is that whp any non-template  $xyz$  has many cascades, so no extra property of the signed decomposition is required to complete the proof. In general, there are still some conditions required for a clique have many cascades, but these are not difficult to ensure in the signed decomposition.

## 8 Concluding remarks

There are many other questions of Design Theory that can be reformulated as asking whether a certain (sparse) hypergraph has a perfect matching. This suggests the (vague) meta-question of formulating and proving a general theorem on the existence of perfect matchings in sparse ‘design-like’ hypergraphs (for some ‘natural’ definition of ‘design-like’ that is sufficiently general to capture a variety of problems in Design Theory). One test for such a statement is that it should capture all variant forms of the basic existence question, such as general hypergraph decompositions (as in [Glock, Kühn, Lo, and Osthus \[2016\]](#)) or resolvable designs (the general form of Kirkman’s original ‘schoolgirl problem’, solved for graphs by [Ray-Chaudhuri and R. M. Wilson \[1971\]](#) but still open for hypergraphs). But could we be even more ambitious?

To focus the ideas, one well-known longstanding open problem is Ryser’s Conjecture [Ryser \[1967\]](#) that every Latin square of odd order has a transversal. (A generalised form

of this conjecture by [Stein \[1975\]](#) was recently disproved by [Pokrovskiy and Sudakov \[2017\]](#).) To see the connection with hypergraph matchings, we associate to any Latin square a tripartite 3-graph in which the parts correspond to rows, columns and symbols, and each cell of the square corresponds to an edge consisting of its own row, column and symbol. A perfect matching in this 3-graph is precisely a transversal of the Latin square. However, there is no obvious common structure to the various possible 3-graphs that may arise in this way, which presents a challenge to the absorbing methods described in this article, and so to formulating a meta-theorem that might apply to Ryser’s Conjecture. The best known lower bound of  $n - O(\log^2 n)$  on a partial transversal (by [Hatami and Shor \[2008\]](#)) has a rather different proof. Another generalisation of Ryser’s Conjecture by [Aharoni and Berger \[2009\]](#) concerning rainbow matchings in properly coloured multigraphs has recently motivated the development of various other methods for such problems not discussed in this article (see e.g. [Gao, Ramadurai, Wanless, and Wormald \[2017\]](#), [Keevash and Yepremyan \[2017\]](#), and [Pokrovskiy \[2016\]](#)).

Recalling the theme of random matchings discussed in the introduction, it is unsurprising that it is hard to say much about random designs, but for certain applications one can extract enough from the proof in [Keevash \[2014\]](#), e.g. to show that whp a random Steiner Triple System has a perfect matching ([Kwan \[2016\]](#)) or that one can superimpose a constant number of Steiner Systems to obtain a bounded codegree high-dimensional expander ([Lubotzky, Luria, and Rosenthal \[2015\]](#)). Does the nascent connection between hypergraph matchings and high-dimensional expanders go deeper?

We conclude by recalling two longstanding open problems from the other end of the Design Theory spectrum, concerning  $q$ -graphs with  $q$  of order  $\sqrt{n}$  (the maximum possible), as opposed to the setting  $n > n_0(q)$  considered in this article (or even the methods of [Kuperberg, Lovett, and Peled \[2017\]](#) which can allow  $q$  to grow as a sufficiently small power of  $n$ ).

### **Hadamard’s Conjecture.** ([Hadamard \[1893\]](#))

There is an  $n \times n$  orthogonal matrix  $H$  with all entries  $\pm n^{-1/2}$  iff  $n$  is 1, 2 or divisible by 4?

### **Projective Plane Prime Power Conjecture.** (folklore)

There is a Steiner system with parameters  $(k^2 + k + 1, k + 1, 2)$  iff  $k$  is a prime power?

## **References**

- Ron Aharoni (2001). “Ryser’s conjecture for tripartite 3-graphs”. *Combinatorica* 21.1, pp. 1–4. MR: [1805710](#) (cit. on p. 3132).
- Ron Aharoni and Eli Berger (2009). “Rainbow matchings in  $r$ -partite  $r$ -graphs”. *Electron. J. Combin.* 16.1, Research Paper 119, 9. MR: [2546322](#) (cit. on p. 3147).

- Ron Aharoni and Penny Haxell (2000). “Hall’s theorem for hypergraphs”. *J. Graph Theory* 35.2, pp. 83–88. MR: [1781189](#) (cit. on p. [3132](#)).
- Peter Allen, Julia Böttcher, Hiep Hàn, Yoshiharu Kohayakawa, and Yury Person (Dec. 2016). “Blow-up lemmas for sparse graphs”. arXiv: [1612.00622](#) (cit. on p. [3137](#)).
- Noga Alon, Jeong Han Kim, and Joel Spencer (1997). “Nearly perfect matchings in regular simple hypergraphs”. *Israel J. Math.* 100, pp. 171–187. MR: [1469109](#) (cit. on p. [3136](#)).
- Arash Asadpour, Uriel Feige, and Amin Saberi (2008). “Santa Claus meets hyper-graph matchings”. In: *Approximation, randomization and combinatorial optimization*. Vol. 5171. Lecture Notes in Comput. Sci. Springer, Berlin, pp. 10–20. MR: [2538773](#) (cit. on p. [3132](#)).
- Ben Barber, Daniela Kühn, Allan Lo, Richard Montgomery, and Deryk Osthus (2017). “Fractional clique decompositions of dense graphs and hypergraphs”. *J. Combin. Theory Ser. B* 127, pp. 148–186. MR: [3704659](#) (cit. on p. [3136](#)).
- Ben Barber, Daniela Kühn, Allan Lo, and Deryk Osthus (2016). “Edge-decompositions of graphs with high minimum degree”. *Adv. Math.* 288, pp. 337–385. MR: [3436388](#) (cit. on pp. [3136](#), [3141](#)).
- Alexander Barvinok and Alex Samorodnitsky (2011). “Computing the partition function for perfect matchings in a hypergraph”. *Combin. Probab. Comput.* 20.6, pp. 815–835. MR: [2847269](#) (cit. on p. [3133](#)).
- Patrick Bennett and Tom Bohman (Oct. 2012). “A natural barrier in random greedy hyper-graph matching”. arXiv: [1210.3581](#) (cit. on pp. [3136](#), [3139](#)).
- Tom Bohman (2009). “The triangle-free process”. *Adv. Math.* 221.5, pp. 1653–1677. MR: [2522430](#) (cit. on p. [3136](#)).
- Tom Bohman, Alan Frieze, and Eyal Lubetzky (2015). “Random triangle removal”. *Adv. Math.* 280, pp. 379–438. MR: [3350225](#) (cit. on p. [3136](#)).
- Tom Bohman and Peter Keevash (2010). “The early evolution of the  $H$ -free process”. *Invent. Math.* 181.2, pp. 291–336. MR: [2657427](#) (cit. on p. [3136](#)).
- (2013). “Dynamic concentration of the triangle-free process”. In: *The Seventh European Conference on Combinatorics, Graph Theory and Applications*. Vol. 16. CRM Series. Ed. Norm., Pisa, pp. 489–495. arXiv: [1302.5963](#). MR: [3185850](#) (cit. on pp. [3136](#), [3142](#)).
- Charles J. Colbourn and Jeffrey H. Dinitz, eds. (2007). *Handbook of combinatorial designs*. Second. Discrete Mathematics and its Applications (Boca Raton). Chapman & Hall/CRC, Boca Raton, FL, pp. xxii+984. MR: [2246267](#) (cit. on p. [3138](#)).
- Colin Cooper, Alan Frieze, Michael Molloy, and Bruce Reed (1996). “Perfect matchings in random  $r$ -regular,  $s$ -uniform hypergraphs”. *Combin. Probab. Comput.* 5.1, pp. 1–14. MR: [1395689](#) (cit. on p. [3133](#)).

- G rard Cornu jols (2001). [Combinatorial optimization](#). Vol. 74. CBMS-NSF Regional Conference Series in Applied Mathematics. Packing and covering. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, pp. xii+132. MR: [1828452](#) (cit. on pp. [3132](#), [3136](#)).
- Ewan Davies, Matthew Jenssen, Will Perkins, and Barnaby Roberts (Apr. 2017). [“Tight bounds on the coefficients of partition functions via stability”](#). arXiv: [1704.07784](#) (cit. on p. [3133](#)).
- Pat Devlin and Jeff Kahn (2017). “Perfect fractional matchings in  $k$ -out hypergraphs”. *Electron. J. Combin.* 24.3, Paper 3.60, 12. MR: [3711102](#) (cit. on p. [3136](#)).
- G. A. Dirac (1952). [“Some theorems on abstract graphs”](#). *Proc. London Math. Soc.* (3) 2, pp. 69–81. MR: [0047308](#) (cit. on p. [3132](#)).
- Fran ois Dross (2016). [“Fractional triangle decompositions in graphs with large minimum degree”](#). *SIAM J. Discrete Math.* 30.1, pp. 36–42. MR: [3440184](#) (cit. on p. [3136](#)).
- Jack Edmonds (1965). [“Paths, trees, and flowers”](#). *Canad. J. Math.* 17, pp. 449–467. MR: [0177907](#) (cit. on p. [3131](#)).
- Jeno Egerv ry (1931). “Matrixok kombinatorius tulajdons gair l”. *Matematikai  s Fizikai Lapok* 38.1931, pp. 16–28 (cit. on p. [3131](#)).
- P. Erd s (1965). “A problem on independent  $r$ -tuples”. *Ann. Univ. Sci. Budapest. E tv s Sect. Math.* 8, pp. 93–95. MR: [0260599](#) (cit. on pp. [3132](#), [3136](#)).
- (1981). [“On the combinatorial problems which I would most like to see solved”](#). *Combinatorica* 1.1, pp. 25–42. MR: [602413](#) (cit. on p. [3133](#)).
- P. Erd s and H. Hanani (1963). “On a limit theorem in combinatorial analysis”. *Publ. Math. Debrecen* 10, pp. 10–13. MR: [0166116](#) (cit. on p. [3136](#)).
- Asaf Ferber, Rani Hod, Michael Krivelevich, and Benny Sudakov (2014). [“A construction of almost Steiner systems”](#). *J. Combin. Des.* 22.11, pp. 488–494. MR: [3263662](#) (cit. on pp. [3138](#), [3143](#)).
- P. Frankl and V. R dl (1985). [“Near perfect coverings in graphs and hypergraphs”](#). *European J. Combin.* 6.4, pp. 317–326. MR: [829351](#) (cit. on p. [3136](#)).
- Peter Frankl and Norihide Tokushige (2016). [“Invitation to intersection problems for finite sets”](#). *J. Combin. Theory Ser. A* 144, pp. 157–211. MR: [3534067](#) (cit. on p. [3132](#)).
- S. Friedland, E. Krop, and K. Markstr m (2008). [“On the number of matchings in regular graphs”](#). *Electron. J. Combin.* 15.1, Research Paper 110, 28. MR: [2438582](#) (cit. on p. [3133](#)).
- Zolt n F redi (1988). [“Matchings and covers in hypergraphs”](#). *Graphs Combin.* 4.2, pp. 115–206. MR: [943753](#) (cit. on p. [3132](#)).
- D. Gale and L. S. Shapley (1962). [“College Admissions and the Stability of Marriage”](#). *Amer. Math. Monthly* 69.1, pp. 9–15. MR: [1531503](#) (cit. on p. [3131](#)).
- Pu Gao, Reshma Ramadurai, Ian Wanless, and Nick Wormald (Sept. 2017). [“Full rainbow matchings in graphs and hypergraphs”](#). arXiv: [1709.02665](#) (cit. on p. [3147](#)).

- Stefan Glock, Daniela Kühn, Allan Lo, and Deryk Osthus (Nov. 2016). “The existence of designs via iterative absorption”. arXiv: [1611.06827](#) (cit. on pp. [3139](#), [3141–3143](#), [3146](#)).
- (June 2017). “Hypergraph  $F$ -designs for arbitrary  $F$ ”. arXiv: [1706.01800](#) (cit. on pp. [3139](#), [3141](#)).
- W. T. Gowers (2007). “Hypergraph regularity and the multidimensional Szemerédi theorem”. *Ann. of Math. (2)* 166.3, pp. 897–946. MR: [2373376](#) (cit. on p. [3137](#)).
- David A. Grable (1999). “More-than-nearly-perfect packings and partial designs”. *Combinatorica* 19.2, pp. 221–239. MR: [1723040](#) (cit. on p. [3136](#)).
- J. E. Graver and W. B. Jurkat (1973). “The module structure of integral designs”. *J. Combinatorial Theory Ser. A* 15, pp. 75–90. MR: [0329930](#) (cit. on pp. [3138](#), [3145](#)).
- Jacques Hadamard (1893). “Resolution d’une question relative aux déterminants”. *Bull. des sciences math.* 17, pp. 240–246 (cit. on p. [3147](#)).
- Philip Hall (1935). “On representatives of subsets”. *Journal of the London Mathematical Society* 10, pp. 26–30 (cit. on p. [3131](#)).
- Jie Han (2015). “Near perfect matchings in  $k$ -uniform hypergraphs”. *Combin. Probab. Comput.* 24.5, pp. 723–732. MR: [3371500](#) (cit. on p. [3135](#)).
- (2017). “Decision problem for perfect matchings in dense  $k$ -uniform hypergraphs”. *Trans. Amer. Math. Soc.* 369.7, pp. 5197–5218. MR: [3632565](#) (cit. on p. [3134](#)).
- Pooya Hatami and Peter W. Shor (2008). “A lower bound for the length of a partial transversal in a Latin square”. *J. Combin. Theory Ser. A* 115.7, pp. 1103–1113. MR: [2450332](#) (cit. on p. [3147](#)).
- Penny Haxell (2016). “Independent transversals and hypergraph matchings—an elementary approach”. In: *Recent trends in combinatorics*. Vol. 159. IMA Vol. Math. Appl. Springer, [Cham], pp. 215–233. MR: [3526410](#) (cit. on p. [3132](#)).
- Penny Haxell and V. Rödl (2001). “Integer and fractional packings in dense graphs”. *Combinatorica* 21.1, pp. 13–38. MR: [1805712](#) (cit. on p. [3137](#)).
- John Robert Henderson (1971). *Permutation decompositions of  $(0,1)$ -matrices and decomposition transversals*. Thesis (Ph.D.)—California Institute of Technology. ProQuest LLC, Ann Arbor, MI, p. 60. MR: [2620362](#) (cit. on p. [3132](#)).
- Mark Jerrum (2003). *Counting, sampling and integrating: algorithms and complexity*. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, pp. xii+112. MR: [1960003](#) (cit. on p. [3133](#)).
- Anders Johansson, Jeff Kahn, and Van H. Vu (2008). “Factors in random graphs”. *Random Structures Algorithms* 33.1, pp. 1–28. MR: [2428975](#) (cit. on p. [3133](#)).
- Jeff Kahn (1996a). “A linear programming perspective on the Frankl–Rödl–Pippenger theorem”. *Random Structures Algorithms* 8.2, pp. 149–157. MR: [1607104](#) (cit. on pp. [3136](#), [3142](#)).

- (1996b). “Asymptotically good list-colorings”. *J. Combin. Theory Ser. A* 73.1, pp. 1–59. MR: [1367606](#) (cit. on p. [3136](#)).
- (2000). “A normal law for matchings”. *Combinatorica* 20.3, pp. 339–391. MR: [1774843](#) (cit. on p. [3133](#)).
- Jeff Kahn and P. Mark Kayll (1997). “On the stochastic independence properties of hardcore distributions”. *Combinatorica* 17.3, pp. 369–391. MR: [1606040](#) (cit. on p. [3133](#)).
- Richard M. Karp (1972). “Reducibility among combinatorial problems”, pp. 85–103. MR: [0378476](#) (cit. on p. [3131](#)).
- Peter Keevash (2011a). “A hypergraph blow-up lemma”. *Random Structures Algorithms* 39.3, pp. 275–376. MR: [2816936](#) (cit. on p. [3137](#)).
- (2011b). “Hypergraph Turán problems”. In: *Surveys in combinatorics 2011*. Vol. 392. London Math. Soc. Lecture Note Ser. Cambridge Univ. Press, Cambridge, pp. 83–139. MR: [2866732](#) (cit. on p. [3134](#)).
- (Jan. 2014). “The existence of designs”. arXiv: [1401.3665](#) (cit. on pp. [3138](#), [3139](#), [3141](#), [3144](#), [3147](#)).
- (Apr. 2015). “Counting designs”. to appear in *J. Eur. Math. Soc.* arXiv: [1504.02909](#) (cit. on pp. [3138](#), [3144](#), [3146](#)).
- Peter Keevash, Fiachra Knox, and Richard Mycroft (2015). “Polynomial-time perfect matchings in dense hypergraphs”. *Adv. Math.* 269, pp. 265–334. MR: [3281137](#) (cit. on pp. [3134](#), [3135](#)).
- Peter Keevash and Richard Mycroft (2015a). “A geometric theory for hypergraph matching”. *Mem. Amer. Math. Soc.* 233.1098, pp. vi+95. MR: [3290271](#) (cit. on pp. [3133](#), [3134](#), [3137](#), [3139](#)).
- (2015b). “A multipartite Hajnal-Szemerédi theorem”. *J. Combin. Theory Ser. B* 114, pp. 187–236. MR: [3354296](#) (cit. on p. [3134](#)).
- Peter Keevash and Liana Yepremyan (Oct. 2017). “Rainbow matchings in properly-coloured multigraphs”. arXiv: [1710.03041](#) (cit. on p. [3147](#)).
- R. Kenyon (2010). “The dimer model”. In: *Exact methods in low-dimensional statistical physics and quantum computing*. Oxford Univ. Press, Oxford, pp. 341–361. MR: [2668650](#) (cit. on p. [3133](#)).
- Jaehoon Kim, Daniela Kühn, Deryk Osthus, and Mykhaylo Tyomkyn (Apr. 2016). “A blow-up lemma for approximate decompositions”. to appear in *Trans. Amer. Math. Soc.* arXiv: [1604.07282](#) (cit. on p. [3137](#)).
- Jeong Han Kim (2001). “Nearly optimal partial Steiner systems”. In: *Brazilian Symposium on Graphs, Algorithms and Combinatorics*. Vol. 7. Electron. Notes Discrete Math. Elsevier Sci. B. V., Amsterdam, p. 4. MR: [2154563](#) (cit. on p. [3136](#)).
- János Komlós, Gábor N. Sárközy, and Endre Szemerédi (1997). “Blow-up lemma”. *Combinatorica* 17.1, pp. 109–123. MR: [1466579](#) (cit. on p. [3137](#)).

- D. König (1931). “Gráfok és mátrixok”. *Matematikai és Fizikai Lapok* 38, pp. 116–119 (cit. on p. [3131](#)).
- A. V. Kostochka and V. Rödl (1998). “Partial Steiner systems and matchings in hypergraphs”. In: *Proceedings of the Eighth International Conference “Random Structures and Algorithms” (Poznan, 1997)*. Vol. 13. 3-4, pp. 335–347. MR: [1662789](#) (cit. on p. [3136](#)).
- Michael Krivelevich (1996). “Perfect fractional matchings in random hypergraphs”. *Random Structures Algorithms* 9.3, pp. 317–334. MR: [1606849](#) (cit. on p. [3136](#)).
- Daniela Kühn and Deryk Osthus (2009). “Embedding large subgraphs into dense graphs”. In: *Surveys in combinatorics 2009*. Vol. 365. London Math. Soc. Lecture Note Ser. Cambridge Univ. Press, Cambridge, pp. 137–167. MR: [2588541](#) (cit. on p. [3132](#)).
- Daniela Kühn and Deryk Osthus (2014). “Hamilton cycles in graphs and hypergraphs: an extremal perspective”. In: *Proc. Intern. Congress of Math, Seoul, Korea*. Vol. 4, pp. 381–406 (cit. on pp. [3132](#), [3141](#)).
- H. W. Kuhn (1955). “The Hungarian method for the assignment problem”. *Naval Res. Logist. Quart.* 2, pp. 83–97. MR: [0075510](#) (cit. on p. [3131](#)).
- Greg Kuperberg, Shachar Lovett, and Ron Peled (2017). “Probabilistic existence of regular combinatorial structures”. *Geom. Funct. Anal.* 27.4, pp. 919–972. MR: [3678505](#) (cit. on pp. [3138](#), [3147](#)).
- Nikolai N. Kuzjurin (1995). “On the difference between asymptotically good packings and coverings”. *European J. Combin.* 16.1, pp. 35–40. MR: [1317200](#) (cit. on p. [3136](#)).
- Matthew Kwan (Nov. 2016). “Almost all Steiner triple systems have perfect matchings”. arXiv: [1611.02246](#) (cit. on p. [3147](#)).
- Lap Chi Lau, R. Ravi, and Mohit Singh (2011). *Iterative methods in combinatorial optimization*. Cambridge Texts in Applied Mathematics. Cambridge University Press, New York, pp. xii+242. MR: [2808916](#) (cit. on pp. [3132](#), [3137](#)).
- László Lovász (1975). “On minimax theorems of combinatorics”. *Mat. Lapok* 26.3-4, 209–264 (1978). MR: [510823](#) (cit. on p. [3132](#)).
- László Lovász and Michael D. Plummer (2009). *Matching theory*. Corrected reprint of the 1986 original [MR0859549]. AMS Chelsea Publishing, Providence, RI, pp. xxxiv+554. MR: [2536865](#) (cit. on p. [3131](#)).
- Alexander Lubotzky, Zur Luria, and Ron Rosenthal (Dec. 2015). “Random Steiner systems and bounded degree coboundary expanders of every dimension”. arXiv: [1512.08331](#) (cit. on p. [3147](#)).
- Zur Luria (May 2017). “New bounds on the number of n-queens configurations”. arXiv: [1705.05225](#) (cit. on p. [3138](#)).
- Brendan Nagle, Vojtěch Rödl, and Mathias Schacht (2006). “The counting lemma for regular  $k$ -uniform hypergraphs”. *Random Structures Algorithms* 28.2, pp. 113–179. MR: [2198495](#) (cit. on p. [3137](#)).



- C St JA Nash-Williams (1970). “An unsolved problem concerning decomposition of graphs into triangles”. *Combinatorial theory and its applications, North Holland 3*, pp. 1179–1183 (cit. on p. [3136](#)).
- Nicholas Pippenger and Joel Spencer (1989). “Asymptotic behavior of the chromatic index for hypergraphs”. *J. Combin. Theory Ser. A* 51.1, pp. 24–42. MR: [993646](#) (cit. on pp. [3136](#), [3141](#)).
- Alexey Pokrovskiy (Sept. 2016). “An approximate version of a conjecture of Aharoni and Berger”. arXiv: [1609.06346](#) (cit. on p. [3147](#)).
- Alexey Pokrovskiy and Benny Sudakov (Nov. 2017). “A counterexample to Stein’s Equi- $n$ -square Conjecture”. arXiv: [1711.00429](#) (cit. on p. [3147](#)).
- Gonzalo Fiz Pontiveros, Simon Griffiths, and Robert Morris (Feb. 2013). “The triangle-free process and  $R(3,k)$ ”. arXiv: [1302.6279](#) (cit. on pp. [3136](#), [3142](#)).
- Jaikumar Radhakrishnan (1997). “An entropy proof of Bregman’s theorem”. *J. Combin. Theory Ser. A* 77.1, pp. 161–164. MR: [1426744](#) (cit. on p. [3138](#)).
- D. K. Ray-Chaudhuri and Richard M. Wilson (1971). “Solution of Kirkman’s schoolgirl problem”, pp. 187–203. MR: [0314644](#) (cit. on p. [3146](#)).
- V. Rödl, M. Schacht, M. H. Siggers, and N. Tokushige (2007). “Integer and fractional packings of hypergraphs”. *J. Combin. Theory Ser. B* 97.2, pp. 245–268. MR: [2290324](#) (cit. on p. [3137](#)).
- Vojtěch Rödl (1985). “On a packing and covering problem”. *European J. Combin.* 6.1, pp. 69–78. MR: [793489](#) (cit. on p. [3136](#)).
- Vojtech Rödl and Andrzej Ruciński (2010). “Dirac-type questions for hypergraphs—a survey (or more problems for Endre to solve)”. In: *An irregular mind*. Vol. 21. Bolyai Soc. Math. Stud. János Bolyai Math. Soc., Budapest, pp. 561–590. MR: [2815614](#) (cit. on p. [3132](#)).
- Vojtech Rödl, Andrzej Ruciński, and Endre Szemerédi (2009). “Perfect matchings in large uniform hypergraphs with large minimum collective degree”. *J. Combin. Theory Ser. A* 116.3, pp. 613–636. MR: [2500161](#) (cit. on pp. [3132](#), [3134](#), [3135](#), [3139](#), [3140](#)).
- Vojtěch Rödl and Mathias Schacht (2007a). “Regular partitions of hypergraphs: counting lemmas”. *Combin. Probab. Comput.* 16.6, pp. 887–901. MR: [2351689](#) (cit. on p. [3137](#)).
- (2007b). “Regular partitions of hypergraphs: regularity lemmas”. *Combin. Probab. Comput.* 16.6, pp. 833–885. MR: [2351688](#) (cit. on pp. [3137](#), [3142](#)).
- Vojtěch Rödl and Jozef Skokan (2004). “Regularity lemma for  $k$ -uniform hypergraphs”. *Random Structures Algorithms* 25.1, pp. 1–42. MR: [2069663](#) (cit. on p. [3137](#)).
- Herbert J Ryser (1967). “Neuere probleme der kombinatorik”. *Vorträge über Kombinatorik, Oberwolfach*, pp. 69–91 (cit. on p. [3146](#)).
- Alexander Schrijver (2003). *Combinatorial optimization: polyhedra and efficiency*. Springer-Verlag (cit. on pp. [3131](#), [3136](#)).



- Joel Spencer (1995). “Asymptotic packing via a branching process”. *Random Structures Algorithms* 7.2, pp. 167–172. MR: [1369062](#) (cit. on p. [3136](#)).
- S. K. Stein (1975). “Transversals of Latin squares and their generalizations”. *Pacific J. Math.* 59.2, pp. 567–575. MR: [0387083](#) (cit. on p. [3147](#)).
- E Steinitz (1894). “Über die Konstruktion der Configurationen  $n$  (sub 3)”. PhD thesis. Ph.D. thesis, Universität Breslau (cit. on p. [3131](#)).
- Endre Szemerédi (1978). “Regular partitions of graphs”. In: *Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976)*. Vol. 260. Colloq. Internat. CNRS. CNRS, Paris, pp. 399–401. MR: [540024](#) (cit. on p. [3137](#)).
- Luc Teirlinck (1987). “Nontrivial  $t$ -designs without repeated blocks exist for all  $t$ ”. *Discrete Math.* 65.3, pp. 301–311. MR: [897654](#) (cit. on p. [3138](#)).
- Van H. Vu (2000). “New bounds on nearly perfect matchings in hypergraphs: higher code-greedy do help”. *Random Structures Algorithms* 17.1, pp. 29–63. MR: [1768848](#) (cit. on p. [3136](#)).
- Michelle L. Wachs (2003). “Topology of matching, chessboard, and general bounded degree graph complexes”. *Algebra Universalis* 49.4. Dedicated to the memory of Gian-Carlo Rota, pp. 345–385. MR: [2022345](#) (cit. on p. [3132](#)).
- David P. Williamson and David B. Shmoys (2011). *The design of approximation algorithms*. Cambridge University Press, Cambridge, pp. xii+504. MR: [2798112](#) (cit. on p. [3132](#)).
- Richard M. Wilson (1972a). “An existence theory for pairwise balanced designs. I. Composition theorems and morphisms”. *J. Combinatorial Theory Ser. A* 13, pp. 220–245. MR: [0304203](#) (cit. on p. [3138](#)).
- (1972b). “An existence theory for pairwise balanced designs. II. The structure of PBD-closed sets and the existence conjectures”. *J. Combinatorial Theory Ser. A* 13, pp. 246–273. MR: [0304204](#) (cit. on p. [3138](#)).
- (1973). “The necessary conditions for  $t$ -designs are sufficient for something”. *Utilitas Math.* 4, pp. 207–215. MR: [0325415](#) (cit. on pp. [3138](#), [3145](#)).
- (1975). “An existence theory for pairwise balanced designs. III. Proof of the existence conjectures”. *J. Combinatorial Theory Ser. A* 18, pp. 71–79. MR: [0366695](#) (cit. on p. [3138](#)).
- (1999). “Signed hypergraph designs and diagonal forms for some incidence matrices”. *Des. Codes Cryptogr.* 17.1-3, pp. 289–297. MR: [1715268](#) (cit. on p. [3138](#)).
- (1973/74). “Nonisomorphic Steiner triple systems”. *Math. Z.* 135, pp. 303–313. MR: [0340046](#).
- Robin Wilson (2003). “The early history of block designs”. *Rend. Sem. Mat. Messina Ser. II* 9(25), 267–276 (2004). MR: [2121481](#) (cit. on p. [3137](#)).
- Raphael Yuster (2007). “Combinatorial and computational aspects of graph packing and graph decomposition”. *Computer Science Review* 1, pp. 12–26 (cit. on p. [3132](#)).

- Yi Zhao (2016). “[Recent advances on Dirac-type problems for hypergraphs](#)”. In: *Recent trends in combinatorics*. Vol. 159. IMA Vol. Math. Appl. Springer, [Cham], pp. 145–165. MR: [3526407](#) (cit. on p. [3132](#)).
- Yufei Zhao (2017). “[Extremal regular graphs: independent sets and graph homomorphisms](#)”. *Amer. Math. Monthly* 124.9, pp. 827–843. arXiv: [1610 . 09210](#). MR: [3722040](#) (cit. on p. [3133](#)).

Received 2017-11-27.

PETER KEEVASH  
MATHEMATICAL INSTITUTE  
UNIVERSITY OF OXFORD  
OXFORD  
UK  
[Peter.Keevash@maths.ox.ac.uk](mailto:Peter.Keevash@maths.ox.ac.uk)  
[keevash@maths.ox.ac.uk](mailto:keevash@maths.ox.ac.uk)



## LIMIT SHAPES AND THEIR ANALYTIC PARAMETERIZATIONS

RICHARD KENYON

### Abstract

A “limit shape” is a form of the law of large numbers, and happens when a large random system, typically consisting of many interacting particles, can be described, after an appropriate normalization, by a certain nonrandom object. Limit shapes occur in, for example, random integer partitions or in random interface models such as the dimer model. Typically limit shapes can be described by some variational formula based on a large deviations estimate. We discuss limit shapes for certain 2-dimensional interface models, and explain how their underlying analytic structure is related to a (conjectural in some cases) conformal invariance property for the models.

### 1 Limit shapes: integer partitions

We illustrate the notion of limit shape with a fundamental example. Given a uniform random integer partition  $\lambda$  of  $n$  for  $n$  large, a theorem of [Vershik and Kerov \[1981\]](#) asserts that, when both axes are scaled by  $\sqrt{n}$ , the graph of  $\lambda$  (that is, the Young diagram associated to  $\lambda$ ) converges with probability tending to 1 to a *nonrandom* curve, given by the equation  $e^{-cx} + e^{-cy} = 1$ , with  $c = \sqrt{\pi^2/6}$ , see [Figure 1](#). This is an example (in fact, one of the first examples) of a *limit shape theorem*: in the limit of large system size, the typical random object will, when appropriately scaled, concentrate on a fixed nonrandom shape. One way to make a more precise formulation of this statement is say that for each  $n$ , the random partition of  $n$  defines a certain probability measure  $\mu_n$  (on the space of non-increasing functions  $f : [0, \infty) \rightarrow [0, \infty)$  of integral 1) and as  $n \rightarrow \infty$  this sequence of measures converges in probability<sup>1</sup> to a point mass on the Vershik-Kerov curve.

---

MSC2010: 82B20.

<sup>1</sup> The topology of convergence for the sequence of random functions can be taken to be uniform convergence on compact subsets of  $(0, \infty)$ .

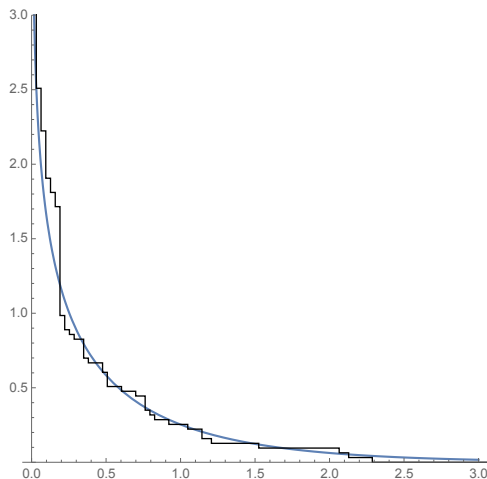


Figure 1: The Vershik-Kerov curve and a (scaled) random partition of 1000.

## 2 Lozenge tilings

There are many other limit shape theorems for random objects. In this talk we will discuss certain two-dimensional generalizations of the Kerov-Vershik result above. Our objects are *lozenge tilings*. A lozenge tiling is a tiling of a region in the plane with lozenges, which are  $60^\circ$  rhombi (in three possible orientations), see [Figure 2](#).

Equivalently, we can think of a lozenge tiling as a projection of a three-dimensional object, a *stepped surface*. A stepped surface is a piecewise linear surface in  $\mathbb{R}^3$  composed of squares from the  $\mathbb{Z}^3$  lattice, and which is monotone in the sense that it projects orthogonally injectively to the plane  $P_{111} = \{x + y + z = 0\}$ . The equivalence between lozenge tilings and stepped surfaces spanning an appropriate boundary is clear. Each lozenge tiling comes equipped with a *height function* which is the scaled distance of its stepped surface to the plane  $P_{111}$ .

Another way to view lozenge tilings is as monotone nonintersecting lattice paths (MNLPS), see [Figure 3](#).

**2.1 Limit shape theorem.** Given a polygonal domain in the plane which can be tiled with lozenges, what is the “shape” of a typical tiling? Here by *typical* we mean chosen uniformly at random from the set of all possible tilings of the region. The first such result was for a different but actually closely related model, domino tilings of the “Aztec diamond”, where Jockusch, Propp and Shor showed the existence of a certain circle in the

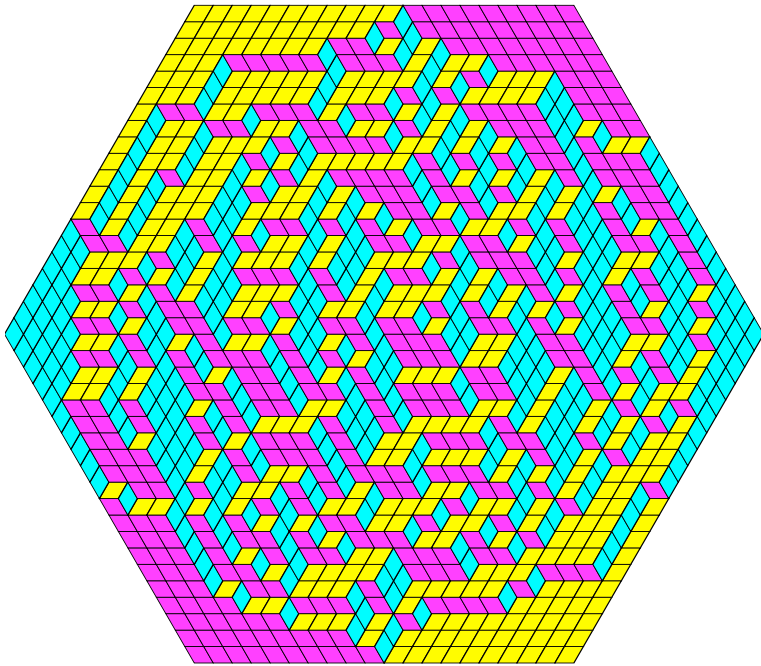


Figure 2: A (uniform) random lozenge tiling of a hexagon.

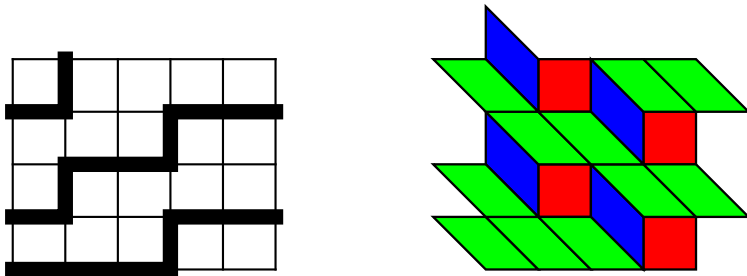


Figure 3: The lozenge tiling model is equivalent to the MNLP model. The lattice paths are the paths running through the centers of the blue and green lozenges (horizontally through the green and vertically through the blue).

limit shape [Jockusch, Propp, and Shor \[1998\]](#); this was extended to the limit shape itself by Cohn, Elkies and Propp in [Cohn, Elkies, and Propp \[1996\]](#). The first limit shape result for lozenge tilings is due to [Cohn, Larsen, and Propp \[1998\]](#) for the special case of the hexagonal boundary conditions of [Figure 2](#).

Henry Cohn, James Propp and myself proved [Cohn, Kenyon, and Propp \[2001\]](#) the following limit shape theorem for lozenges with general boundary conditions. Given a Jordan domain  $U \subseteq \mathbb{R}^2$  and a function  $u : \partial U \rightarrow \mathbb{R}$  satisfying a certain Lipschitz condition, there is a unique Lipschitz extension  $h : U \rightarrow \mathbb{R}$  (satisfying  $h|_{\partial U} = u$ ) which is the “limit shape for lozenge tilings with boundary values  $u$ ” in the following sense. Let  $U_n \subset \mathbb{R}^2$  be a polygonal domain, approximating  $U$ , which can be tiled by lozenges scaled by  $1/n$ , and such that the height function along the boundary of  $U_n$  approximates  $u$  as  $n \rightarrow \infty$ . Then for any  $\varepsilon > 0$ , for sufficiently large  $n$ , a uniform random lozenge tiling of  $U_n$  will have, with probability at least  $1 - \varepsilon$ , height function lying within  $\varepsilon$  of  $h$ . Moreover, there is a variational formula for the limiting surface  $h$ :  $h$  is the unique function minimizing the integral

$$(1) \quad \iint_U \sigma(\nabla h) \, dx \, dy$$

where  $\sigma$  is an explicit function (see [Figure 4](#)), the *surface tension*. Here for convenience

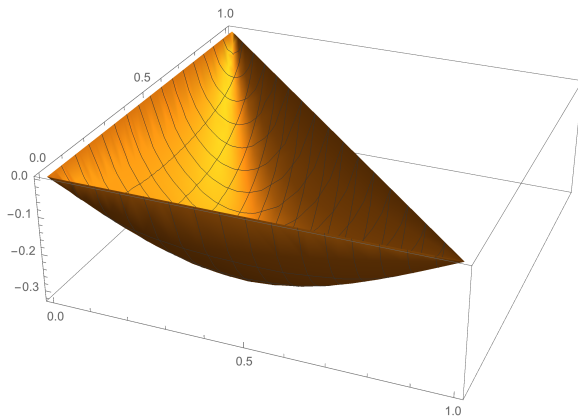


Figure 4: The surface tension function  $\sigma$  for the lozenge tiling model.

we parameterize points of  $P_{111}$  by their  $x$  and  $y$  coordinates, and rather than have  $h$  be the distance to the plane  $P_{111}$ , we define  $h(x, y)$  to be the  $z$  coordinate of the point

on the surface. In this case the gradient  $\nabla h = (h_x, h_y)$  lies in the unit triangle  $\mathfrak{N} = \text{conv}\{(0, 0), (1, 0), (0, 1)\}$ . The function  $\sigma$  is taken to be  $\infty$  when  $\nabla h$  is not in  $\mathfrak{N}$ ; this constrains  $h$  to have gradient in  $\mathfrak{N}$ , which it must if it arises as a limit of height functions for lozenge tilings. The Lipschitz condition on  $u$  referred to in the statement is that there exists an extension  $h$  with gradient in  $\mathfrak{N}$ .

The function  $\sigma = \sigma(s, t)$  itself is rather remarkable. If we take a triangle in  $\mathbb{C}$  with vertices  $0, 1, z$  with  $\text{Im } z > 0$  and angles at  $0$  and  $1$  equal to  $\pi s$  and  $\pi t$  respectively, then  $\sigma(s, t) = D(z)$ , where  $D$  is the *Bloch-Wigner dilogarithm*

$$D(z) = \frac{1}{\pi} (\arg(1 - z) \log |z| + \text{Im Li}(z)).$$

Here  $\text{Li}(z) = -\int_0^z \log(1 - \xi) \frac{d\xi}{\xi}$  is the standard dilogarithm.

The limit shapes arising in the above theorem are generally not analytic: near the corners of  $\mathfrak{N}$  the Hessian of  $\sigma$  becomes singular; this feature cause the limit shapes to form *facets* in the directions of the coordinate planes. These can be seen in [Figure 2](#): there are regions near the corners of solid color, containing only one tile type, which persist in the limit; in fact these regions are the exterior regions of the inscribed circle (one of the results of [Cohn, Larsen, and Propp \[1998\]](#).) The boundary of these facets is called the *frozen boundary*. Our limit shapes are only  $C^1$ , not  $C^2$ , across the frozen boundary. The region encircled by the frozen boundary is the *temperate zone*; the limit shape is analytic in the temperate zone.

**2.2 Analytic parameterization.** The surface tension function  $\sigma$  is sufficiently complicated that it is not immediately clear how to solve the variational problem of minimizing (1). Any minimizer will solve the associated Euler-Lagrange equation, which in this case is

$$\text{div}_{x,y}(\nabla \sigma(\nabla h)) = 0.$$

It is pretty gnarly when written out in Cartesian coordinates. Several years after the publication of the limit shape theorem, in joint work with Andrei Okounkov, we rewrote the Euler-Lagrange equation in terms of the variable  $z$  (the apex of the triangle mentioned above in the definition of  $\sigma$ ). In this variable the equation becomes significantly simpler

$$(2) \quad \frac{z_x}{z} - \frac{z_y}{1 - z} = 0,$$

which is a version of the *complex Burgers' equation* (substituting  $\phi = -\frac{z}{1-z}$  the equation is the actual complex Burgers' equation  $\phi_x + \phi\phi_y = 0$ ). This equation can be solved by the method of complex characteristics; in fact it is easy to check that if  $Q(z)$  is any analytic function and  $z = z(x, y)$  is defined implicitly by  $xz + y(1 - z) = Q(z)$  then  $z$  satisfies (2).



This gives a remarkable way to parameterize all limit shapes for lozenge tilings in terms of analytic functions. It is reminiscent of the Weierstrass-Enneper parameterization of minimal surfaces, except that here we are minimizing a different functional: not the surface area, but the function  $\sigma$  which depends on the slope.

One difficulty that this parameterization shares with the Weierstrass-Enneper parameterization is the relation between the analytic input  $Q(z)$  and the boundary data  $u$ . For given boundary data  $(\partial U, u)$ , how do we find  $Q$ ? This is a nontrivial problem which is not solved in general. See the next section for a large family of algebraic solutions.

If instead of solving for  $z$  as a function of  $x, y$  we solve for  $x, y$  as a function of  $z$ , we obtain a *linear* PDE

$$zx_{\bar{z}} + (1 - z)y_{\bar{z}} = 0.$$

This linearity results in an interesting semigroup property of limit shapes: given two limit shape surfaces  $\Sigma_1, \Sigma_2$  we can add them to get a third: this “addition” is a form of Minkowski sum: we add corresponding  $(x, y)$  coordinates where the normals are equal: the inverse of the gauss map for  $\Sigma_3$  is the sum of the inverse gauss maps for  $\Sigma_1$  and  $\Sigma_2$ .

**2.3 Rational parameterization.** In the special case that  $U$  is a polygon with edges in the directions of the cube roots of 1, as is the case for the hexagon in [Figure 2](#), there is an explicit method for computing the limit shape, discussed in [Kenyon and Okounkov \[2007\]](#). A quick way to find the limit shape, when  $U$  has  $3n$  sides which alternate in directions  $1, e^{2\pi i/3}, e^{4\pi i/3}$ , is to first find the frozen boundary, which has a *rational parameterization* in terms of  $t \in \hat{\mathbb{R}} = \mathbb{R} \cup \{\infty\}$ : it is the envelope of lines of the form

$$Ax \frac{\prod_{i=1}^n (t - a_i)}{\prod_{i=1}^n (t - c_i)} + By \frac{\prod_{i=1}^n (t - b_i)}{\prod_{i=1}^n (t - c_i)} = 1$$

where  $a_1, b_1, c_1, a_2, \dots, c_n$  are in cyclic order in  $\hat{\mathbb{R}}$ . One simply needs to find the numbers  $a_i, b_i, c_i$  and  $A, B$  so that this envelope contains the lines corresponding to the sides of  $U$ , in order. There is a unique solution up to precomposition by a real Möbius transformation of  $t$ .

As an example, for the hexagon of [Figure 2](#), the pencil of lines

$$\left\{ \frac{t-2}{t+1}x + \frac{1-t}{t}y + 1 = 0 \mid t \in \hat{\mathbb{R}} \right\}$$

contains the lines  $x = 0, y = 0, x - y = 1, x = 2, y = 2, x - y = -1$  in cyclic order, as  $t$  runs over  $\hat{\mathbb{R}}$ . As such it defines the frozen boundary. See [Figure 5](#) (this picture is distorted by a linear mapping as compared to [Figure 2](#) because it is plotted in Cartesian coordinates for clarity).

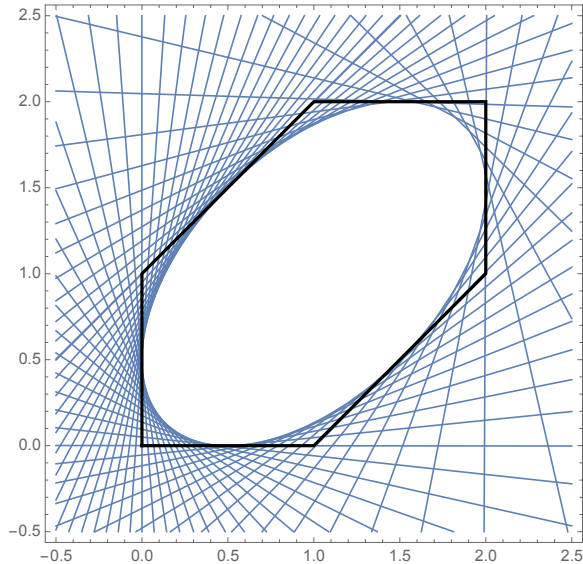


Figure 5: The pencil of lines defining the limit shape for the hexagon.

Once the frozen boundary is found, giving  $x = x(t)$ ,  $y = y(t)$ , the slope of the line defines  $z = z(t)$  (precisely, the slope is  $z(t)/(1 - z(t))$ ). Extending  $t$  to the upper half plane gives  $z = z(x, y)$  at interior points as well, from which the surface can be found by integrating.

### 3 Interacting lozenge tilings

Given the success of the calculations for the lozenge tiling model, it is natural to try to extend the results to other similar random tiling models. This was accomplished for general bipartite planar dimer models in the papers [Cohn, Kenyon, and Propp \[2001\]](#), [Kenyon, Okounkov, and Sheffield \[2006\]](#), and [Kenyon and Okounkov \[2007\]](#). The tractability of these models arises from their “determinantal” nature: the fundamental tool is the result of [Kasteleyn \[1963\]](#) who showed how to compute partition functions (weighted sums of configurations) for the planar dimer model with determinants.

If we move out of the class of determinantal models, things quickly become very challenging. No combinatorial methods are available for general tiling models. A few models have been shown to be “solvable”, to a certain extent, via so-called Bethe Ansatz methods, see e.g. [Baxter \[1982\]](#), [Kalugin \[1994\]](#), and [de Gier and Nienhuis \[1996\]](#). One of the most important of these Bethe-Ansatz solvable models is the six-vertex model, see [Figure 6](#). A

special case of this model (the “symmetric” case when  $a_1 = a_2, a_3 = a_4, a_5 = a_6$ ) was solved by Lieb in 1967 [Lieb \[1967\]](#) and partial results on the general six-vertex model were obtained by [Sutherland, C. N. Yang, and C. P. Yang \[1967\]](#). The general six-vertex model is still unsolved, however, in the sense that the asymptotic free energy and surface tension have not been computed.

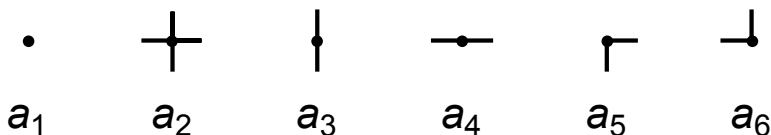


Figure 6: The six-vertex model consists of edge subsets of  $\mathbb{Z}^2$  whose local configurations at a vertex are one of these six types. The probability of a configuration is proportional to the product of its vertex weights.

In joint work with Jan de Gier and Sam Watson [de Gier, Kenyon, and Watson \[2018\]](#), we extend the lozenge tiling model to a model of interacting lozenges, known as the “5-vertex model” (this model was also studied, with partial results, by Huang et al [Huang, Wu, Kunz, and Kim \[1996\]](#)). We define a new probability measure by assigning a configuration a weight  $r$  for every adjacency between a blue and green rhombi, see [Figure 7](#); equivalently, in terms of the MNLP model, we assign each corner of a lattice path a weight  $r$ . This measure is a special case of the 6-vertex model, where one of the six possible vertex configurations is disallowed. For each choice of  $r$ , there is a two-parameter family of measures  $\mu_{s,t}$  on configurations, where  $s$  and  $t$  give the average slope. These three parameters  $r, s, t$  describe in fact the whole phase space of the model, which is 3-dimensional (scaling all weights has no effect on the measure, and for the boundary conditions we care about, the free energy depends on the weight of configurations 4 and 5 only via their product, so we may as well assume their weights are equal).

The analysis of the five-vertex model is considerably more involved than that of the lozenge tiling model, however we can get a complete picture of the limit shape theory including explicit analytic parameterizations of limit shapes. The methods used here still do not extend to the full six-vertex model, unfortunately. Our method relies on the Bethe Ansatz; the first part of the calculation was given by Sutherland et al [Sutherland, C. N. Yang, and C. P. Yang \[1967\]](#) in 1969.

**3.1 Electric field variables.** Suppose as in [Figure 7](#) we assign weights  $e^X$  and  $e^Y$  to vertical and horizontal edges, or equivalently (referring to [Figure 3](#)), to blue and green lozenges, so that a configuration (on a finite region) has probability  $\frac{1}{Z} e^{N_b X + N_g Y} r^{N_{bg}}$

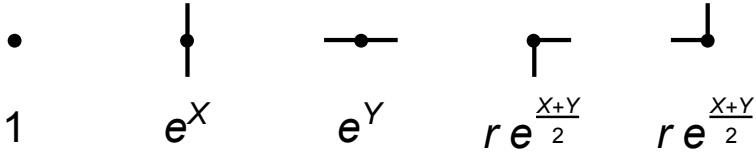


Figure 7: The five-vertex model is a special case of the 6-vertex model where no vertex has four incident edges (that is, weight  $a_2$  of Figure 6 is zero). It is convenient to parameterize the vertex weights as shown. Setting instead one of  $a_1, a_3, a_4$  equal to zero results in an equivalent model. If instead we set one of  $a_5$  or  $a_6$  equal to zero, we get a different 5-vertex model with different properties, quite degenerate, which we will not discuss here.

where  $N_b$  is the number of blue lozenges,  $N_g$  is the number of green lozenges,  $N_{bg}$  is the number of adjacencies between blue and green lozenges, and  $Z$  is a normalizing constant. For configuration with periodic boundary conditions, that is, on a torus, these “electric field” variables  $X, Y$  can be used to adjust the relative numbers of red, blue and green lozenges, thereby adjusting the *expected slope*  $(s, t) = \nabla h$ . For fixed  $r$  the phase space of the model can be parameterized either by the variables  $s, t$ , which describe the average slope, or  $X, Y$  which define the tile weights.

For fixed  $r$  the relationship between  $(X, Y)$  and  $(s, t)$  is essential in understanding the model, because we have the following equations relating the free energy  $F(X, Y)$  and the surface tension  $\sigma(s, t)$ :

$$(3) \quad \begin{aligned} \nabla F(X, Y) &= (s, t), \\ \nabla \sigma(s, t) &= (X, Y), \end{aligned}$$

which says that  $F$  and  $\sigma$  are Legendre dual to each other.

Although we didn’t mention this above, the electric field variables  $X, Y$  also play a role in the uniform lozenge tiling model, which is the special case  $r = 1$ ; in this case  $e^X, e^Y$  are weights assigned to blue and green lozenges (red lozenges have weight 1). The quantities  $e^X, e^Y, 1$  are in fact the side lengths of the  $0, 1, z$  triangle.

Knowing the map from  $(X, Y)$  to  $(s, t)$ , we can integrate (3) to get  $\sigma$ .

**3.2 Bethe Ansatz.** The 5-vertex measure can not be solved with determinants unless  $r = 1$  (in which case it is equivalent to the uniform lozenge tiling model). However it is still solvable on a cylinder via the Bethe Ansatz method. Let us briefly describe this method here. One considers configurations on a grid on a cylinder  $\mathbb{Z}/N \times \mathbb{Z}$  of circumference  $N$ . The states of the model are configurations of occupied vertical bonds on a given horizontal row; there are  $2^N$  possible states. The  $2^N \times 2^N$  transfer matrix

$T$  is indexed by such row configurations and the entry  $T(x, y)$  is the weighted sum of configurations between a row configuration  $x$  and the configuration  $y$  occurring on the subsequent row. Thus  $T^k(x, y)$  is the weighted number of configurations starting from row 0 in configuration  $x$  and ending  $k$  rows later in configuration  $y$ . The logarithm of the Perron eigenvalue of  $T$  is the free energy per row; dividing by  $N$  and taking the limit  $N \rightarrow \infty$  gives the normalized free energy per site of the model. From this free energy the surface tension can be computed by an appropriate Legendre transform as discussed above.

**3.3 Surface tension.** For the five vertex model the limit shape theorem and formula (1) still hold, however the surface tension function  $\sigma$  is more complicated to write down explicitly. We can give the surface tension derivative as follows: Let  $w$  be a point in the upper half plane and consider the triangle with vertices  $0, 1, w$ . Draw in the line from  $w$  to  $1/(1-r^2)$  as indicated in Figure 8.

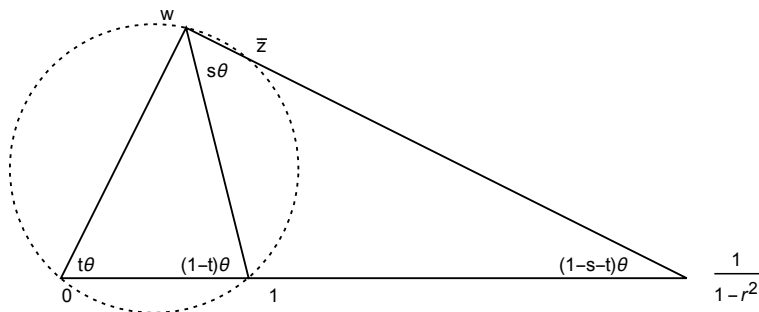


Figure 8

Given  $w$ , let  $t$  be defined by the equation  $\frac{t}{1-t} = \frac{\arg w}{\arg(\frac{1}{1-w})}$ , that is,  $t/(1-t)$  is the ratio of base angles at 0 and 1 of the triangle  $\{0, 1, w\}$ . Let  $s$  be defined by the equation that  $(1-s-t)/t$  is the ratio of the base angles of the triangle  $\{0, \frac{1}{1-r^2}, w\}$ . Consider the circle through  $0, 1, w$  and let  $\bar{z}$  be the other intersection of this circle with the line from  $w$  to  $1/(1-r^2)$ . Then  $w, z$  satisfy the equation

$$1 - w - z + (1 - r^2)wz = 0.$$

(This equation defines the “spectral curve” of the model.) Let  $X = -\log(1 - r^2) - B(w)$  and  $Y = -\log(1 - r^2) - B(\bar{z})$  where

$$(4) \quad B(u) = \frac{1}{\pi} (\arg u \log |1 - u| + \operatorname{Im} \operatorname{Li}(u))$$

is another variant of the dilogarithm. This gives  $X, Y, s, t$  as functions of  $w$ , and thus we can find implicitly  $X, Y$  as functions of  $s, t$ , from which we arrive at a formula for  $\sigma$  and the free energy, see [Figure 9](#) and [Figure 11](#).

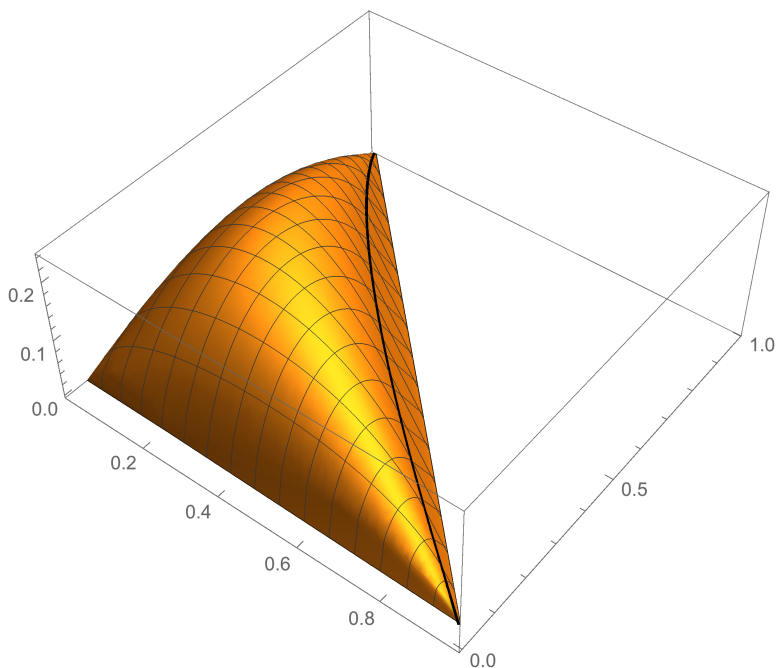


Figure 9: Minus surface tension as a function of  $(s, t) \in \mathfrak{N}$  with  $r = .8$ . The black line is graph of surface tension on the hyperbola bounding  $\mathfrak{N}^*$ ; the entropy is linear in  $\mathfrak{N} \setminus \mathfrak{N}^*$ .

The surface tension  $\sigma$  has an interesting feature that for  $r < 1$  it is not strictly convex, and in fact is only piecewise analytic: it is strictly convex and analytic on the region  $\mathfrak{N}^* \subseteq \mathfrak{N}$  bounded by the axes and the hyperbola

$$\frac{1-r^2}{r^2}xy + x + y - 1 = 0.$$

In the region  $\mathfrak{N} \setminus \mathfrak{N}^*$ ,  $\sigma$  is linear.

The free energy, which is the Legendre dual of  $\sigma$ , is shown (for  $r = .8$ ) in [Figure 11](#). It is piecewise analytic with four pieces, see [Figure 12](#).

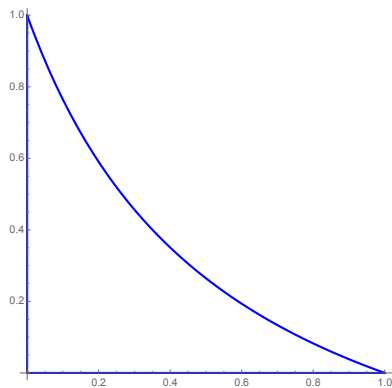


Figure 10: The region  $\mathfrak{N}^*$  is bounded by the axes and the hyperbola  $(\frac{1-r^2}{r^2})xy + x + y - 1 = 0$  (shown here for  $r = 0.6$ ).

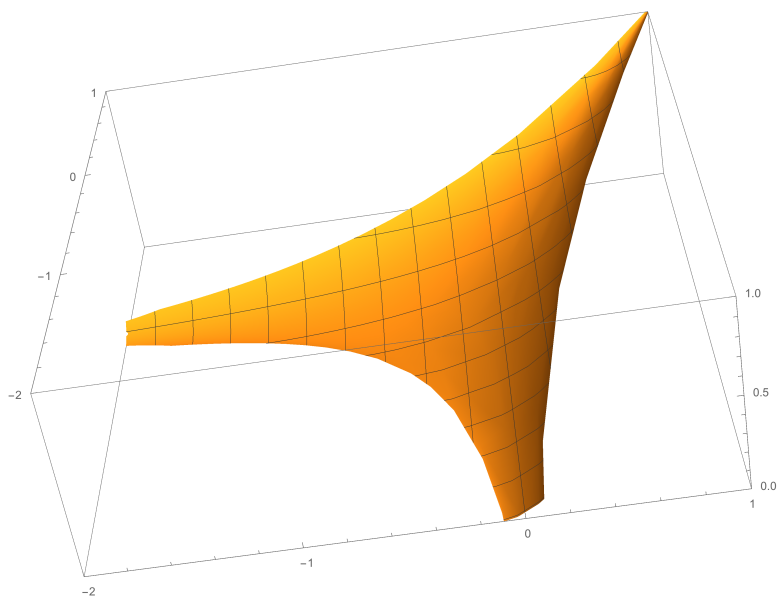


Figure 11: The free energy  $F(X, Y)$  for  $r = .8$ . It is linear in each complementary component of the curved region shown (except for a slope change along the line  $y = x$  for  $x > -\log(1 - r^2)$ ).

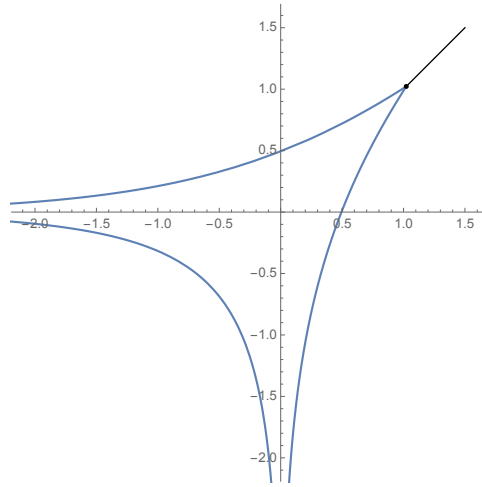


Figure 12: For  $r < 1$ , the region bounding the analytic part of the free energy is bounded by the three curves  $1 - e^X - e^Y + (1 - r^2)e^{X+Y} = 0$ ,  $-1 + e^X - r^2e^Y = 0$ , and  $-1 + e^Y - r^2e^X = 0$ . The line  $x = y$  separates the external phases.

The graph of the free energy is also the limit shape of a 3D partition, weighted according to the 5-vertex measure. That is, consider all 3D partitions of  $n$ , with probability measure proportional to  $r^{N_{bg}}$  where  $N_{bg}$  is the number of blue-green lozenge adjacencies. Then for large  $n$ , when rescaled by  $n^{1/3}$  (so that the total volume is 1) the limit shape of the 3D partition is given by the graph in [Figure 11](#) (once we apply a linear coordinate change so that the two facets become vertical:  $(x, y, z) \rightarrow (x - z, y - z, z)$ ); see [Figure 13](#).

**3.4 Euler-Lagrange equation.** Because  $(X, Y) = \nabla\sigma$ , the Euler-Lagrange equation for the surface tension minimizer can be written in a very simple form

$$X_x + Y_y = 0.$$

This can be supplemented with another equation

$$s_y = t_x$$

(the equality of mixed partials of the height function). It is a small miracle (i.e. something which we can prove but don't fully understand) that these two equations can be combined into a single equation for the complex variable  $w$ . The equation is more symmetric when written in terms of  $z$  and  $w$  (but remember that  $1 - w - z + (1 - r^2)wz = 0$ )

$$(5) \quad \frac{\partial B(w)}{\partial w} w_y - \frac{\partial B(z)}{\partial z} z_x = 0$$



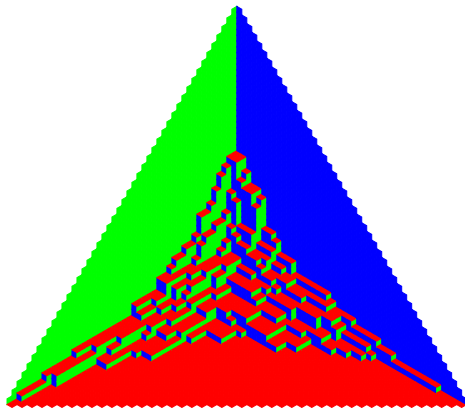


Figure 13: A 3D partition (that is, 3D Young diagram) weighted according to the 5-vertex measure with  $r = .7$ .

where  $B$  is the function of (4).

It is worth rewriting (5) as an equation for  $x$  and  $y$  (real variables) as a function of  $w$  (a complex variable), instead of  $w$  as a function of  $x$  and  $y$ . After some manipulation this becomes the first order linear equation

$$(6) \quad A(w)x_{\bar{w}} - A(z)y_{\bar{w}} = 0$$

where  $A(w)$  is the function  $A(w) := -w \arg w - (1 - w) \arg(1 - w)$ .

The Equation (6) is first order but cannot be solved by standard methods (of complex characteristics); it is really a coupled  $2 \times 2$  system of real PDEs. Another miracle is that one can integrate it explicitly to find an explicit parameterization of solutions with analytic functions. Rather than derive the solution, we'll just give the answer here. Let  $u = 1 - (1 - r^2)\bar{w}$ , and consider the equation

$$(7) \quad x - \frac{r^2}{u\bar{u}}y = \frac{\operatorname{Im} h(u)}{\operatorname{Im} u}.$$

For *any* analytic function  $h$ , solve this for  $y$ , plug into (6), and arrive at an equation for  $x$  of the form

$$x_u + A_1x + A_2 = 0$$

for functions  $A_1(u), A_2(u)$ , which can be integrated by standard techniques. The special form of (7) guarantees that  $x$  will be real.

This leads to the following explicit parameterization of solutions to (5):

$$(8) \quad x(w) = \frac{-1}{(r^2 A(w) - |u|^2 A(z))} \operatorname{Im} \left( \int_{u_0}^u \frac{r^2 h(u)}{(1-u)(u-r^2)} du + \frac{|u|^2 h(u) A(z)}{\operatorname{Im}(u)} \right)$$

and  $y$  is defined by (7).

Plugging in for example  $h(u) = 0$  leads to (inserting an appropriate constant of integration in (8))

$$(x, y) = \left( \frac{1}{r^2 A(w) - u\bar{u} A(z)}, \frac{u\bar{u}}{r^2 (r^2 A(w) - u\bar{u} A(z))} \right).$$

## 4 Open problems

There are many open problems and mysteries in the above calculations, some of which we are still working on. Here are four important ones.

1. Is there a nice parameterization of limit shapes for the five-vertex model (in polygonal domains with sides in directions of cubic roots of 1) in terms of rational curves, like there is for the lozenge tiling model?
2. Should we expect that for the six-vertex model we can similarly combine the Euler-Lagrange equation  $X_s + Y_t = 0$  and the mixed-partial equation  $s_y - t_x = 0$  into a single ODE for a complex variable  $z$ ? Is this some general property of conformally invariant models? And if so, is there a similar parameterization of solutions with analytic functions?
3. For the lozenge tiling,  $\sigma$  has the interesting property that its Hessian determinant  $\det H(\sigma)$  is a constant. Is there an analog of this Monge-Ampère relation for  $\sigma$  for the 5-vertex model?
4. Under what conditions on a real analytic function  $f(z, \bar{z})$  can the PDE

$$z_x = f(z, \bar{z}) z_y$$

be solved (for a complex function  $z = z(x, y)$  of two real variables) explicitly? When  $f$  is analytic or antianalytic (that is, a function of  $z$  or  $\bar{z}$  only), it can easily be solved by characteristics. The case of (5) is neither but we were still (after some significant head-scratching) able to find solutions by some sort of generalized characteristics.

## References

- Rodney J. Baxter (1982). *Exactly solved models in statistical mechanics*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], London, p. 486. MR: [690578](#) (cit. on p. [3161](#)).
- Henry Cohn, Noam Elkies, and James Propp (1996). “[Local statistics for random domino tilings of the Aztec diamond](#)”. *Duke Math. J.* 85.1, pp. 117–166. MR: [1412441](#) (cit. on p. [3158](#)).
- Henry Cohn, Richard Kenyon, and James Propp (2001). “[A variational principle for domino tilings](#)”. *J. Amer. Math. Soc.* 14.2, pp. 297–346. MR: [1815214](#) (cit. on pp. [3158](#), [3161](#)).
- Henry Cohn, Michael Larsen, and James Propp (1998). “[The shape of a typical boxed plane partition](#)”. *New York J. Math.* 4, pp. 137–165. MR: [1641839](#) (cit. on pp. [3158](#), [3159](#)).
- Jan de Gier, Richard Kenyon, and Sam Watson (2018). “Limit shapes for the asymmetric five vertex model”. *preprint* (cit. on p. [3162](#)).
- Jan de Gier and Bernard Nienhuis (1996). “Exact solution of an octagonal random tiling model”. *Phys. Rev. Lett.* 76 (cit. on p. [3161](#)).
- H. Y. Huang, F. Y. Wu, H. Kunz, and D. Kim (1996). “[Interacting dimers on the honeycomb lattice: an exact solution of the five-vertex model](#)”. *Phys. A* 228.1-4, pp. 1–32. MR: [1399280](#) (cit. on p. [3162](#)).
- William Jockusch, James Propp, and Peter Shor (1998). “[Random Domino Tilings and the Arctic Circle Theorem](#)”. arXiv: [math/9801068](#) (cit. on p. [3158](#)).
- P. A. Kalugin (1994). “[The square-triangle random-tiling model in the thermodynamic limit](#)”. *Journal of Physics A: Mathematical and General* 27.11, p. 3599 (cit. on p. [3161](#)).
- P. W. Kasteleyn (1963). “[Dimer statistics and phase transitions](#)”. *J. Mathematical Phys.* 4, pp. 287–293. MR: [0153427](#) (cit. on p. [3161](#)).
- Richard Kenyon and Andrei Okounkov (2007). “[Limit shapes and the complex Burgers equation](#)”. *Acta Math.* 199.2, pp. 263–302. MR: [2358053](#) (cit. on pp. [3160](#), [3161](#)).
- Richard Kenyon, Andrei Okounkov, and Scott Sheffield (2006). “[Dimers and amoebae](#)”. *Ann. of Math. (2)* 163.3, pp. 1019–1056. MR: [2215138](#) (cit. on p. [3161](#)).
- Elliott H. Lieb (Oct. 1967). “[Residual Entropy of Square Ice](#)”. *Phys. Rev.* 162 (1), pp. 162–172 (cit. on p. [3162](#)).
- B. Sutherland, C. N. Yang, and C. P. Yang (Sept. 1967). “[Exact Solution of a Model of Two-Dimensional Ferroelectrics in an Arbitrary External Electric Field](#)”. *Phys. Rev. Lett.* 19 (10), pp. 588–591 (cit. on p. [3162](#)).
- A. M. Vershik and S. V. Kerov (1981). “Asymptotic theory of the characters of a symmetric group”. *Funktsional. Anal. i Prilozhen.* 15.4, pp. 15–27, 96. MR: [639197](#) (cit. on p. [3155](#)).

Received 2018-02-21.

RICHARD KENYON

[rkenyon@math.brown.edu](mailto:rkenyon@math.brown.edu)



## COMPLEXITY PROBLEMS IN ENUMERATIVE COMBINATORICS

IGOR PAK

### Abstract

We give a broad survey of recent results in enumerative combinatorics and their complexity aspects.

### Introduction

The subject of Enumerative Combinatorics is both classical and modern. It is classical as the basic counting questions go back millennia, yet it is modern in the use of a large variety of the latest ideas and technical tools from across many areas of mathematics. The remarkable successes from the last few decades have been widely publicized, yet they come at a price, as one wonders if there is anything left to explore. In fact, are there enumerative problems which cannot be resolved with existing technology? In this paper we present many challenges in the field from the Computational Complexity point of view, and describe how recent results fit into the story.

Let us first divide the problems into three major classes. This division is not as neat as it may seem as there are problems which fit into multiple or none of the classes, especially if they come from other areas. Still, it would provide us with a good starting point.

- (1) **Formula.** Let  $\mathcal{P}$  be a set of combinatorial objects, think of trees, words, permutations, Young tableaux, etc. Such objects often come with a parameter  $n$  corresponding to the size of the objects. Let  $\mathcal{P}_n$  be the set of objects of size  $n$ . Find a formula for  $|\mathcal{P}_n|$ .
- (2) **Bijection.** Now let  $\mathcal{P}$  and  $\mathcal{Q}$  be two sets of (possibly very different) combinatorial objects. Say, you know (or at least suspect) that  $|\mathcal{P}_n| = |\mathcal{Q}_n|$ . Find an explicit bijection  $\varphi : \mathcal{P}_n \rightarrow \mathcal{Q}_n$ .

---

The author was partially supported by the NSF.

*MSC2010:* primary 05A15; secondary 05A10, 05A17, 68R05, 68Q17, 05C30.

(3) **Combinatorial interpretation.** Now suppose there is an integer sequence  $\{a_n\}$  given by a formula. Say, you know (or at least suspect) that  $a_n \geq 0$  for all  $n$ . Find a combinatorial interpretation of  $a_n$ , i.e. a set of combinatorial objects  $\mathcal{P}$  such that  $|\mathcal{P}_n| = a_n$ .

People in the area are well skilled in both resolving and justifying these problems. Indeed, a formula is a good thing to have in case one needs to compute  $|\mathcal{P}_n|$  explicitly for large  $n$ , find the asymptotics, gauge the structural complexity of the objects, etc. A bijection between a complicated set  $\mathcal{P}$  and a simpler set  $\mathcal{Q}$  is an even better thing to have, as it allows one to better understand the nature of  $\mathcal{P}$ , do a refined counting of  $\mathcal{P}_n$  with respect to various statistics, generate elements of  $\mathcal{P}_n$  at random, etc. Finally, a combinatorial interpretation is an excellent first step which allows one to proceed to (1) and then (2), or at least obtain some useful estimates for  $a_n$ .

Here is the troubling part, which comes in the form of inquisitive questions in each case:

(1') What is a formula? What happens if there is no formula? Can you prove there isn't one? How do you even formalize the last question if you don't know the answer to the first?

(2') There are, obviously,  $|\mathcal{P}_n|!$  bijections  $\varphi : \mathcal{P}_n \rightarrow \mathcal{Q}_n$ , so you must want a particular one, or at least one with certain properties? Is there a "canonical" bijection, or at least the one you want best? What if there isn't a good bijection by whatever measure, can you prove that? Can you even formalize that?

(3') Again, what do you do in the case when there isn't a combinatorial interpretation? Can you formally prove a negative result so that others stop pursuing these problems?

We have a few formal answers to these questions, at least in some interesting special cases.<sup>1</sup> As the reader will see, the complexity approach does bring some clarity to these matters. But to give the answers we first need to explain the nature of combinatorial objects in each case, and to review the literature. That is the goal of this survey.

## 1 What is a formula?

**1.1 Basic examples.** We start with the Fibonacci numbers [Sloane \[n.d., A000045\]](#):

$$(1-1) \quad F_n = F_{n-1} + F_{n-2}, \quad F_0 = F_1 = 1$$

$$(1-2) \quad F_n = \sum_{i=0}^{\lfloor n/2 \rfloor} \binom{n-i}{i}$$

---

<sup>1</sup>Due to space limitations, we address (3) and (3') in the full version of the paper.

$$(1-3) \quad F_n = \frac{1}{\sqrt{5}} \left( \phi^n + (-\phi)^{-n} \right), \quad \text{where } \phi = \frac{1 + \sqrt{5}}{2}$$

$$(1-4) \quad F_n = (A^n)_{2,2}, \quad \text{where } A = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix}.$$

Equation (1-1) is usually presented as a definition, but can also be used to compute  $F_n$  in  $\text{poly}(n)$  time. Equation (1-2) is useful to place Fibonacci numbers in the hierarchy of integer sequences (see below). Equation (1-3) is useful to obtain asymptotics, and equation (1-4) gives a fast algorithm for computing  $F_n$  (by repeated squaring). *The moral:* there is no one notion of a “good formula”, as different equations have different uses.

Let us consider a few more plausible formula candidates:

$$(1-5) \quad D_n = \lfloor [n!/e] \rfloor, \quad \text{where } \lfloor [x] \rfloor \text{ denotes the nearest integer}$$

$$(1-6) \quad C_n = [t^n] \frac{1 - \sqrt{1 - 4t}}{2t}$$

$$(1-7) \quad E_n = n! \cdot [t^n] y(t), \quad \text{where } 2y' = 1 + y^2, \quad y(0) = 1$$

$$(1-8) \quad T_n = (n-1)! \cdot [t^n] z(t), \quad \text{where } z = t e^{t e^{t e^{t e^{\dots}}}}$$

Here  $D_n$  is the number of *derangements* (fixed-point-free permutations in  $S_n$ ),  $C_n$  is the *Catalan number* (the number of binary trees with  $n$  vertices),  $E_n$  is the *Euler number* (the number of *alternating permutations*  $\sigma(1) < \sigma(2) > \sigma(3) < \sigma(4) > \dots$  in  $S_n$ ),  $T_n$  is the *Cayley number* (the number of spanning trees in  $K_n$ ), and  $[t^n] F(t)$  denotes the coefficient of  $t^n$  in  $F(t)$ .

In each case, there are better formulas for applications:

$$(1-9) \quad D_n = n! \sum_{k=0}^n \frac{(-1)^k}{k!}$$

$$(1-10) \quad C_n = \frac{1}{n+1} \binom{2n}{n}$$

$$(1-11) \quad E_n = n! \cdot [t^n] y(t), \quad \text{where } y(t) = \tan(t) + \sec(t)$$



$$(1-12) \quad T_n = n^{n-2}$$

In all four cases, the corresponding formulas are equivalent by mathematical reasoning. Whether or not you accept (1-5)–(1-8) as formulas, it is their meaning that's important, not their form.

Finally, consider the following equations for the *number of partitions*  $p(n)$ , and *n-th prime number*  $p_n$ :

$$(1-13) \quad p(n) = [t^n] \prod_{i=1}^{\infty} \frac{1}{1-t^i}$$

$$(1-14) \quad p_n = \sum_{m=2}^{n^2} m \left\langle 1 + \left| n - \langle \gamma_m \rangle \sum_{r=2}^m \langle \gamma_r \rangle \right| \right\rangle, \text{ where } \langle x \rangle := \left\lfloor \frac{1}{x} \right\rfloor, \quad \gamma_r := \sum_{d=1}^{\sqrt{r}} \left\lfloor \frac{\lfloor r/d \rfloor}{r/d} \right\rfloor.$$

Equation (1-13) is due to Euler (1748), and had profound implications in number theory and combinatorics, initiating the whole area of *partition theory* (see e.g. [R. Wilson and Watkins \[2013\]](#)). Equation (1-14) is from [Tsangaris \[2007\]](#). Esthetic value aside, both equations are largely unhelpful for computing purposes and follow directly from definitions. Indeed, the former is equivalent to the standard counting algorithm (*dynamic programming*), while the latter is an iterated divisibility testing in disguise.

In summary, we see that the notion of “good formula” is neither syntactic nor semantic. One needs to make a choice depending on the application.

**1.2 Wilfian formulas.** In his pioneer 1982 paper [Wilf \[1982\]](#), Wilf proposed to judge a formula from the complexity point of view. He suggested two definitions of “good formulas” for computing an integer sequence  $\{a_n\}$ :

(W1) There is an algorithm which computes  $a_n$  in time  $\text{poly}(n)$ .

(W2) There is an algorithm which computes  $a_n$  in time  $o(a_n)$ .

In the literature, such algorithms are called sometimes *Wilfian formulas*. Note that (W1) is aimed to apply for sequences  $\{a_n\}$  of at most exponential growth  $a_n = \exp O(n^c)$ , while (W2) for  $\{a_n\}$  of at most polynomial growth (see e.g. [Garrabrant and Pak \[2017\]](#) and [Flajolet and Sedgewick \[2009\]](#) for more on growth of sequences).

Going over our list of examples we conclude that (1-1), (1-2), (1-4), (1-9), (1-10) and (1-12) are all transparently Wilfian of type (W1). Equations (1-3), (1-6), (1-7) and (1-11) are Wilfian of type (W1) in a less obvious but routine way (see below). Equations (1-3) and (1-5) do give rise to ad hoc  $\text{poly}(n)$  algorithms, but care must be applied when dealing with irrational numbers. E.g., one must avoid circularity, such as when computing  $\{p_n\}$  by

using the *prime constant*  $\sum_n 1/2^{p_n}$ , see Sloane [n.d., A051006]. Finally, equation (1-8) is not Wilfian of type (W1), while (1-14) is not Wilfian of type (W2).

Let us add two more notions of a “good formula” in the same spirit, both of which are somewhat analogous but more useful than (W2):

(W3) There is an algorithm which computes  $a_n$  in time  $\text{poly}(\log n)$ .

(W4) There is an algorithm which computes  $a_n$  in time  $n^{o(1)}$ .

Now, for a *combinatorial sequence*  $\{a_n\}$  one can ask if there is a Wilfian formula. In the original paper Wilf [1982] an explicit example is given:

**Conjecture 1.1** (Wilf). *Let  $a_n$  be the number of unlabeled graphs on  $n$  vertices. Then  $\{a_n\}$  has no Wilfian formula of type (W1).*

See Sloane [n.d., A000088] for this sequence. Note that by the classical result Erdős and Rényi [1963] (see also Babai [1995, §1.6]), we have  $a_n \sim 2^{\binom{n}{2}}/n!$ , so the problem is not approximating  $a_n$ , but computing it exactly. For comparison, the sequence  $\{c_n\}$  of the number of connected (labeled) graphs does have a Wilfian formula:

$$c_n = 2^{\binom{n}{2}} - \frac{1}{n} \sum_{k=1}^{n-1} k \binom{n}{k} 2^{\binom{n-k}{2}} c_k$$

(see Sloane [n.d., A001187] and Harary and Palmer [1973, p. 7]).

The idea behind Conjecture 1.1 is that the *Pólya theory* formulas (see e.g. Harary and Palmer [ibid.]) are fundamentally not Wilfian. We should mention that we do not believe the conjecture in view of Babai’s recent quasipolynomial time algorithm for GRAPH ISOMORPHISM Babai [2016]. While the connection is indirect, it is in fact conceivable that both problems can be solved in  $\text{poly}(n)$  time.

**Open Problem 1.2.** *Let  $\pi(n)$  denote the number of primes  $\leq n$ . Does  $\{\pi(n)\}$  have a Wilfian formula of type (W4)?*

The *prime-counting function*  $\pi(n)$  has a long history. Initially Wilf asked about formula of type (W2), and such formula was found in Lagarias, V. S. Miller, and Odlyzko [1985]. Note that even the parity of  $\pi(n)$  is hard to compute Tao, Croot, and Helfgott [2012].

**1.3 Complexity setting and graph enumeration.** Let  $\mathcal{O}_n$  denote the set of certain *combinatorial objects* of size  $n$ . This means one can decide if  $X \in \mathcal{O}_n$  in time  $\text{poly}(n)$ . The problem of computing  $a_n := |\mathcal{O}_n|$  is in #EXP because the input  $n$  has *bit-length*  $O(\log n)$ .<sup>2</sup> This is a counting version of the decision problem NEXP.

<sup>2</sup>To bring the problem into the (usual) polynomial hierarchy, the input  $n$  should be given in *unary*, cf. Gol-dreich [2008] and Moore and Mertens [2011].

For example, let  $a_n = |\mathcal{P}_n|$  be the set of (labeled) planar 3-regular 3-connected graphs on  $n$  vertices. Graphs in  $\mathcal{P}_n$  are exactly graphs of simple 3-dimensional polytopes. Since testing each property can be done in  $\text{poly}(n)$  time, the decision problem is naturally in NEXP, and the counting problem is in #EXP. In fact, the decision problem is trivially in P, since such graphs exist for all even  $n \geq 4$  and don't exist for odd  $n$ . Furthermore, Tutte's formula for the number of rooted plane triangulations gives a simple product formula for  $a_n$ , and thus can be computed in  $\text{poly}(n)$  time, see [Tutte \[1998, Ch. 10\]](#).

On the one hand, counting the number of non-Hamiltonian graphs in  $\mathcal{P}_n$  is not naturally in #EXP, since testing non-Hamiltonicity is CO-NP-complete in this case [Garey, Johnson, and Tarjan \[1976\]](#). On the other hand, the corresponding decision problem (the existence of such graphs) is again in P by Tutte's disproof of Tait's conjecture, see [Tutte \[1998, Ch. 2\]](#).

Note that GRAPH ISOMORPHISM is in P for trees, planar graphs and graphs of bounded degree, see e.g. [Babai \[1995, §6.2\]](#). The discussion above suggests the following counterpart of Wilf's [Conjecture 1.1](#).

**Conjecture 1.3.** *Let  $a_n$  be the number of unlabeled plane triangulations with  $n$  vertices,  $b_n$  the number of 3-connected planar graphs with  $n$  vertices, and  $t_n$  the number of unlabeled trees with  $n$  vertices. Then  $\{a_n\}$ ,  $\{b_n\}$  and  $\{t_n\}$  can be computed in  $\text{poly}(n)$  time.*

We are very optimistic about this conjecture. For triangulations and trees, there is some recent evidence in [Kang and Sprüssel \[2018\]](#) and the theory of species [Bergeron, Labelle, and Leroux \[1998\]](#), respectively. See also [Noy, Requilé, and Rué \[2018\]](#) for further positive results on enumeration of (labeled) planar graphs.

Denote by  $a_n$  the number of 3-regular labeled graphs on  $2n$  vertices. The sequence  $\{a_n\}$  can be computed in polynomial time via the following recurrence relation, see [Sloane \[n.d., A002829\]](#).

(1-15)

$$\begin{aligned} 3(3n-7)(3n-4) \cdot a_n &= 9(n-1)(2n-1)(3n-7)(3n^2-4n+2) \cdot a_{n-1} \\ &+ (n-1)(2n-3)(2n-1)(108n^3-441n^2+501n-104) \cdot a_{n-2} \\ &+ 2(n-2)(n-1)(2n-5)(2n-3)(2n-1)(3n-1)(9n^2-42n+43) \cdot a_{n-3} \\ &- 2(n-3)(n-2)(n-1)(2n-7)(2n-5)(2n-3)(2n-1)(3n-4)(3n-1) \cdot a_{n-4} \end{aligned}$$

**Conjecture 1.4.** *Fix  $k \geq 1$  and let  $a_n$  be the number of unlabeled  $k$ -regular graphs with  $n$  vertices. Then  $\{a_n\}$  can be computed in  $\text{poly}(n)$  time.*

For  $k = 1, 2$  the problem is elementary, but for  $k = 3$  is related to enumeration of certain 2-groups (cf. [Luks \[1982\]](#)).

Consider now the problem of computing the number  $f(m, n)$  of triangulations of an integer  $[m \times n]$  grid (see Figure 1). This problem is a distant relative of Catalan numbers  $C_n$  in (1-10) which Euler proved counts the number of triangulations of a convex  $(n+2)$ -gon, see R. P. Stanley [2015], and is one of the large family of triangulation problems, see De Loera, Rambau, and Santos [2010]. Kaibel and Ziegler prove in Kaibel and Ziegler [2003] that  $f(m, n)$  can be computed in  $\text{poly}(n)$  time for every fixed  $m$ , but report that their algorithm is expensive even for relatively small  $m$  and  $n$  (see Sloane [n.d., A082640]).

**Question 1.5.** *Can  $\{f(n, n)\}$  can be computed in  $\text{poly}(n)$  time?*

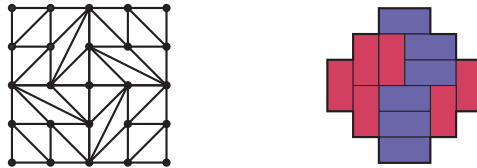


Figure 1: Grid triangulation of  $[5 \times 5]$  and a domino tiling.

**1.4 Computability setting and polyomino tilings.** Let  $a_n$  be the number of domino tilings on a  $[2n \times 2n]$  square. Kasteleyn and Temperley–Fisher classical determinant formula (1961) for the number of perfect matchings of planar graphs gives a  $\text{poly}(n)$  time algorithm for computing  $\{a_n\}$ , see e.g. Kenyon [2004] and Lovász and Plummer [1986]. This foundational result opens the door to potential generalizations, but, unfortunately, most of them turn out to be computationally hard.

First, one can ask about computing the number  $b_n$  of 3-dimensional domino tilings of a  $[2n \times 2n \times 2n]$  box. Or how about the seemingly simpler problem of counting the number  $c_n$  of 3-dimensional domino tilings of a “slim”  $[2 \times n \times n]$  box? We don’t know how to solve either problem, but both are likely to be difficult. The negative results include #P-completeness of the counting problem for general and slim regions Pak and Yang [2013] and Valiant [1979], and topological obstacles, see Freire, Klivans, Milet, and Saldanha [2017] and Pak and Yang [2013, Prop. 8.1].

Consider now a fixed finite set  $\mathbf{T} = \{\tau_1, \dots, \tau_k\}$  of general polyomino tiles on a square grid:  $\tau_i \subset \mathbb{Z}^2$ ,  $1 \leq i \leq k$ . To tile a region  $\Gamma \subset \mathbb{Z}^2$ , one must cover it with copies of the tiles without overlap. These copies must be parallel translations of  $\tau_i$  (rotations and reflections are not allowed). There exist NP-complete tileability problems even for a fixed set of few small tiles. We refer to Pak [2003] for short survey of the area.

For a fixed  $\mathbf{T}$ , denote by  $g(m, n)$  the number of tilings of  $[m \times n]$  with  $\mathbf{T}$ . Is  $g(m, n)$  computable in polynomial time? The following conjecture suggests otherwise.

**Conjecture 1.6.** *There exists a finite set of tiles  $\mathbf{T}$  such that counting the number of tilings of  $[n \times n]$  with  $\mathbf{T}$  is #EXP-complete.*

In fact, until we started writing this survey, we always believed this result to be known, only to realize that standard references such as [van Emde Boas \[1997\]](#) fall a bit short. Roughly, one needs to embed a #EXP-complete language into a counting tilings problem of a rectangle. This may seem like a classical idea (see e.g. [Moore and Mertens \[2011, §5.3.4, §7.6.5\]](#)), which worked well for many related problems. For example, the RECTANGULAR TILEABILITY asks: given a finite set of tiles  $\mathbf{T}$ , does there exist integers  $m$  and  $n$ , such that  $\mathbf{T}$  tiles  $[m \times n]$ .

**Theorem 1.7** ([Yang \[2014\]](#)). *The RECTANGULAR TILEABILITY problem is undecidable.*

In the proof, Yang embeds the HALTING PROBLEM into RECTANGULAR TILEABILITY. So can one embed a NEXP-complete problem into tileability of  $[m \times n]$  rectangle? The answer is yes if  $\mathbf{T}$  is allowed to be part of the input. In fact, even Levin's original 1973 paper introducing NP-completeness proposed this approach [L. A. Levin \[1973\]](#). The following result should come as a surprise, perhaps.

**Theorem 1.8** ([Lam \[2008\]](#)). *Given  $\mathbf{T}$ , the tileability of  $[m \times n]$  can be decided in  $O(\log m + \log n)$  time.*

The proof is nonconstructive; it is based on *Hilbert's Basis Theorem* and the algebraic approach by F. W. Barnes. A combination of [Theorem 1.7](#) and [Theorem 1.8](#) implies that the constant implied by the  $O(\cdot)$  notation is not computable as a function of  $\mathbf{T}$ . Roughly, we do know that a linear time algorithm exists, but given  $\mathbf{T}$  it is undecidable to find it. [Theorem 1.8](#) also explains why [Conjecture 1.6](#) remains open – most counting results in the area use parsimonious reductions (think bijections between solutions of two problems), and in this case a different approach is required.

## 2 Classes of combinatorial sequences

**2.1 Algebraic and D-algebraic approach.** Combinatorial sequences  $\{a_n\}$  are traditionally classified depending on the algebraic properties of their GFs

$$A(t) = \sum_{n=0}^{\infty} a_n t^n.$$

We list here only four major classes:

- Rational:**  $A(t) = P(t)/Q(t)$ , for some  $P, Q \in \mathbb{Z}[t]$ ,  
**Algebraic:**  $c_0 A^k + c_1 A^{k-1} + \dots + c_k = 0$ , for some  $k \in \mathbb{N}$ ,  $c_i \in \mathbb{Z}[t]$ ,  
**D-finite:**  $c_0 A + c_1 A' + \dots + c_k A^{(k)} = b$ , for some  $k \in \mathbb{N}$ ,  $b, c_i \in \mathbb{Z}[t]$ ,  
**D-algebraic:**  $Q(t, A, A', \dots, A^{(k)}) = 0$ , for some  $k \in \mathbb{N}$ ,  $Q \in \mathbb{Z}[t, x_0, x_1, \dots, x_k]$ .

Note that rational GFs are exactly those  $\{a_n\}$  which satisfy linear recurrence:

$$c_0 a_n = c_1 a_{n-1} + \dots + c_k a_{n-k}, \text{ for some } k \in \mathbb{N}, c_i \in \mathbb{Z}.$$

Such sequences  $\{a_n\}$  are called *C-recursive*. For example, Fibonacci numbers satisfy (1-1) and have GF  $(1 - t - t^2)^{-1}$ . Similarly, *Catalan numbers* have algebraic GF by (1-6). D-finite GFs (also called *holonomic*) are exactly those  $\{a_n\}$  which satisfy polynomial recurrence

$$c_0(n) a_n = c_1(n) a_{n-1} + \dots + c_k(n) a_{n-k}, \text{ for some } k \in \mathbb{N}, c_i \in \mathbb{Z}[n].$$

Such sequences  $\{a_n\}$  are called *P-recursive*. Examples include  $\{n!\}$ , derangement numbers  $\{D_n\}$  by (1-9), the number of 3-regular graphs by (1-15), and the numbers  $\{r_n\}$  of involutions in  $S_n$ , which satisfy  $r_n = r_{n-1} + (n-1)r_{n-2}$ , see Sloane [n.d., A000085]. Finally, *D-algebraic* GFs (also called *ADE* and *hyperalgebraic*) include Euler numbers by the equation (1-7).

**Theorem 2.1** (see e.g. R. P. Stanley [1999], Ch. 6).

*Rational*  $\subset$  *Algebraic*  $\subset$  *D-finite*  $\subset$  *D-algebraic*.

Here only the inclusion *Algebraic*  $\subset$  *D-finite* is nontrivial. The following observation explains the connection to the subject.

**Proposition 2.2.** *Sequences with D-algebraic GFs have Wilfian formulas of type (W1).*

In other words, if one wants to show that a sequence does not have a Wilfian formula, then proving that it is *D-transcendental*, i.e. non-D-algebraic, is a good start.<sup>3</sup> Unfortunately, even proving that a sequence is non-P-recursive is often challenging (see below).

**Example 2.3** (Bell numbers). Let  $B_n$  denotes the number of *set partitions* of  $\{1, \dots, n\}$ , see R. P. Stanley [ibid.] and Sloane [n.d., A000110]. Let

$$y(t) = \sum_{n=0}^{\infty} \frac{B_n t^n}{n!}, \quad z(t) = \sum_{n=0}^{\infty} B_n t^n$$

<sup>3</sup>To simplify exposition and for the lack of better terminology, here and in the future we refer to sequences by the properties of their GFs.

be the exponential and ordinary GFs of Bell numbers, respectively. On the one hand, we have:

$$y(t) = e^{e^t - 1}, \quad y''y - (y')^2 - y'y = 0.$$

Thus,  $y(t)$  is D-algebraic, and the proposition implies that  $\{B_n\}$  can be computed in  $\text{poly}(n)$  time. On the other hand,  $z(t)$  is D-transcendental by Klazar's theorem [Klazar \[2003\]](#).

This also implies that  $y(t)$  is not D-finite. Indeed, observe by definition, that if a sequence  $\{a_n\}$  is P-recursive, then so is  $\{n!a_n\}$ , which implies the result by taking  $a_n = B_n/n!$  (cf. [Lipshitz and Rubel \[1986\]](#)). Of course, there is a more direct way to prove that  $y(t)$  is not D-finite by repeated differentiation or via the asymptotics, see below. This suggests the following advanced generalization of Klazar's theorem.

**Open Problem 2.4** (P.–Yeliussizov). *Suppose  $\{a_n/n!\}$  is D-algebraic but not P-recursive. Does this imply that  $\{a_n\}$  is D-transcendental?*

Before we proceed to more combinatorial examples, let us mention that D-transcendental GFs are the subject of *Differential Galois Theory*, which goes back to Liouville, Lie, Picard and Vessiot in the 19th century (see e.g. [Ritt \[1950\]](#)), and continues to be developed [van der Put and Singer \[1997\]](#). Some natural GFs are known to be D-transcendental, e.g.  $\Gamma(z)$ ,  $\zeta(z)$ , etc., but there are too few methods to prove this in most cases of interest. Here are some of our favorite open problems along these lines, unapproachable with existing tools.

**Conjecture 2.5.**  $\sum_{n \geq 1} p_n t^n$  and  $\sum_{n \geq 1} \pi(n) t^n$  are D-transcendental.

Here  $p_n$  is  $n$ -th prime,  $\pi(n)$  is the number of primes  $\leq n$ , as above. Both GFs are known to be non-D-finite, as shown by [Flajolet, Gerhold, and Salvy \[2004/06\]](#) by asymptotic arguments. The authors quip: “*Almost anything is non-holonomic unless it is holonomic by design*”. Well, maybe so. But the same applies for D-transcendence where the gap between what we believe and what we can prove is much wider. The reader should think of such open problems as irrationality of  $e + \pi$  and  $\zeta(5)$ , and imagine a similar phenomenon in this case.

**Conjecture 2.6.**  $\sum_{n \geq 0} t^{n^3}$  is D-transcendental.

This problem should be compared with Jacobi's 1848 theorem that  $\sum_{n \geq 0} t^{n^2}$  is D-algebraic. To understand the difference, the conjecture is saying that there are no good formulas governing the number of ways to write  $n$  as a sum of  $k$  cubes, for any  $k$ , the kind of formulas that exist for sums of two, four and six squares, see [Hardy and Wright \[2008, §XX\]](#).

**2.2 Asymptotic tools.** The following result is the best tool we have for proving that a combinatorial sequence is not P-recursive. Note that deriving such asymptotics can be very difficult; we refer to [Flajolet and Sedgewick \[2009\]](#) and [Pemantle and M. C. Wilson \[2013\]](#) for recent comprehensive monographs on the subject.

**Theorem 2.7.** *Let  $\{a_n\}$  be a P-recursive sequence, s.t.  $a_n \in \mathbb{Q}$ ,  $C_1^n < a_n < C_2^n$  for some  $C_2 > C_1 > 0$  and all  $n \geq 1$ . Then:*

$$a_n \sim \sum_{i=1}^m K_i \lambda_i^n n^{\alpha_i} (\log n)^{\beta_i},$$

where  $K_i \in \mathbb{R}_+$ ,  $\lambda_i \in \overline{\mathbb{Q}}$ ,  $\alpha_i \in \mathbb{Q}$ , and  $\beta_i \in \mathbb{N}$ .

The theorem is a combination of several known results [Garrafrant and Pak \[2017\]](#). Briefly, the generating series  $\mathfrak{A}(t)$  is a  $G$ -function in a sense of Siegel (1929), which by the works of André, Bombieri, Chudnovsky, Dwork and Katz, must satisfy an ODE which has only regular singular points and rational exponents. We then apply the Birkhoff–Trjitzinsky claim/theorem, which in the regular case has a complete and self-contained proof in [Flajolet and Sedgewick \[2009\]](#) (see Theorem VII.10 and subsequent comments).

**Example 2.8** (Euler numbers  $E_n$ ). Recall that

$$E_n \sim \frac{4}{\pi} \left( \frac{2}{\pi} \right)^n n!$$

(see e.g. [Flajolet and Sedgewick \[ibid., p. 7\]](#)). Then  $\{E_n\}$  is not P-recursive, since otherwise  $E_n/n! \sim K\lambda^N$  with a transcendental exponent  $\lambda = (2/\pi) \notin \overline{\mathbb{Q}}$ .

**Example 2.9** ( $n$ -th prime  $p_n$ ). Following [Flajolet, Gerhold, and Salvy \[2004/06\]](#), recall that  $p_n = n \log n + n \log \log n + O(n)$ . Observe that the *harmonic number*  $h_n$  is P-recursive by definition:

$$h_n = h_{n-1} + \frac{1}{n} = 1 + \frac{1}{2} + \dots + \frac{1}{n} = \log n + O(1).$$

Then  $\{p_n\}$  is not P-recursive, since otherwise so is

$$p_n - n h_n = n \log \log n + O(n),$$

which is impossible by [Theorem 2.7](#).



**2.3 Lattice walks.** Let  $\Gamma = (V, E)$  be a graph and let  $v_0, v_1 \in V$  be two fixed vertices. Let  $a_n$  be the number of walks  $v_0 \rightarrow v_1$  in  $\Gamma$  of length  $n$ . This is a good model which leads to many interesting sequences. For example, Fibonacci number  $F_n$  is the number of walks  $1 \rightarrow 1$  of length  $n$  in the graph on  $\{1, 2\}$ , with edges  $(1, 1)$ ,  $(1, 2)$  and  $(2, 1)$ .

For general finite graphs we get C-recursive sequences  $\{a_n\}$  with rational GFs. For the usual walks  $0 \rightarrow 0$  on  $\mathbb{N}$  we get Catalan numbers  $a_{2n} = C_n$  as in (1-10), while for  $\pm 1$  walks in  $\mathbb{Z}$  we get  $a_{2n} = \binom{2n}{n}$ , both algebraic sequences. Similarly, for  $(0, \pm 1)$ ,  $(\pm 1, 0)$  walks in  $\mathbb{Z}^2$ , we get  $a_{2n} = \binom{2n}{n}^2$ , which is P-recursive but not algebraic. In higher dimensions or for more complicated graphs, there is no such neat formula.

**Theorem 2.10.** *Let  $S \subset \mathbb{Z}^d$  be a fixed finite set of steps, and let  $a_n$  be the number of walks  $O \rightarrow O$  in  $\mathbb{Z}^d$  of length  $n$ , with steps in  $S$ . Then  $\{a_n\}$  is P-recursive.*

This result is classical and follows easily from R. P. Stanley [1999, §6.3]. It suggests that to obtain more interesting sequences one needs to look elsewhere. Notably, one can consider natural lattice walks on some portion of  $\mathbb{Z}^d$ . There is a tremendous number of results in the literature, remarkable both in scope and beauty.

In recent years, M. Bousquet-Mélou and her coauthors initiated a broad study of the subject, and now have classified all walks in the first quadrant which start and end at the origin  $O$ , and have a fixed set  $S$  of steps with both coordinates in  $\{0, \pm 1\}$ . There are in principle  $2^8 - 1 = 255$  such walks, but some of them are trivial and some are the same up to symmetries. After the classification was completed, some resulting sequences are proved algebraic (say, *Kreweras walks* and *Gessel walks*), very surprisingly so, some are D-finite (not a surprise given Theorem 2.10), some are D-algebraic (this required development of new tools), and some are D-transcendental (it is amazing that this can be done at all).

**Example 2.11** (Case 16). Let  $S = \{(1, 1), (-1, -1), (-1, 0), (0, -1)\}$ , and let  $a_n$  be the number of walks  $O \rightarrow O$  in the first quadrant of length  $n$ , with steps in  $S$ , see Sloane [n.d., A151353]. It was shown in Bostan, Raschel, and Salvy [2014, Case 16] that

$$a_n \sim K \lambda^n n^\alpha,$$

where  $\lambda \approx 3.799605$  is a root of  $x^4 + x^3 - 8x^2 - 36x - 11 = 0$ , and  $\alpha \approx -2.318862$  satisfies  $c = -\cos(\pi/\alpha)$ , and  $c$  is a root of

$$y^4 - \frac{9}{2}y^3 + \frac{27}{4}y^2 - \frac{35}{8}y + \frac{17}{16} = 0$$

Since  $\alpha \notin \mathbb{Q}$ , Theorem 2.7 implies that  $\{a_n\}$  is not P-recursive.

We refer to Bousquet-Mélou [2006] and Bousquet-Mélou and Mishna [2010] for a comprehensive overview of the background and early stages of this far reaching project, and

to Bernardi, Bousquet-Mélou, and Raschel [2017] and Bostan, Bousquet-Mélou, Kauers, and Melczer [2016] for some recent developments which are gateways to references. Finally, let us mention a remarkable recent development Dreyfus, Hardouin, Roques, and Singer [2017], which proves D-transcendence for many families of lattice walks. Let us single out just one of the many results in that paper:

**Theorem 2.12** (Dreyfus, Hardouin, Roques, and Singer [ibid.], Thm. 5.8). *Sequence  $\{a_n\}$  defined in Example 2.11 is D-transcendental.*

In conclusion, let us mention that  $\{a_n\}$  can be computed in polynomial time straight from definition using dynamic programming, since the number of points reachable after  $n$  steps is  $\text{poly}(n)$ . This leads us to consider walks with constraints or graphs of superpolynomial growth.

**Conjecture 2.13.** *Let  $a_n$  denotes the number of self-avoiding walks  $O \rightarrow O$  in  $\mathbb{Z}^2$  of length  $n$ . Then sequence  $\{a_n\}$  has no Wilfian formula of type (W1).*

We refer to Guttmann [2009] for an extensive investigation of *self-avoiding walks* and its relatives, and the review of the literature.

**2.4 Walks on Cayley graphs.** Let  $G = \langle S \rangle$  be a finitely generated group  $G$  with a generating set  $S$ . Let  $a_n = a_n(G, S)$  be the number of words in  $S$  of length  $n$  equal to 1; equivalently, the number of walks  $1 \rightarrow 1$  of length  $n$ , in the Cayley graph  $\Gamma = \Gamma(G, S)$ . In this case  $\{a_n\}$  is called the *cogrowth sequence* and its GF  $A(t)$  the *cogrowth series*. They were introduced by Pólya in 1921 in probabilistic context of random walks on graphs, and by Kesten in the context of amenability Kesten [1959].

The cogrowth sequence  $\{a_n\}$  is C-recursive if only if  $G$  is finite Kouksov [1998]. It is algebraic for the dihedral group Humphries [1997], for the free group Haiman [1993] and for free products of finite groups Kuksov [1999], all with standard generators. The cogrowth sequence is P-recursive for many abelian groups Humphries [1997], and for the *Baumslag-Solitar groups*  $G = \text{BS}(k, k)$  in the standard presentation  $\text{BS}(k, \ell) = \langle x, y \mid x^k y = y x^\ell \rangle$ , see Elder, Rechnitzer, Janse van Rensburg, and Wong [2014].

**Theorem 2.14** (Garrafrant and Pak [2017]). *Sequence  $\{a_n(G, S)\}$  is not P-recursive for all symmetric  $S = S^{-1}$ , and the following classes of groups  $G$ :*

- (1) *virtually solvable groups of exponential growth with finite Prüfer rank,*
- (2) *amenable linear groups of superpolynomial growth,*
- (3) *groups of weakly exponential growth*

$$Ae^{n^\alpha} < \gamma_{G,S}(n) < Be^{n^\beta},$$

where  $A, B > 0$ , and  $0 < \alpha, \beta < 1$ ,

(4) the Baumslag–Solitar groups  $BS(k, 1)$ ,  $k \geq 2$ ,

(5) the lamplighter groups  $L(d, H) = H \wr \mathbb{Z}^d$ , where  $H$  is a finite abelian group and  $d \geq 1$ .

Since  $G \simeq \mathbb{Z} \ltimes \mathbb{Z}^2$  with a free action of  $\mathbb{Z}$ , is linear of exponential growth, by (2) we obtain a solution to the question originally asked by Kontsevich, see [R. Stanley \[2014\]](#).

**Corollary 2.15** ([Garrabrant and Pak \[2017\]](#)). *There is a linear group  $G$  and a symmetric generating set  $S$ , s.t. the sequence  $\{a_n(G, S)\}$  is not P-recursive.*

The proof of [Theorem 2.14](#) is a combination of many results by different authors. For example, for  $G = BS(k, 1)$ ,  $k \geq 2$ , and every symmetric  $\langle S \rangle = G$ , there exist  $C_1, C_2 > 0$  which depend on  $S$ , s.t.

$$(2-1) \quad |S|^n e^{-C_1 \sqrt[3]{n}} \leq a_n(G, S) \leq |S|^n e^{-C_2 \sqrt[3]{n}},$$

see [Woess \[2000, §15.C\]](#). The result now follows from [Theorem 2.7](#).

It may seem from [Theorem 2.14](#) that the properties of  $\{a_n(G, S)\}$  depend only on  $G$ , but that is false. In fact, for  $G = F_k \times F_\ell$  there are generating sets with both P-recursive and non-P-recursive sequences [Garrabrant and Pak \[2017\]](#). For groups in the theorem, this is really a byproduct of probabilistic tools used in establishing the asymptotics such as (2-1). In fact, the probabilities of return of the random walk  $a_n(G, S)/|S|^n$  always have the same growth under *quasi-isometry*, see e.g. [Woess \[2000\]](#).<sup>4</sup>

In a forthcoming paper [Garrabrant and Pak \[n.d.\]](#) we construct an explicit but highly artificial non-symmetric set  $S \subset F_k \times F_\ell$  with D-transcendental cogrowth sequence. In [Kassabov and Pak \[n.d.\]](#) we use the tools in [Kassabov and Pak \[2013\]](#) to prove that groups have an uncountable set of *spectral radii*

$$\rho(G, S) := \lim_{n \rightarrow \infty} a_n(G, S)^{1/n}.$$

Since the set of D-algebraic sequence is countable, this implies the existence of D-transcendental Cayley graphs with symmetric  $S$ , but such proof is nonconstructive.

**Open Problem 2.16.** *Find an explicit construction of  $\Gamma(G, S)$  when  $S$  is symmetric, and  $\{a_n(G, S)\}$  is D-transcendental.*

Sequences  $\{a_n\}$  have been computed in very few special cases. For example, for  $\text{PSL}(2, \mathbb{Z}) = \mathbb{Z}_2 * \mathbb{Z}_3$  with the natural symmetric generating set, the cogrowth series

---

<sup>4</sup>While the leading term in the asymptotics remains the same, lower order terms can change for different  $S$ , see [Woess \[2000, §17.B\]](#).

$A(t)$  is computed in [Kuksov \[1999\]](#):

$$A(t) = \frac{(1+t)(-t+t^2-8t^3+3t^4-9t^5+(2-t+6t^2)\sqrt{\mathcal{R}(t)})}{2(1-3t)(1+3t^2)(1+3t+3t^2)(1-t+3t^2)},$$

where  $\mathcal{R}(t) = 1 - 2t + t^2 - 6t^3 - 8t^4 - 18t^5 + 9t^6 - 54t^7 + 81t^8$ .

There are more questions than answers here. For example, can cogrowth sequence be computed for nilpotent groups?

Before we conclude, let us note that everywhere above we are implicitly assuming that  $G$  either has a faithful rational representation, e.g.  $G = \text{BS}(k, 1)$  as in (4) above, or more generally has the *word problem* solvable in polynomial time (cf. [Lipton and Zalcstein \[1977\]](#)). The examples include the *Grigorchuk group*  $\mathbb{G}$ , which is an example of 3, see [Grigorchuk and Pak \[2008\]](#) and the lamplighter groups  $L(d, H)$  as in (5). Note that in general the word problem can be superpolynomial or even unsolvable, see e.g. [C. F. Miller I. \[1992\]](#), in which case  $\{a_n\}$  is no longer a combinatorial sequence.

**2.5 Partitions.** Let  $p(n)$  be the number of integer partitions of  $n$ , as in (1-13). We have the *Hardy–Ramanujan formula*:

$$(2-2) \quad p(n) \sim \frac{1}{4n\sqrt{3}} e^{\pi\sqrt{\frac{2n}{3}}} \quad \text{as } n \rightarrow \infty.$$

(see e.g. [Flajolet and Sedgewick \[2009\]](#), p. VIII.6). [Theorem 2.7](#) implies that  $\{p(n)\}$  is not P-recursive. On the other hand, it is known that

$$F(t) := \sum_{n=0}^{\infty} p(n)t^n = \prod_{i=1}^{\infty} \frac{1}{1-t^i}$$

satisfies the following ADE:

$$4F^3 F'' + 5t F^3 F''' + t^2 F^3 F^{(4)} - 16F^2 (F')^2 - 15t F^2 F' F'' - 39t^2 F^2 (F'')^2 + 10t F (F')^3 + 12t^2 F (F')^2 F'' + 6t^2 (F')^4 = 0$$

(cf. [Zagier \[2008\]](#)). A quantitative version of [Proposition 2.2](#) then implies that  $\{p(n)\}$  can be computed in time  $O^*(n^{4.5})$ , where  $O^*$  indicates  $\log n$  terms. For comparison, the dynamic programming takes  $O(n^{2.5})$  time, where  $O(\sqrt{n})$  comes as the cost of addition. Similarly, *Euler's recurrence* famously used by MacMahon (1915) to compute  $p(200)$ , gives an  $O(n^2)$  algorithm:

$$p(n) = p(n-1) + p(n-2) - p(n-5) - p(n-7) + p(n-12) + p(n-15) - \dots$$

(cf. [Calkin, Davis, James, Perez, and Swannack \[2007\]](#)). There is also an efficient implementation [Johansson \[2012\]](#) based on the Hardy–Ramanujan–Rademacher sharp asymptotic formula which extends (2-2) to  $o(1)$  additive error. It would be interesting to analyze this algorithm perhaps using Lehmer’s estimates used in [DeSalvo and Pak \[2015\]](#).

Now, for a subset  $\mathcal{Q} \subseteq \{1, 2, \dots\}$ , denote by  $p_{\mathcal{Q}}(n)$  the number of partitions of  $n$  into parts in  $\mathcal{Q}$ . The dynamic programming algorithm is easy to generalize to every  $\{p_{\mathcal{Q}}(n)\}$  where the membership  $a \in \mathcal{Q}$  can be decided in  $\text{poly}(\log a)$  time, giving a Wilfian formula of type (W1). This is polynomially optimal for partitions into primes [Sloane \[n.d., A000607\]](#) or squares [Sloane \[ibid., A001156\]](#), but not for sparse sequences.

**Proposition 2.17.** *Let  $\mathcal{Q} = \{a_1, a_2, \dots\}$ , such that  $a_k \geq c^k$ , for some  $c > 1$  and all  $k \geq 1$ . Then  $p_{\mathcal{Q}}(n) = n^{O(\log n)}$ .*

Thus,  $p_{\mathcal{Q}}(n)$  as in the proposition could in principle have a Wilfian formula of type (W3). Notable examples include the number  $q(n)$  of *binary partitions* (partitions of  $n$  into powers of 2), see [Sloane \[ibid., A000123\]](#), *partitions into Fibonacci numbers* [Sloane \[ibid., A003107\]](#), and *s-partitions* defined as partitions into  $\{1, 3, 7, \dots, 2^k - 1, \dots\}$  [Sloane \[ibid., A000929\]](#).

**Theorem 2.18** ([Pak and Yeliussizov \[n.d.\]](#)). *Let  $\mathcal{Q} = \{a_1, a_2, \dots\}$ , and suppose  $a_k/a_{k-1}$  is an integer  $\geq 2$ , for all  $k > 1$ . Then  $\{p_{\mathcal{Q}}(n)\}$  can be computed in time  $\text{poly}(\log n)$ .*

This covers binary partitions, partitions into factorials [Sloane \[n.d., A064986\]](#), etc. We conjecture that partitions into Fibonacci numbers and s-partitions also have Wilfian formulas of type (W3). Cf. [N. Robbins \[1996\]](#) for an algorithm for partitions into distinct Fibonacci numbers. Other partitions functions such as partitions into Catalan numbers [Sloane \[n.d., A033552\]](#) and partitions into partition numbers [Sloane \[ibid., A007279\]](#), could prove less tractable. We should mention that connection between algebraic properties of GFs and complexity goes only one way:

**Theorem 2.19.** *The sequence  $\{q(n)\}$  of the number of binary partitions is D-transcendental.*

This follows from the *Mahler equation*

$$Q(t) - tQ(t) - Q(t^2) = 0, \quad \text{where} \quad Q(t) = \sum_{n=0}^{\infty} q(n)t^n,$$

see [Dreyfus, Hardouin, and Roques \[2015\]](#). We conjecture that  $\{a_n\}$  and  $\{b_n\}$  from [Conjecture 1.3](#) satisfy similar functional equations, and are also D-transcendental.

**2.6 Pattern avoidance.** Let  $\sigma \in S_n$  and  $\omega \in S_k$ . Permutation  $\sigma$  is said to *contain* the pattern  $\omega$  if there is a subset  $X \subseteq \{1, \dots, n\}$ ,  $|X| = k$ , such that  $\sigma|_X$  has the same relative order as  $\omega$ . Otherwise,  $\sigma$  is said to *avoid*  $\omega$ .

Fix a set of patterns  $\mathcal{F} \subset S_k$ . Denote by  $A_n(\mathcal{F})$  the number of permutations  $\sigma \in S_n$  *avoiding* all patterns  $\omega \in \mathcal{F}$ . The sequence  $\{A_n(\mathcal{F})\}$  is the fundamental object of study in the area of *pattern avoidance*, extensively analyzed from analytic, asymptotic and combinatorial points of view.

The subject was initiated by MacMahon (1915) and Knuth (1973), who showed that  $A_n(123) = A_n(213) = C_n$ , the  $n$ -th Catalan number (1-10). The *Erdős–Szekeres theorem* (1935) on longest increasing and decreasing subsequences in a permutation can also be phrased in this language:  $A_n(12 \cdots k, \ell \cdots 21) = 0$ , for all  $n > (k-1)(\ell-1)$ .

To give a flavor of subsequent developments, let us mention a few more of our most favorite results. Simion–Schmidt (1985) proved  $A_n(123, 132, 213) = F_{n+1}$ , the Fibonacci numbers. Similarly, Shapiro–Stephens (1991) proved  $A_n(2413, 3142) = S(n)$ , the Schröder numbers Sloane [n.d., A006318]. The celebrated Marcus–Tardos theorem Marcus and Tardos [2004] states that  $\{A_n(\omega)\}$  is at most exponential, for all  $\omega \in S_k$ , with a large base of exponent for random  $\omega \in S_k$  Fox [2013]. We refer to Kitaev [2011], Klazar [2010], and Vatter [2015] for many results on the subject, history and background.

The *Noonan–Zeilberger conjecture* Noonan and Zeilberger [1996], first posed as a question by Gessel [1990], stated that the sequence  $\{A_n(\mathcal{F})\}$  is P-recursive for all  $\mathcal{F} \subset S_k$ . It was recently disproved:

**Theorem 2.20** (Garrabrant and Pak [2015]). *There is  $\mathcal{F} \subset S_{80}$ ,  $|\mathcal{F}| < 30,000$ , such that  $\{A_n(\mathcal{F})\}$  is not P-recursive.*

We extend this result in a forthcoming paper Garrabrant and Pak [n.d.], where we construct a D-transcendent pattern avoiding sequence  $\{A_n(\mathcal{F})\}$ , for some  $\mathcal{F} \subset S_{80}$ . Both proofs involve embedding of Turing Machines into the problem modulo 2. We also prove the following result on complexity of counting pattern avoiding permutations, our only result forbidding Wilfian formulas:

**Theorem 2.21** (Garrabrant and Pak [2015]). *If  $\text{EXP} \neq \oplus \text{EXP}$ , then  $A_n(\mathcal{F}) \bmod 2$  cannot be computed in  $\text{poly}(n)$  time.*

In other words, counting parity of pattern avoiding permutations is likely hard. We conjecture that  $A_n(\mathcal{F})$  is  $\# \text{EXP}$ -complete, but we are not very close to proving this.

**Theorem 2.22** (Garrabrant and Pak [ibid.]). *The problem whether  $A_n(\mathcal{F}) = A_n(\mathcal{F}')$  mod 2 for all  $n$ , is undecidable.*

The theorem implies that in some cases even a large amount of computational evidence in pattern avoidance is misleading. For example, there exists two sets of patterns  $\mathfrak{F}, \mathfrak{F}' \in S_k$ , so that the first time they have different parity is for  $n > \text{tower of } 2\text{s of length } 2^k$ .

Finally, let us mention an ongoing effort to find a small set of patterns  $\mathfrak{F}$ , so that  $\{A_n(\mathfrak{F})\}$  is not P-recursive. Is one permutation enough? It is known that  $\{A_n(1342)\}$  is algebraic [Bóna \[1997\]](#), while  $\{A_n(1234)\}$  is P-recursive [Gessel \[1990\]](#). One of the most challenging problems is to analyze  $\{A_n(1324)\}$ , the only 4-pattern remaining. The asymptotics obtained experimentally in [Conway, Guttman, and Zinn-Justin \[2018\]](#) based on the values for  $n \leq 50$ , suggests:

$$A_n(1324) \sim B \lambda^n \mu^{\sqrt{n}} n^\alpha,$$

where  $\lambda = 11.600 \pm 0.003$ ,  $\mu = 0.0400 \pm 0.0005$ ,  $\alpha = -1.1 \pm 0.1$ . If true, [Theorem 2.7](#) would imply that  $\{A_n(1324)\}$  is not P-recursive. While this remains out of reach, the following problem could be easier.

**Open Problem 2.23.** *Can  $\{A_n(1324)\}$  be computed in  $\text{poly}(n)$  time?<sup>5</sup> More generally, can one find a single permutation  $\pi$  such that  $\{A_n(\pi)\}$  cannot be computed in  $\text{poly}(n)$  time? Is the computation of  $\{A_n(\pi)\}$  easier or harder for random permutations  $\pi \in S_k$ ?*

### 3 Bijections

**3.1 Counting and sampling via bijections.** There is an ocean of bijections between various combinatorial objects. They have a variety of uses: to establish a theorem, to obtain refined counting, to simplify the proof, to make the proof amenable for generalizations, etc. Last but not least, some especially beautiful bijections are often viewed as a piece of art, an achievement in its own right, a result to be taught and admired.

From the point of view of this survey, bijections  $\varphi : \mathcal{R}_n \rightarrow \mathcal{B}_n$  are simply algorithms which require complexity analysis. There are two standard applications of such bijections. First, their existence allows us to reduce counting of  $\{|\mathcal{R}_n|\}$  to counting of  $\{|\mathcal{B}_n|\}$ . For example, the classical *Prüfer's algorithm* allows counting of spanning trees in  $K_n$ , reducing it to Cayley's formula (1-12).

Second and more recent application is to *random sampling* of combinatorial objects. Oftentimes, one of the sets has a much simpler structure which allows (nearly) uniform sampling. To compare the resulting algorithm with other competing approaches one then needs a worst case and/or average case analysis of the complexity of the bijection.

<sup>5</sup>In 2005, Doron Zeilberger expressed doubts that  $A_{1000}(1324)$  can be computed even by Hashem. This sentiment has been roundly criticized on both mathematical and theological grounds (see [Steingrímsson \[2013\]](#)).

Of course, most bijections in the literature are so straightforward that their analysis is elementary, think of the Prüfer’s algorithm or the classical “plane trees into binary trees” bijection [de Bruijn and Morselt \[1967\]](#). But this is also what makes them efficient. For example, the bijections for planar maps are amazing in their elegance, and have some important applications to statistical physics; we refer to [Schaeffer \[2015\]](#) for an extensive recent survey and numerous references.

Finally, we should mention a number of *perfect sampling* algorithms, some of which in the right light can also be viewed as bijections. These include most notably general techniques such as *Boltzmann samplers* [Duchon, Flajolet, Louchard, and Schaeffer \[2004\]](#) and *coupling from the past* [D. A. Levin, Peres, and Wilmer \[2017\]](#). Note also two beautiful ad hoc algorithms: *Wilson’s LERW* [D. B. Wilson \[1996\]](#) and the *Aldous–Broder algorithm* for sampling uniform spanning trees in a graph (both of which are highly nontrivial already for  $K_n$ ), see e.g. [D. A. Levin, Peres, and Wilmer \[2017\]](#).

**3.2 Partition bijections.** Let  $q(n)$  denote the number of *concave partitions* defined by  $\lambda_i - \lambda_{i+1} \geq \lambda_{i+1} - \lambda_{i+2}$  for all  $i$ . Then  $\{q(n)\}$  can be computed in  $\text{poly}(n)$  time. To see this, recall *Corteel’s bijection* between convex partitions and partitions into triangular numbers [Sloane \[n.d., A007294\]](#). We then have:

$$\sum_{n=1}^{\infty} q(n)t^n = \prod_{k=2}^{\infty} \frac{1}{1 - t^{\binom{k}{2}}},$$

see [Canfield, Corteel, and Hitczenko \[2001\]](#). This bijection can be described as a linear transformation which can be computed in polynomial time [Corteel and Savage \[2004\]](#) and [Pak \[2004a\]](#). More importantly, the bijections allow random sampling of concave partitions, leading to their limit shape [Canfield, Corteel, and Hitczenko \[2001\]](#) and [DeSalvo and Pak \[2016\]](#).

On the opposite extreme, there is a similar *Hickerson’s bijection* between  $s$ -partitions and partitions with  $\lambda_i \geq 2\lambda_{i+1}$  for all  $i \geq 1$ , see [Canfield, Corteel, and Hitczenko \[2001\]](#) and [Pak \[2004a\]](#). Thus, both sets are equally hard to count, but somehow this makes the problem more interesting.

The [Garsia and Milne \[1981\]](#) celebrated *involution principle* combines the Schur and Sylvester’s bijections in an iterative manner, giving a rather complicated bijective proof of the *Rogers–Ramanujan identity*:

$$(3-1) \quad 1 + \sum_{k=1}^{\infty} \frac{t^{k^2}}{(1-t)(1-t^2)\cdots(1-t^k)} = \prod_{i=0}^{\infty} \frac{1}{(1-t^{5i+1})(1-t^{5i+4})}.$$

To be precise, they constructed a bijection  $\Psi_n : \mathcal{P}_n \rightarrow \mathcal{Q}_n$ , where  $\mathcal{P}$  is the set of partitions into parts  $\lambda_i \geq \lambda_{i+1} + 2$ , and  $\mathcal{Q}$  is the set of partitions into parts  $\pm 1 \pmod{5}$ . In [Pak](#)



[2006, §8.4.5] we conjecture that  $\Psi_n$  requires  $\exp n^{\Omega(1)}$  iterations in the worst case. Partial evidence in favor of this conjecture is our analysis of O’Hara’s bijection in [Konvalinka and Pak \[2009\]](#), with a  $\exp \Omega(\sqrt[3]{n})$  worst case lower bound. On the other hand, the iterative proof in [Boulet and Pak \[2006\]](#) for (3-1) requires only  $O(n)$  iterations.

**3.3 Plane partitions and Young tableaux.** Denote by  $sp(n)$  the number of 3-dimensional (or solid) partitions. MacMahon famously proved in 1912 that

$$\sum_{n=0}^{\infty} sp(n) t^n = \prod_{k=1}^{\infty} \frac{1}{(1-t^k)^k},$$

which gives a  $poly(n)$  time algorithm for computing  $\{sp(n)\}$ . A variation on the celebrated *Hillman-Grassl* and *RSK* bijections proves this result and generalizes it [Pak \[2006\]](#). The application to sampling of this bijection have been analyzed in [Bodini, Fusy, and Pivoteau \[2010\]](#). On the other hand, there is a strong evidence that the RSK-based algorithms cannot be improved. While we are far from proving this, let us note that in [Pak and Vallejo \[2010\]](#) we show linear time reductions between all major bijections in the area, so a speedup of one of them implies a speedup of all.

A remarkable *Krattenthaler’s bijection* allows enumerations of solid partitions which fit into  $[n \times n \times n]$  box [Krattenthaler \[1999\]](#). This bijection is based on top of the *NPS algorithm*, which has also been recently analyzed [Neumann and Sulzgruber \[2015\]](#) and [Schneider and Sulzgruber \[2017\]](#). Curiously, there are no analogous results in  $d \geq 4$  dimensions, making counting such  $d$ -dimensional partitions an open problem (cf. [Govindarajan \[2013\]](#)).

**3.4 Complexity of bijections.** Let us now discuss the questions (2’) in the introduction, about the nature of bijections  $\varphi : \mathcal{P}_n \rightarrow \mathcal{Q}_n$  from an algorithmic point of view.

If we think of  $\varphi$  as a map, we would want both  $\varphi$  and  $\varphi^{-1}$  to be computable in polynomial time. If that’s all we want, it is not hard to show that such  $\varphi$  can always be constructed whenever there is a polynomial time algorithm to compute  $|\mathcal{P}_n| = |\mathcal{Q}_n|$ . For example, the dynamic programming plus the *divide and conquer* allows a construction of  $poly(n)$  time bijection  $\varphi_n : \mathcal{P}_n \rightarrow \mathcal{Q}_n$ , proving Rogers–Ramanujan identity (3-1). Since such construction would require prior knowledge of  $|\mathcal{P}_n| = |\mathcal{Q}_n|$ , from a combinatorial point of view this is unsatisfactory.

Alternatively, one can think of a bijection as an algorithm which computes a given map  $\varphi_n$  as above in  $poly(n)$  time. This is a particularly unfriendly setting as one would essentially need to prove new *lower bounds* in complexity. Worse, we proved in [Konvalinka and Pak \[2009\]](#) that in some cases O’Hara’s algorithm requires superpolynomial

time, while the map given by the algorithm can be computed in  $\text{poly}(n)$  time using *integer programming*. Since this is the only nice bijective proof of the *Andrews identities* that we know (see Pak [2006]), this suggests that either we don't understand the nature of these identities or have a very restrictive view on what constitutes a combinatorial bijection. Or, perhaps, the complexity approach is simply inapplicable in this combinatorial setting.

There are other cases of unquestionably successful bijections which are inferior to other algorithms from complexity point of view. For example, stretching the definitions a bit, Wilson's LERW algorithm for generating random (directed) spanning trees requires exponential time on directed graphs D. B. Wilson [1996], while a straightforward algorithm based on the *matrix-tree theorem* is polynomial, of course.

Finally, even when the bijection is nice and efficient, it might still have no interesting properties, so the only application is the proof of the theorem. One example is an iterative bijection for the Rogers–Ramanujan identity (3-1) which is implied by the proof in Boulet and Pak [2006]. We don't know if it respects any natural statistics which would imply a stronger result. Thus, we left it in the form of a combinatorial proof to make the underlying algebra clear.

**3.5 Probabilistic/asymptotic approach.** Suppose both sets of combinatorial objects  $\mathcal{P}_n$  and  $\mathcal{Q}_n$  have well-defined *limit shapes*  $\pi$  and  $\omega$ , as  $n \rightarrow \infty$ . Such limits exists for various families of trees Drmota [2009], graphs Lovász [2012], partitions DeSalvo and Pak [2016], permutations Hoppen, Kohayakawa, Moreira, Ráth, and Menezes Sampaio [2013], solid partitions Okounkov [2016], Young tableaux Romik [2015], etc.<sup>6</sup> For a sequence  $\{\varphi_n\}$  of bijections  $\varphi_n : \mathcal{P}_n \rightarrow \mathcal{Q}_n$ , one can ask about the *limit bijection*  $\Phi : \pi \rightarrow \omega$ , defined as  $\lim_{n \rightarrow \infty} \varphi_n$ . We can then require that  $\Phi$  satisfies certain additional properties. This is the approach taken in Pak [2004b] to prove the following result:

**Theorem 3.1.** *The Rogers–Ramanujan identity (3-1) has no geometric bijection.*

Here the *geometric bijections* are defined as compositions of certain piecewise  $\text{GL}(2, \mathbb{Z})$  maps acting on Ferrers diagrams, which are viewed as subsets of  $\mathbb{Z}^2$ . We first prove that the limits of such bijections are *asymptotically stable*, i.e. act piecewise linearly on the limit shapes. The rest of the proof follows from existing results on the limit shapes  $\pi$  and  $\omega$  on both sides of (3-1), which forbid a piecewise linear map  $\Phi : \pi \rightarrow \omega$ , see DeSalvo and Pak [2016].

The next story is incomplete, yet the outlines are becoming clear. Let  $\text{ASM}(n)$  be the number of *alternating sign matrices* of order  $n$ , defined as the number of  $n \times n$  matrices where every row has entries in  $\{0, \pm 1\}$ , with row and column sums equal to 1, and all signs alternate in each row and column. Let  $\text{FSLT}(n)$  be the number of the *fully symmetric*

<sup>6</sup>Here the notion of a “limit shape” is used very loosely as it means very different things in each case.

*lozenge tilings*, defined as lozenge tilings of the regular  $2n$ -hexagon with the full group of symmetries  $D_6$ . Such tilings are in easy bijection with solid partitions which fit into  $[2n \times 2n \times 2n]$  box, have full group of symmetries  $S_3$ , and are self-complementary within the box (cf. [Section 3.3](#)). Finally, let  $\text{TSPP}(n)$  be the number of *triangular shifted plane partitions* defined as plane partitions  $(b_{ij})_{1 \leq i \leq j}$  of shifted shape  $(n-1, n-2, \dots, 1)$ , and entries  $n-i \leq b_{ij} \leq n$  for  $1 \leq i \leq j \leq n-1$ .

The following identity is justly celebrated:

$$(3-2) \quad \text{ASM}(n) = \text{FSLT}(n) = \text{TSPP}(n) = \frac{1! 4! 7! \cdots (3n-2)!}{n! (n+1)! \cdots (2n-1)!}$$

Here the second equality is known to have a bijective proof [Mills, D. P. Robbins, and Rumsey \[1986\]](#). Finding bijective proof of the third equality is a major open problem. See [Bressoud \[1999\]](#) and [Krattenthaler \[2016\]](#) for the history of the problem and [Sloane \[n.d., A005130\]](#) for further references.

**Claim 3.2.** *The equality  $\text{ASM}(n) = \text{FSLT}(n)$  has no geometric bijection.*

We now know (conjecturally) what the *frozen regions* in each case are: the circle for FSLTs and a rather involved sextic equation for ASMs. The latter is an ingenious conjecture in [Colomo and Pronko \[2010\]](#) (see also [Colomo and Sportiello \[2016\]](#)), while the former is a natural conjecture about the *Arctic Circle* which remains when the symmetries are introduced (cf. [Panova \[2015\]](#)).<sup>7</sup> We are not sure in this case what do we mean by a “geometric bijection”. But any natural definition should imply that the two shapes are incompatible. It would be interesting to formalize this even before both frozen regions are fully established.

There is another aspect of this asymptotic approach, which allows us to distinguish between different equinumerous collections of combinatorial objects with respect to some (transitive) notions of a “good” (canonical) bijection, and thus divide them into equivalence classes. This method would allow us to understand the nature of these families and ignore superficial differences within the same class.

The prototypical example of this is a collection of over 200 objects enumerating Catalan numbers [R. P. Stanley \[2015\]](#), but there are other large such collections: for Motzkin numbers, Schröder numbers, Euler numbers (1-11), etc. A natural approach would be to use the symmetry properties or the topology, but such examples are rare (see, however, [Armstrong, Stump, and Thomas \[2013\]](#) and [West \[1995\]](#) for two “canonical” bijections between Catalan objects).

<sup>7</sup>While the frozen region hasn’t been established for FSLTs, it is known that if exists it must be a circle (Greta Panova, personal communication).

In [Miner and Pak \[2014\]](#), we studied the limit averages of permutation matrices corresponding to  $A_n(\mathcal{F})$ . We showed that the limit surfaces corresponding to  $A_n(123)$  and  $A_n(213)$  are quite different, even though their sizes are Catalan numbers (see also [Hoffman, Rizzolo, and Slivken \[2017\]](#) and [Madras and Pehlivan \[2016\]](#)). This partly explains a well known phenomenon: there are *nine*(!) different bijections between these two families described in [Kitaev \[2011\]](#), each with its own special properties – there is simply no “canonical” bijection in this case. See also [Dokos and Pak \[2014\]](#) for the analysis of another interesting family of Catalan objects.

**Acknowledgments.** We are very grateful to Matthias Aschenbrenner, Artëm Chernikov, Stephen DeSalvo, Persi Diaconis, Sam Dittmer, Jacob Fox, Scott Garrabrant, Bon-Soon Lin, Danny Nguyen, Pasha Pylyavskyy, Vic Reiner, Richard Stanley and Jed Yang for many helpful conversations. Special thanks to Alejandro Morales, Greta Panova and Damir Yeliussizov for many collaborations, for reading the first draft of this paper, their comments and useful suggestions.

## References

- Drew Armstrong, Christian Stump, and Hugh Thomas (2013). “A uniform bijection between nonnesting and noncrossing partitions”. *Trans. Amer. Math. Soc.* 365.8, pp. 4121–4151. MR: [3055691](#) (cit. on p. [3192](#)).
- László Babai (1995). “Automorphism groups, isomorphism, reconstruction”. In: *Handbook of combinatorics, Vol. 1, 2*. Elsevier Sci. B. V., Amsterdam, pp. 1447–1540. MR: [1373683](#) (cit. on pp. [3175](#), [3176](#)).
- (2016). “Graph isomorphism in quasipolynomial time [extended abstract]”. In: *STOC 2016 — Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, pp. 684–697. MR: [3536606](#) (cit. on p. [3175](#)).
- F. Bergeron, G. Labelle, and P. Leroux (1998). *Combinatorial species and tree-like structures*. Vol. 67. Encyclopedia of Mathematics and its Applications. Translated from the 1994 French original by Margaret Readdy, With a foreword by Gian-Carlo Rota. Cambridge University Press, Cambridge, pp. xx+457. MR: [1629341](#) (cit. on p. [3176](#)).
- Olivier Bernardi, Mireille Bousquet-Mélou, and Kilian Raschel (Aug. 2017). “Counting quadrant walks via Tutte’s invariant method”. arXiv: [1708.08215](#) (cit. on p. [3183](#)).
- Olivier Bodini, Éric Fusy, and Carine Pivoteau (2010). “Random sampling of plane partitions”. *Combin. Probab. Comput.* 19.2, pp. 201–226. MR: [2593621](#) (cit. on p. [3190](#)).
- Miklós Bóna (1997). “Exact enumeration of 1342-avoiding permutations: a close link with labeled trees and planar maps”. *J. Combin. Theory Ser. A* 80.2, pp. 257–272. MR: [1485138](#) (cit. on p. [3188](#)).

- Alin Bostan, Mireille Bousquet-Mélou, Manuel Kauers, and Stephen Melczer (2016). “On 3-dimensional lattice walks confined to the positive octant”. *Ann. Comb.* 20.4, pp. 661–704. MR: [3572381](#) (cit. on p. [3183](#)).
- Alin Bostan, Kilian Raschel, and Bruno Salvy (2014). “Non-D-finite excursions in the quarter plane”. *J. Combin. Theory Ser. A* 121, pp. 45–63. MR: [3115331](#) (cit. on p. [3182](#)).
- Cilanne Boulet and Igor Pak (2006). “A combinatorial proof of the Rogers-Ramanujan and Schur identities”. *J. Combin. Theory Ser. A* 113.6, pp. 1019–1030. MR: [2244131](#) (cit. on pp. [3190](#), [3191](#)).
- Mireille Bousquet-Mélou (2006). “Rational and algebraic series in combinatorial enumeration”. In: *International Congress of Mathematicians. Vol. III*. Eur. Math. Soc., Zürich, pp. 789–826. MR: [2275707](#) (cit. on p. [3182](#)).
- Mireille Bousquet-Mélou and Marni Mishna (2010). “Walks with small steps in the quarter plane”. In: *Algorithmic probability and combinatorics*. Vol. 520. Contemp. Math. Amer. Math. Soc., Providence, RI, pp. 1–39. MR: [2681853](#) (cit. on p. [3182](#)).
- Mireille Bousquet-Mélou and Gilles Schaeffer (2002). “Walks on the slit plane”. *Probab. Theory Related Fields* 124.3, pp. 305–344. MR: [1939650](#).
- David M. Bressoud (1999). *Proofs and confirmations*. MAA Spectrum. The story of the alternating sign matrix conjecture. Mathematical Association of America, Washington, DC; Cambridge University Press, Cambridge, pp. xvi+274. MR: [1718370](#) (cit. on p. [3192](#)).
- N. G. de Bruijn and B. J. M. Morselt (1967). “A note on plane trees”. *J. Combinatorial Theory* 2, pp. 27–34. MR: [0205872](#) (cit. on p. [3189](#)).
- Neil Calkin, Jimena Davis, Kevin James, Elizabeth Perez, and Charles Swannack (2007). “Computing the integer partition function”. *Math. Comp.* 76.259, pp. 1619–1638. MR: [2299791](#) (cit. on p. [3186](#)).
- Rod Canfield, Sylvie Corteel, and Pawel Hitczenko (2001). “Random partitions with non-negative  $r$ -th differences”. *Adv. in Appl. Math.* 27.2-3. Special issue in honor of Dominique Foata’s 65th birthday (Philadelphia, PA, 2000), pp. 298–317. MR: [1868967](#) (cit. on p. [3189](#)).
- F. Colomo and A. G. Pronko (2010). “The limit shape of large alternating sign matrices”. *SIAM J. Discrete Math.* 24.4, pp. 1558–1571. MR: [2746708](#) (cit. on p. [3192](#)).
- F. Colomo and A. Sportiello (2016). “Arctic curves of the six-vertex model on generic domains: the tangent method”. *J. Stat. Phys.* 164.6, pp. 1488–1523. MR: [3541190](#) (cit. on p. [3192](#)).
- Andrew R. Conway, Anthony J. Guttmann, and Paul Zinn-Justin (2018). “1324-avoiding permutations revisited”. *Adv. in Appl. Math.* 96, pp. 312–333. arXiv: [1709.01248](#). MR: [3767512](#) (cit. on p. [3188](#)).
- Sylvie Corteel and Carla D. Savage (2004). “Partitions and compositions defined by inequalities”. *Ramanujan J.* 8.3, pp. 357–381. MR: [2111689](#) (cit. on p. [3189](#)).

- Jesús A. De Loera, Jörg Rambau, and Francisco Santos (2010). *Triangulations*. Vol. 25. Algorithms and Computation in Mathematics. Structures for algorithms and applications. Springer-Verlag, Berlin, pp. xiv+535. MR: [2743368](#) (cit. on p. [3177](#)).
- Stephen DeSalvo and Igor Pak (2015). “[Log-concavity of the partition function](#)”. *Ramanujan J.* 38.1, pp. 61–73. MR: [3396486](#) (cit. on p. [3186](#)).
- (Nov. 2016). “[Limit shapes via bijections](#)”. arXiv: [1611.06073](#) (cit. on pp. [3189](#), [3191](#)).
- Theodore Dokos and Igor Pak (2014). “The expected shape of random doubly alternating Baxter permutations”. *Online J. Anal. Comb.* 9, p. 12. MR: [3238333](#) (cit. on p. [3193](#)).
- Thomas Dreyfus, Charlotte Hardouin, and Julien Roques (July 2015). “[Hypertranscendence of solutions of Mahler equations](#)”. arXiv: [1507.03361](#) (cit. on p. [3186](#)).
- Thomas Dreyfus, Charlotte Hardouin, Julien Roques, and Michael F. Singer (Feb. 2017). “[On the nature of the generating series of walks in the quarter plane](#)”. arXiv: [1702.04696](#) (cit. on p. [3183](#)).
- Michael Drmota (2009). *Random trees*. An interplay between combinatorics and probability. SpringerWienNewYork, Vienna, pp. xviii+458. MR: [2484382](#) (cit. on p. [3191](#)).
- Philippe Duchon, Philippe Flajolet, Guy Louchard, and Gilles Schaeffer (2004). “[Boltzmann samplers for the random generation of combinatorial structures](#)”. *Combin. Probab. Comput.* 13.4-5, pp. 577–625. MR: [2095975](#) (cit. on p. [3189](#)).
- Murray Elder, Andrew Rechnitzer, Esaías J. Janse van Rensburg, and Thomas Wong (2014). “[The cogrowth series for  \$BS\(N, N\)\$  is D-finite](#)”. *Internat. J. Algebra Comput.* 24.2, pp. 171–187. MR: [3192369](#) (cit. on p. [3183](#)).
- Peter van Emde Boas (1997). “The convenience of tilings”. In: *Complexity, logic, and recursion theory*. Vol. 187. Lecture Notes in Pure and Appl. Math. Dekker, New York, pp. 331–363. MR: [1455142](#) (cit. on p. [3178](#)).
- P. Erdős and A. Rényi (1963). “[Asymmetric graphs](#)”. *Acta Math. Acad. Sci. Hungar* 14, pp. 295–315. MR: [0156334](#) (cit. on p. [3175](#)).
- Philippe Flajolet, Stefan Gerhold, and Bruno Salvy (2004/06). “[On the non-holonomic character of logarithms, powers, and the  \$n\$ th prime function](#)”. *Electron. J. Combin.* 11.2, Article 2, 16. MR: [2195433](#) (cit. on pp. [3180](#), [3181](#)).
- Philippe Flajolet and Robert Sedgewick (2009). *Analytic combinatorics*. Cambridge University Press, Cambridge, pp. xiv+810. MR: [2483235](#) (cit. on pp. [3174](#), [3181](#), [3185](#)).
- Jacob Fox (2013). “[Stanley–Wilf limits are typically exponential](#)”. arXiv: [1310.8378](#) (cit. on p. [3187](#)).
- Juliana Freire, Caroline J. Klivans, Pedro H. Milet, and Nicolau C. Saldanha (Feb. 2017). “[On the connectivity of spaces of three-dimensional tilings](#)”. arXiv: [1702.00798](#) (cit. on p. [3177](#)).

- M. R. Garey, D. S. Johnson, and R. Endre Tarjan (1976). “[The planar Hamiltonian circuit problem is NP-complete](#)”. *SIAM J. Comput.* 5.4, pp. 704–714. MR: [0444516](#) (cit. on p. [3176](#)).
- S. Garrabrant and I. Pak (n.d.). In preparation (cit. on pp. [3184](#), [3187](#)).
- Scott Garrabrant and Igor Pak (May 2015). “[Pattern avoidance is not P-recursive](#)”. arXiv: [1505.06508](#) (cit. on p. [3187](#)).
- (2017). “[Words in linear groups, random walks, automata and P-recursiveness](#)”. *J. Comb. Algebra* 1.2, pp. 127–144. MR: [3634780](#) (cit. on pp. [3174](#), [3181](#), [3183](#), [3184](#)).
- A. M. Garsia and S. C. Milne (1981). “[A Rogers-Ramanujan bijection](#)”. *J. Combin. Theory Ser. A* 31.3, pp. 289–339. MR: [635372](#) (cit. on p. [3189](#)).
- Ira M. Gessel (1990). “[Symmetric functions and P-recursiveness](#)”. *J. Combin. Theory Ser. A* 53.2, pp. 257–285. MR: [1041448](#) (cit. on pp. [3187](#), [3188](#)).
- Oded Goldreich (2008). *Computational complexity*. A conceptual perspective. Cambridge University Press, Cambridge, pp. xxiv+606. MR: [2400985](#) (cit. on p. [3175](#)).
- I. P. Goulden and D. M. Jackson (1986). “[Labelled graphs with small vertex degrees and P-recursiveness](#)”. *SIAM J. Algebraic Discrete Methods* 7.1, pp. 60–66. MR: [819706](#).
- Suresh Govindarajan (2013). “[Notes on higher-dimensional partitions](#)”. *J. Combin. Theory Ser. A* 120.3, pp. 600–622. MR: [3007138](#) (cit. on p. [3190](#)).
- Rostislav Grigorchuk and Igor Pak (2008). “Groups of intermediate growth: an introduction”. *Enseign. Math. (2)* 54.3–4, pp. 251–272. MR: [2478087](#) (cit. on p. [3185](#)).
- Anthony J. Guttmann, ed. (2009). *Polygons, polyominoes and polycubes*. Vol. 775. Lecture Notes in Physics. [Editor name on title page: Anthony J. Guttmann]. Springer, Dordrecht, pp. xx+490. MR: [2797390](#) (cit. on p. [3183](#)).
- Mark Haiman (1993). “[Noncommutative rational power series and algebraic generating functions](#)”. *European J. Combin.* 14.4, pp. 335–339. MR: [1226580](#) (cit. on p. [3183](#)).
- Frank Harary and Edgar M. Palmer (1973). *Graphical enumeration*. Academic Press, New York-London, pp. xiv+271. MR: [0357214](#) (cit. on p. [3175](#)).
- G. H. Hardy and E. M. Wright (2008). *An introduction to the theory of numbers*. Sixth. Revised by D. R. Heath-Brown and J. H. Silverman, With a foreword by Andrew Wiles. Oxford University Press, Oxford, pp. xxii+621. MR: [2445243](#) (cit. on p. [3180](#)).
- Christopher Hoffman, Douglas Rizzolo, and Erik Slivken (2017). “[Pattern-avoiding permutations and Brownian excursion Part I: shapes and fluctuations](#)”. *Random Structures Algorithms* 50.3, pp. 394–419. MR: [3632417](#) (cit. on p. [3193](#)).
- Carlos Hoppen, Yoshiharu Kohayakawa, Carlos Gustavo Moreira, Balázs Ráth, and Rudini Menezes Sampaio (2013). “[Limits of permutation sequences](#)”. *J. Combin. Theory Ser. B* 103.1, pp. 93–113. MR: [2995721](#) (cit. on p. [3191](#)).
- Stephen P. Humphries (1997). “[Cogrowth of groups and the Dedekind-Frobenius group determinant](#)”. *Math. Proc. Cambridge Philos. Soc.* 121.2, pp. 193–217. MR: [1426519](#) (cit. on p. [3183](#)).



- Fredrik Johansson (2012). “Efficient implementation of the Hardy–Ramanujan–Rademacher formula”. *LMS J. Comput. Math.* 15, pp. 341–359. MR: [2988821](#) (cit. on p. [3186](#)).
- Volker Kaibel and Günter M. Ziegler (2003). “Counting lattice triangulations”. In: *Surveys in combinatorics, 2003 (Bangor)*. Vol. 307. London Math. Soc. Lecture Note Ser. Cambridge Univ. Press, Cambridge, pp. 277–307. MR: [2011739](#) (cit. on p. [3177](#)).
- Mihyun Kang and Philipp Sprüssel (2018). “Symmetries of Unlabelled Planar Triangulations”. *The Electronic Journal of Combinatorics* 25.1, pp. 1–34 (cit. on p. [3176](#)).
- Martin Kassabov and Igor Pak (n.d.). “Continuum many spectral radii of finitely generated groups”. In preparation (cit. on p. [3184](#)).
- (2013). “Groups of oscillating intermediate growth”. *Ann. of Math. (2)* 177.3, pp. 1113–1145. MR: [3034295](#) (cit. on p. [3184](#)).
- Richard Kenyon (2004). “An introduction to the dimer model”. In: *School and Conference on Probability Theory*. ICTP Lect. Notes, XVII. Abdus Salam Int. Cent. Theoret. Phys., Trieste, pp. 267–304. MR: [2198850](#) (cit. on p. [3177](#)).
- Harry Kesten (1959). “Symmetric random walks on groups”. *Trans. Amer. Math. Soc.* 92, pp. 336–354. MR: [0109367](#) (cit. on p. [3183](#)).
- Sergey Kitaev (2011). *Patterns in permutations and words*. Monographs in Theoretical Computer Science. An EATCS Series. With a foreword by Jeffrey B. Remmel. Springer, Heidelberg, pp. xxii+494. MR: [3012380](#) (cit. on pp. [3187](#), [3193](#)).
- Martin Klazar (2003). “Bell numbers, their relatives, and algebraic differential equations”. *J. Combin. Theory Ser. A* 102.1, pp. 63–87. MR: [1970977](#) (cit. on p. [3180](#)).
- (2010). “Some general results in combinatorial enumeration”. In: *Permutation patterns*. Vol. 376. London Math. Soc. Lecture Note Ser. Cambridge Univ. Press, Cambridge, pp. 3–40. MR: [2732822](#) (cit. on p. [3187](#)).
- Matjaž Konvalinka and Igor Pak (2009). “Geometry and complexity of O’Hara’s algorithm”. *Adv. in Appl. Math.* 42.2, pp. 157–175. MR: [2493975](#) (cit. on p. [3190](#)).
- Dmitri Kouksov (1998). “On rationality of the cogrowth series”. *Proc. Amer. Math. Soc.* 126.10, pp. 2845–2847. MR: [1487319](#) (cit. on p. [3183](#)).
- C. Krattenthaler (1999). “Another involution principle-free bijective proof of Stanley’s hook-content formula”. *J. Combin. Theory Ser. A* 88.1, pp. 66–92. MR: [1713492](#) (cit. on p. [3190](#)).
- (2016). “Plane partitions in the work of Richard Stanley and his school”. In: *The mathematical legacy of Richard P. Stanley*. Amer. Math. Soc., Providence, RI, pp. 231–261. MR: [3618037](#) (cit. on p. [3192](#)).
- Dmitri Kuksov (1999). “Cogrowth series of free products of finite and free groups”. *Glasg. Math. J.* 41.1, pp. 19–31. MR: [1689726](#) (cit. on pp. [3183](#), [3185](#)).
- J. C. Lagarias, V. S. Miller, and A. M. Odlyzko (1985). “Computing  $\pi(x)$ : the Meissel-Lehmer method”. *Math. Comp.* 44.170, pp. 537–560. MR: [777285](#) (cit. on p. [3175](#)).



- Thomas Lam (2008). “Tiling with Commutative Rings”. *arvard College Math. Review* 2.1, p. 7 (cit. on p. 3178).
- David A. Levin, Yuval Peres, and Elizabeth L. Wilmer (2017). *Markov chains and mixing times*. Second edition of [MR2466937], With a chapter on “Coupling from the past” by James G. Propp and David B. Wilson. American Mathematical Society, Providence, RI, pp. xvi+447. MR: 3726904 (cit. on p. 3189).
- L. A. Levin (1973). “Universal enumeration problems”. *Problemy Peredači Informacii* 9.3, pp. 115–116. MR: 0340042 (cit. on p. 3178).
- Leonard Lipshitz and Lee A. Rubel (1986). “A gap theorem for power series solutions of algebraic differential equations”. *Amer. J. Math.* 108.5, pp. 1193–1213. MR: 859776 (cit. on p. 3180).
- Richard J. Lipton and Yecheskel Zalcstein (1977). “Word problems solvable in logspace”. *J. Assoc. Comput. Mach.* 24.3, pp. 522–526. MR: 0445901 (cit. on p. 3185).
- L. Lovász and M. D. Plummer (1986). *Matching theory*. Vol. 121. North-Holland Mathematics Studies. Annals of Discrete Mathematics, 29. North-Holland Publishing Co., Amsterdam; North-Holland Publishing Co., Amsterdam, pp. xxvii+544. MR: 859549 (cit. on p. 3177).
- László Lovász (2012). *Large networks and graph limits*. Vol. 60. American Mathematical Society Colloquium Publications. American Mathematical Society, Providence, RI, pp. xiv+475. MR: 3012035 (cit. on p. 3191).
- Eugene M. Luks (1982). “Isomorphism of graphs of bounded valence can be tested in polynomial time”. *J. Comput. System Sci.* 25.1, pp. 42–65. MR: 685360 (cit. on p. 3176).
- Neal Madras and Lerna Pehlivan (2016). “Structure of random 312-avoiding permutations”. *Random Structures Algorithms* 49.3, pp. 599–631. MR: 3545829 (cit. on p. 3193).
- Adam Marcus and Gábor Tardos (2004). “Excluded permutation matrices and the Stanley–Wilf conjecture”. *J. Combin. Theory Ser. A* 107.1, pp. 153–160. MR: 2063960 (cit. on p. 3187).
- Charles F. Miller III (1992). “Decision problems for groups—survey and reflections”. In: *Algorithms and classification in combinatorial group theory (Berkeley, CA, 1989)*. Vol. 23. Math. Sci. Res. Inst. Publ. Springer, New York, pp. 1–59. MR: 1230627 (cit. on p. 3185).
- W. H. Mills, David P. Robbins, and Howard Rumsey Jr. (1986). “Self-complementary totally symmetric plane partitions”. *J. Combin. Theory Ser. A* 42.2, pp. 277–292. MR: 847558 (cit. on p. 3192).
- Sam Miner and Igor Pak (2014). “The shape of random pattern-avoiding permutations”. *Adv. in Appl. Math.* 55, pp. 86–130. MR: 3176717 (cit. on p. 3193).
- Cristopher Moore and Stephan Mertens (2011). *The nature of computation*. Oxford University Press, Oxford, pp. xviii+985. MR: 2849868 (cit. on p. 3175, 3178).

- Christoph Neumann and Robin Sulzgruber (2015). “A complexity theorem for the Novelli–Pak–Stoyanovskii algorithm”. *J. Combin. Theory Ser. A* 135, pp. 85–104. MR: [3366471](#) (cit. on p. 3190).
- John Noonan and Doron Zeilberger (1996). “The enumeration of permutations with a prescribed number of “forbidden” patterns”. *Adv. in Appl. Math.* 17.4, pp. 381–407. MR: [1422065](#) (cit. on p. 3187).
- Marc Noy, Clément Requilé, and Juanjo Rué (Feb. 2018). “Further results on random cubic planar graphs”. arXiv: [1802.06679](#) (cit. on p. 3176).
- Andrei Okounkov (2016). “Limit shapes, real and imagined”. *Bull. Amer. Math. Soc. (N.S.)* 53.2, pp. 187–216. MR: [3474306](#) (cit. on p. 3191).
- I. Pak and D. Yeliussizov (n.d.). In preparation (cit. on p. 3186).
- Igor Pak (2003). “Tile invariants: new horizons”. *Theoret. Comput. Sci.* 303.2-3. Tilings of the plane, pp. 303–331. MR: [1990769](#) (cit. on p. 3177).
- (2004a). “Partition identities and geometric bijections”. *Proc. Amer. Math. Soc.* 132.12, pp. 3457–3462. MR: [2084064](#) (cit. on p. 3189).
  - (2004b). “The nature of partition bijections II”. *Asymptotic stability*. Preprint, p. 32 (cit. on p. 3191).
  - (2006). “Partition bijections, a survey”. *Ramanujan J.* 12.1, pp. 5–75. MR: [2267263](#) (cit. on pp. 3189–3191).
- Igor Pak and Ernesto Vallejo (2010). “Reductions of Young tableau bijections”. *SIAM J. Discrete Math.* 24.1, pp. 113–145. MR: [2600656](#) (cit. on p. 3190).
- Igor Pak and Jed Yang (2013). “The complexity of generalized domino tilings”. *Electron. J. Combin.* 20.4, Paper 12, 23. MR: [3139397](#) (cit. on p. 3177).
- Greta Panova (2015). “Lozenge tilings with free boundaries”. *Lett. Math. Phys.* 105.11, pp. 1551–1586. MR: [3406712](#) (cit. on p. 3192).
- Robin Pemantle and Mark C. Wilson (2013). *Analytic combinatorics in several variables*. Vol. 140. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, pp. xiv+380. MR: [3088495](#) (cit. on p. 3181).
- Marius van der Put and Michael F. Singer (1997). *Galois theory of difference equations*. Vol. 1666. Lecture Notes in Mathematics. Springer-Verlag, Berlin, pp. viii+180. MR: [1480919](#) (cit. on p. 3180).
- Joseph Fels Ritt (1950). *Differential Algebra*. American Mathematical Society Colloquium Publications, Vol. XXXIII. American Mathematical Society, New York, N. Y., pp. viii+184. MR: [0035763](#) (cit. on p. 3180).
- Neville Robbins (1996). “Fibonacci partitions”. *Fibonacci Quart.* 34.4, pp. 306–313. MR: [1394758](#) (cit. on p. 3186).
- Dan Romik (2015). *The surprising mathematics of longest increasing subsequences*. Institute of Mathematical Statistics Textbooks. Cambridge University Press, New York, pp. xi+353. MR: [3468738](#) (cit. on p. 3191).

- Gilles Schaeffer (2015). “Planar maps”. In: *Handbook of enumerative combinatorics*. Discrete Math. Appl. (Boca Raton). CRC Press, Boca Raton, FL, pp. 335–395. MR: [3409346](#) (cit. on p. [3189](#)).
- Carsten Schneider and Robin Sulzgruber (2017). “Asymptotic and exact results on the complexity of the Novelli-Pak-Stoyanovskii algorithm”. *Electron. J. Combin.* 24.2, Paper 2.28, 33. MR: [3665561](#) (cit. on p. [3190](#)).
- Neil JA Sloane (n.d.). *The on-line encyclopedia of integer sequences* (cit. on pp. [3172](#), [3175–3177](#), [3179](#), [3182](#), [3186](#), [3187](#), [3189](#), [3192](#)).
- Richard P. Stanley (1999). *Enumerative combinatorics. Vol. 2*. Vol. 62. Cambridge Studies in Advanced Mathematics. With a foreword by Gian-Carlo Rota and appendix 1 by Sergey Fomin. Cambridge University Press, Cambridge, pp. xii+581. MR: [1676282](#) (cit. on pp. [3179](#), [3182](#)).
- (2015). *Catalan numbers*. Cambridge University Press, New York, pp. viii+215. MR: [3467982](#) (cit. on pp. [3177](#), [3192](#)).
- RP Stanley (2014). “D-finiteness of certain series associated with group algebras”. *Oberwolfach Rep* 11. in Enumerative Combinatorics, (M. Bousquet-Mélou et al., eds.), p. 708 (cit. on p. [3184](#)).
- Einar Steingrímsson (2013). “Some open problems on permutation patterns.” *Surveys in combinatorics* 409, pp. 239–263 (cit. on p. [3188](#)).
- Terence Tao, Ernest Croot III, and Harald Helfgott (2012). “Deterministic methods to find primes”. *Math. Comp.* 81.278, pp. 1233–1246. MR: [2869058](#) (cit. on p. [3175](#)).
- Panayiotis G. Tsangaris (2007). “Formulae for the  $k$ th prime number”. *Bull. Greek Math. Soc.* 53, pp. 147–149. MR: [2466504](#) (cit. on p. [3174](#)).
- W. T. Tutte (1998). *Graph theory as I have known it*. Vol. 11. Oxford Lecture Series in Mathematics and its Applications. With a foreword by U. S. R. Murty. The Clarendon Press, Oxford University Press, New York, pp. vi+156. MR: [1635397](#) (cit. on p. [3176](#)).
- L. G. Valiant (1979). “Completeness classes in algebra”. In: *Conference Record of the Eleventh Annual ACM Symposium on Theory of Computing (Atlanta, Ga., 1979)*. ACM, New York, pp. 249–261. MR: [564634](#) (cit. on p. [3177](#)).
- Vincent Vatter (2015). “Permutation classes”. In: *Handbook of enumerative combinatorics*. Discrete Math. Appl. (Boca Raton). CRC Press, Boca Raton, FL, pp. 753–833. MR: [3409353](#) (cit. on p. [3187](#)).
- Julian West (1995). “Generating trees and the Catalan and Schröder numbers”. *Discrete Math.* 146.1-3, pp. 247–262. MR: [1360119](#) (cit. on p. [3192](#)).
- A. Wigderson (2018). Manuscript.
- Herbert S. Wilf (1982). “What is an answer?” *Amer. Math. Monthly* 89.5, pp. 289–292. MR: [653502](#) (cit. on pp. [3174](#), [3175](#)).
- David Bruce Wilson (1996). “Generating random spanning trees more quickly than the cover time”. In: *Proceedings of the Twenty-eighth Annual ACM Symposium on the*

- Theory of Computing (Philadelphia, PA, 1996)*. ACM, New York, pp. 296–303. MR: [1427525](#) (cit. on pp. [3189](#), [3191](#)).
- Robin Wilson and John J. Watkins, eds. (2013). *Combinatorics: ancient and modern*. Oxford University Press, Oxford, pp. x+381. MR: [3204727](#) (cit. on p. [3174](#)).
- Wolfgang Woess (2000). *Random walks on infinite graphs and groups*. Vol. 138. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, pp. xii+334. MR: [1743100](#) (cit. on p. [3184](#)).
- Jed Yang (2014). “Rectangular tileability and complementary tileability are undecidable”. *European J. Combin.* 41, pp. 20–34. MR: [3219249](#) (cit. on p. [3178](#)).
- Don Zagier (2008). “Elliptic modular forms and their applications”. In: *The 1-2-3 of modular forms*. Universitext. Springer, Berlin, pp. 1–103. MR: [2409678](#) (cit. on p. [3185](#)).

Received 2018-03-05.

DEPARTMENT OF MATHEMATICS  
UCLA  
LOS ANGELES CA 90095

IGOR PAK  
[pak@math.ucla.edu](mailto:pak@math.ucla.edu)



# POSITIVE GRASSMANNIAN AND POLYHEDRAL SUBDIVISIONS

ALEXANDER POSTNIKOV

## Abstract

The nonnegative Grassmannian is a cell complex with rich geometric, algebraic, and combinatorial structures. Its study involves interesting combinatorial objects, such as positroids and plabic graphs. Remarkably, the same combinatorial structures appeared in many other areas of mathematics and physics, e.g., in the study of cluster algebras, scattering amplitudes, and solitons. We discuss new ways to think about these structures. In particular, we identify plabic graphs and more general Grassmannian graphs with polyhedral subdivisions induced by 2-dimensional projections of hypersimplices. This implies a close relationship between the positive Grassmannian and the theory of fiber polytopes and the generalized Baues problem. This suggests natural extensions of objects related to the positive Grassmannian.

## 1 Introduction

The geometry of the Grassmannian  $Gr(k, n)$  is related to combinatorics of the hypersimplex  $\Delta_{kn}$ . Gelfand, Goresky, MacPherson, and Serganova [1987] studied the hypersimplex as the moment polytope for the torus action on the complex Grassmannian. In this paper we highlight new links between geometry of the *positive Grassmannian* and combinatorics of the hypersimplex  $\Delta_{kn}$ .

Gelfand, Goresky, MacPherson, and Serganova [ibid.] studied the *matroid stratification* of the Grassmannian  $Gr(k, n, \mathbb{C})$ , whose strata are the realization spaces of *matroids*. They correspond to *matroid polytopes* living inside  $\Delta_{kn}$ . In general, matroid strata are not cells. In fact, according to Mnëv’s universality theorem Mnëv [1988], the matroid

---

MSC2010: primary 05E00; secondary 52B, 52C, 13F60, 81T.

Keywords: Total positivity, positive Grassmannian, hypersimplex, matroids, positroids, cyclic shifts, Grassmannian graphs, plabic graphs, polyhedral subdivisions, triangulations, zonotopal tilings, associahedron, fiber polytopes, Baues poset, generalized Baues problem, flips, cluster algebras, weakly separated collections, scattering amplitudes, amplituhedron, membranes.

strata can be as complicated as any algebraic variety. Thus the matroid stratification of the Grassmannian can have arbitrarily bad behavior.

There is, however, a semialgebraic subset of the real Grassmannian  $Gr(k, n, \mathbb{R})$ , called the *nonnegative Grassmannian*  $Gr^{\geq 0}(k, n)$ , where the matroid stratification exhibits a well behaved combinatorial and geometric structure. Its structure, which is quite rich and nontrivial, can nevertheless be described in explicit terms. In some way, the nonnegative Grassmannian is similar to a polytope.

The notion of a *totally positive matrix*, that is a matrix with all positive minors, originated in pioneering works of [Gantmacher and Krein \[1935\]](#), and [Schoenberg \[1930\]](#). Since then such matrices appeared in many areas of pure and applied mathematics. [Lusztig \[1994, 1998a\]](#) and [Lusztig \[1998b\]](#) generalized the theory of total positivity in the general context of Lie theory. He defined the positive part for a reductive Lie group  $G$  and a generalized flag variety  $G/P$ . [K. C. Rietsch \[1998\]](#) and [K. Rietsch \[1999\]](#) studied its cellular decomposition. Lusztig’s theory of total positivity has close links with his theory of canonical bases [Lusztig \[1990, 1992\]](#) and [Lusztig \[1993\]](#) and Fomin-Zelevinsky’s cluster algebras [Fomin and A. Zelevinsky \[2002a,b, 2003\]](#), [Berenstein, Fomin, and A. Zelevinsky \[2005\]](#), and [Fomin and A. Zelevinsky \[2007\]](#).

[Postnikov \[2006\]](#) initiated a combinatorial approach to the study of the positive Grassmannian. The positive (resp., nonnegative) Grassmannian  $Gr^{>0}(k, n)$  ( $Gr^{\geq 0}(k, n)$ ) was described as the subset of the Grassmannian  $Gr(k, n, \mathbb{R})$  where all Plücker coordinates are positive (resp., nonnegative). This “elementary” definition agrees with Lusztig’s general notion [Lusztig \[1998a\]](#) of the positive part of  $G/P$  in the case when  $G/P = Gr(k, n)$ .

The *positroid cells*, defined as the parts of matroid strata inside the nonnegative Grassmannian, turned out to be indeed cells. (The term “positroid” is an abbreviation for “positive matroid.”) The positroid cells form a CW-complex. Conjecturally, it is a *regular* CW-complex, and the closure of each positroid cell is homeomorphic to a closed ball. This positroid stratification of  $Gr^{\geq 0}(k, n)$  is the common refinement of  $n$  *cyclically shifted* Schubert decompositions [Postnikov \[2006\]](#). Compare this with the result [Gelfand, Goresky, MacPherson, and Serganova \[1987\]](#) that the matroid stratification of  $Gr(k, n)$  is the common refinement of  $n!$  *permuted* Schubert decompositions. The *cyclic shift* plays a crucial role in the study of the positive Grassmannian. Many objects associated with the positive Grassmannian exhibit *cyclic symmetry*.

Positroid cells were identified in [Postnikov \[2006\]](#) with many combinatorial objects, such as decorated permutation, Grassmann necklaces, etc. Moreover, an explicit birational subtraction-free parametrization of each cell was described in terms of *plabic graphs*, that is, *planar bicolored* graphs, which are certain graphs embedded in a disk with vertices colored in two colors.

Remarkably, the combinatorial structures that appeared in the study of the positive Grassmannian also surfaced and played an important role in many different areas of mathematics and physics. J. Scott [2005] and J. S. Scott [2006] and Oh, Postnikov, and D. E. Speyer [2015] linked these objects with *cluster algebra* structure on the Grassmannian and with Leclerc-Zelevinsky’s quasi-commuting families of *quantum minors* and *weakly separated collections*. Corteel and L. K. Williams [2007] applied Le-diagrams (which correspond to positroids) to the study of the *partially asymmetric exclusion process* (PASEP). Knutson, Lam, and D. E. Speyer [2013] proved that the cohomology classes of the *positroid varieties* (the complexifications of the positroid cells) are given by the *affine Stanley symmetric functions*, which are dual to Lapointe-Lascoux-Morse *k-Schur functions*. They also linked positroids with *theory of juggling*. Plabic graphs appeared in works of Chakravarty and Kodama [2009], Kodama and L. K. Williams [2011], and Kodama and L. Williams [2014] as *soliton solutions* of the Kadomtsev-Petviashvili (KP) equation, which describes nonlinear waves. Last but not least, plabic graphs appeared under the name of *on-shell diagrams* in the work by Arkani-Hamed, Bourjaily, Cachazo, Goncharov, Postnikov, and Trnka [2016] on *scattering amplitudes* in  $\mathfrak{N} = 4$  supersymmetric Yang-Mills (SYM) theory. They play a role somewhat similar to Feynman diagrams, however, unlike Feynman diagrams, they represent on-shell processes and do not require introduction of virtual particles.

In this paper, we review some of the main constructions and results from Postnikov [2006], Postnikov, D. Speyer, and L. Williams [2009], and Oh, Postnikov, and D. E. Speyer [2015] related to the positive Grassmannian. We extend these constructions in the language of *Grassmannian graphs*. The parametrization of a positroid cell in  $Gr^{\geq 0}(k, n)$  given by a Grassmannian graph can be thought of as a way to “glue” the positroid cell out of “little positive Grassmannians” associated with vertices of the graph. The idea to think about parametrizations of cells as gluings of Grassmannians came originally from physics Arkani-Hamed, Bourjaily, Cachazo, Goncharov, Postnikov, and Trnka [2016], where vertices of on-shell diagrams (i.e., plabic graphs) were viewed as little Grassmannians  $Gr(1, 3)$  and  $Gr(2, 3)$ .

We link this construction of parametrizations of  $Gr^{>0}(k, n)$  given by Grassmannian graphs with the study of polyhedral subdivisions induced by 2-dimensional cyclic projections  $\pi : \Delta_{kn} \rightarrow Q$  of the hypersimplex. Reduced Grassmannian graphs parametrizing the positive Grassmannian  $Gr^{>0}(k, n)$  turn out to be in bijection with  $\pi$ -induced polyhedral subdivisions. Thus gluing of Grassmannians from smaller Grassmannians is equivalent to subdividing polytopes into smaller polytopes. The study of  $\pi$ -induced subdivisions for projections of polytopes is the subject of Billera-Sturmfels’ theory Billera and Sturmfels [1992] of *fiber polytopes* and the *generalized Baues problem* (GBP) posed by Billera, Kapranov, and Sturmfels [1994]. We also mention the result of Galashin [2016] where plabic graphs are identified with sections of *zonotopal tilings*, and the construction from



the work [T. Lam \[2018\]](#) with Lam on *polypositroids* where plabic graphs are viewed as *membranes*, which are certain 2-dimensional surfaces in higher dimensional spaces.

The correspondence between parametrizations of the positive Grassmannian and polyhedral subdivisions leads to natural generalizations and conjectures. We discuss a possible extension of constructions of this paper to “higher positive Grassmannians” and amplituhedra of [Arkani-Hamed and Trnka \[2014\]](#).

I thank Federico Ardila, Nima Arkani-Hamed, Arkady Berenstein, David Bernstein, Lou Billera, Jacob Bourjaily, Freddy Cachazo, Miriam Farber, Sergey Fomin, Pavel Galashin, Israel Moiseevich Gelfand, Oleg Gleizer, Alexander Goncharov, Darij Grinberg, Alberto Grünbaum, Xuhua He, Sam Hopkins, David Ingberman, Tamás Kálmán, Mikhail Kapranov, Askold Khovanskii, Anatol Kirillov, Allen Knutson, Gleb Koshevoy, Thomas Lam, Joel Lewis, Gaku Liu, Ricky Liu, George Lusztig, Thomas McConville, Karola Mészáros, Alejandro Morales, Gleb Nenashev, Suho Oh, Jim Propp, Vic Reiner, Vladimir Retakh, Konni Rietsch, Tom Roby, Yuval Roichman, Paco Santos, Jeanne Scott, Boris Shapiro, Michael Shapiro, David Speyer, Richard Stanley, Bernd Sturmfels, Dylan Thurston, Jaroslav Trnka, Wuttisak Trongsiriwat, Vladimir Voevodsky, Lauren Williams, Hwanchul Yoo, Andrei Zelevinsky, and Günter Ziegler for insightful conversations. These people made a tremendous contribution to the study of the positive Grassmannian and related combinatorial, algebraic, geometric, topological, and physical structures. Many themes we discuss here are from past and future projects with various subsets of these people.

## 2 Grassmannian and matroids

Fix integers  $0 \leq k \leq n$ . Let  $[n] := \{1, \dots, n\}$  and  $\binom{[n]}{k}$  be the set of  $k$ -element subsets of  $[n]$ .

The *Grassmannian*  $Gr(k, n) = Gr(k, n, \mathbb{F})$  over a field  $\mathbb{F}$  is the variety of  $k$ -dimensional linear subspaces in  $\mathbb{F}^n$ . More concretely,  $Gr(k, n)$  is the space of  $k \times n$ -matrices of rank  $k$  modulo the left action of  $GL(k) = GL(k, \mathbb{F})$ . Let  $[A] = GL(k)A$  be the element of  $Gr(k, n)$  represented by matrix  $A$ .

Maximal minors  $\Delta_I(A)$  of such matrices  $A$ , where  $I \in \binom{[n]}{k}$ , form projective coordinates on  $Gr(k, n)$ , called the *Plücker coordinates*. For  $[A] \in Gr(k, n)$ , let

$$\mathfrak{M}(A) := \{I \in \binom{[n]}{k} \mid \Delta_I(A) \neq 0\}.$$

The sets of the form  $\mathfrak{M}(A)$  are a special kind of *matroids*, called  $\mathbb{F}$ -*realizable matroids*. *Matroid strata* are the realization spaces of realizable matroids  $\mathfrak{M} \subset \binom{[n]}{k}$ :

$$S_{\mathfrak{M}} := \{[A] \in Gr(k, n) \mid \mathfrak{M}(A) = \mathfrak{M}\}.$$

The *matroid stratification* is the disjoint decomposition

$$Gr(k, n) = \bigsqcup_{\mathfrak{M} \text{ realizable matroid}} S_{\mathfrak{M}}.$$

The *Gale order* “ $\preceq$ ” (or the coordinatewise order) is the partial order on  $\binom{[n]}{k}$  given by  $\{i_1 < \dots < i_k\} \preceq \{j_1 < \dots < j_k\}$ , if  $i_r \leq j_r$  for  $r \in [k]$ . Each matroid  $\mathfrak{M}$  has a unique minimal element  $I_{\min}(\mathfrak{M})$  with respect to the Gale order.

For  $I \in \binom{[n]}{k}$ , the *Schubert cell*  $\Omega_I \subset Gr(k, n)$  is given by

$$\Omega_I := \{[A] \in Gr(k, n) \mid I = I_{\min}(\mathfrak{M}(A))\} = \bigsqcup_{\mathfrak{M}: I = I_{\min}(\mathfrak{M})} S_{\mathfrak{M}}.$$

They form the *Schubert decomposition*  $Gr(k, n) = \bigsqcup \Omega_I$ . Clearly, for a realizable matroid  $\mathfrak{M}$ , we have  $S_{\mathfrak{M}} \subset \Omega_I$  if and only if  $I = I_{\min}(\mathfrak{M})$ .

The *symmetric group*  $S_n$  acts on  $Gr(k, n)$  by permutations

$$w([v_1, \dots, v_n]) = [v_{w(1)}, \dots, v_{w(n)}]$$

of columns of  $[A] = [v_1, \dots, v_n] \in Gr(k, n)$ .

It is clear that, see [Gelfand, Goresky, MacPherson, and Serganova \[1987\]](#), the matroid stratification of  $Gr(k, n)$  is the common refinement of the  $n!$  permuted Schubert decompositions. In other words, each matroid stratum  $S_{\mathfrak{M}}$  is an intersection of permuted Schubert cells:

$$S_{\mathfrak{M}} = \bigcap_{w \in S_n} w(\Omega_{I_w}).$$

Indeed, if we know the minimal elements of a set  $\mathfrak{M} \subset \binom{[n]}{k}$  with respect to all  $n!$  orderings of  $[n]$ , we know the set  $\mathfrak{M}$  itself.

### 3 Positive Grassmannian and positroids

Fix the field  $\mathbb{F} = \mathbb{R}$ . Let  $Gr(k, n) = Gr(k, n, \mathbb{R})$  be the real Grassmannian.

**Definition 3.1.** [Postnikov \[2006, Definition 3.1\]](#) The *positive Grassmannian*  $Gr^{>0}(k, n)$  (resp., *nonnegative Grassmannian*  $Gr^{\geq 0}(k, n)$ ) is the semialgebraic set of elements  $[A] \in Gr(k, n)$  represented by  $k \times n$  matrices  $A$  with all positive maximal minors  $\Delta_I(A) > 0$  (resp., all nonnegative maximal minors  $\Delta_I(A) \geq 0$ ).

This definition agrees with Lusztig’s general definition [Lusztig \[1998a\]](#) of the positive part of a generalized flag variety  $G/P$  in the case when  $G/P = Gr(k, n)$ .

**Definition 3.2.** Postnikov [2006, Definition 3.2] A *positroid cell*  $\Pi_{\mathfrak{m}} \subset Gr^{\geq 0}(k, n)$  is a nonempty intersection of a matroid stratum with the nonnegative Grassmannian:

$$\Pi_{\mathfrak{m}} := S_{\mathfrak{m}} \cap Gr^{\geq 0}(k, n).$$

A *positroid* of rank  $k$  is a collection  $\mathfrak{m} \subset \binom{[n]}{k}$  such that  $\Pi_{\mathfrak{m}}$  is nonempty. The *positroid stratification* of the nonnegative Grassmannian is the disjoint decomposition of  $Gr^{\geq 0}(k, n)$  into the positroid cells:

$$Gr^{\geq 0}(k, n) = \bigsqcup_{\mathfrak{m} \text{ is a positroid}} \Pi_{\mathfrak{m}}.$$

Clearly, positroids, or positive matroids, are a special kind of matroids. The positive Grassmannian  $Gr^{>0}(k, n)$  itself is the *top positroid cell*  $\Pi_{\binom{[n]}{k}}$  for the uniform matroid  $\mathfrak{m} = \binom{[n]}{k}$ .

The *cyclic shift* is the map  $\tilde{c} : Gr(k, n) \rightarrow Gr(k, n)$  acting on elements  $[A] = [v_1, \dots, v_n] \in Gr(k, n)$  by

$$\tilde{c} : [v_1, \dots, v_n] \mapsto [v_2, v_3, \dots, v_n, (-1)^{k-1}v_1].$$

The shift  $\tilde{c}$  induces the action of the *cyclic group*  $\mathbb{Z}/n\mathbb{Z}$  on the Grassmannian  $Gr(k, n)$ , that preserves its positive part  $Gr^{>0}(k, n)$ . Many of the objects associated with the positive Grassmannian exhibit *cyclic symmetry*. This cyclic symmetry is a crucial ingredient in the study of the positive Grassmannian.

**Theorem 3.3.** Postnikov [ibid., Theorem 3.7] The *positroid stratification* is the common refinement of  $n$  cyclically shifted Schubert decompositions restricted to  $Gr^{\geq 0}(k, n)$ . In other words, each positroid cell  $\Pi_{\mathfrak{m}}$  is given by the intersection of the nonnegative parts of  $n$  cyclically shifted Schubert cells:

$$\Pi_{\mathfrak{m}} = \bigcap_{i=0}^{n-1} \tilde{c}^i(\Omega_{I_i} \cap Gr^{\geq 0}(k, n)).$$

So the positroid cells require intersecting  $n$  cyclically shifted Schubert cells, which is a smaller number than  $n!$  permuted Schubert cells needed for general matroid strata. In fact, the positroid cells  $\Pi_{\mathfrak{m}}$  (unlike matroid strata) are indeed cells.

**Theorem 3.4.** Postnikov [ibid., Theorem 3.5], Postnikov, D. Speyer, and L. Williams [2009, Theorem 5.4] The positroid cells  $\Pi_{\mathfrak{m}}$  are homeomorphic to open balls. The cell decomposition of  $Gr^{\geq 0}(k, n)$  into the positroid cells  $\Pi_{\mathfrak{m}}$  is a CW-complex.

**Conjecture 3.5.** Postnikov [2006, Conjecture 3.6] The *positroid stratification* of the nonnegative Grassmannian  $Gr^{\geq 0}(k, n)$  is a regular CW-complex. In particular, the closure  $\overline{\Pi}_{\mathfrak{m}}$  of each positroid cell in  $Gr^{\geq 0}(k, n)$  is homeomorphic to a closed ball.

This conjecture was motivated by a similar conjecture of Fomin and Zelevinsky on double Bruhat cells [Fomin and A. Zelevinsky \[1999\]](#). Up to homotopy-equivalence this conjecture was proved by [K. Rietsch and L. Williams \[2010\]](#). A major step towards this conjecture was recently achieved by Galashin, Karp, and Lam, who proved it for the top cell.

**Theorem 3.6.** *Galashin, Karp, and Lam [2017, Theorem 1.1] The nonnegative Grassmannian  $Gr^{\geq 0}(k, n)$  is homeomorphic to a closed ball of dimension  $k(n - k)$ .*

By [Theorem 3.3](#), positroids  $\mathfrak{M}$  and positroid cells  $\Pi_{\mathfrak{M}} \subset Gr^{\geq 0}(k, n)$  correspond to certain sequences  $(I_0, I_1, \dots, I_{n-1})$ . Let us describe this bijection explicitly.

**Definition 3.7.** [Postnikov \[2006, Definition 16.1\]](#) A *Grassmann necklace*  $\mathfrak{g} = (J_1, J_2, \dots, J_n)$  of type  $(k, n)$  is a sequence of elements  $J_i \in \binom{[n]}{k}$  such that, for any  $i \in [n]$ , either  $J_{i+1} = (J_i \setminus \{i\}) \cup \{j\}$  or  $J_{i+1} = J_i$ , where the indices  $i$  are taken  $(\bmod n)$ .

The *cyclic permutation*  $c \in S_n$  is given by  $c : i \mapsto i + 1 \pmod{n}$ . The action of the symmetric group  $S_n$  on  $[n]$  induces the  $S_n$ -action on  $\binom{[n]}{k}$  and on subsets of  $\binom{[n]}{k}$ . Recall that  $I_{\min}(\mathfrak{M})$  is the minimal element of a matroid  $\mathfrak{M}$  in the Gale order. For a matroid  $\mathfrak{M}$ , let

$$\mathfrak{g}(\mathfrak{M}) := (J_1, \dots, J_n), \text{ where}$$

$$J_{i+1} = c^i(I_{\min}(c^{-i}(\mathfrak{M}))), \text{ for } i = 0, \dots, n - 1.$$

**Theorem 3.8.** *Postnikov [ibid., Theorem 17.1] The map  $\mathfrak{M} \mapsto \mathfrak{g}(\mathfrak{M})$  is a bijection between positroids  $\mathfrak{M}$  of rank  $k$  on the ground set  $[n]$  and Grassmann necklaces of type  $(k, n)$ .*

The sequence  $(I_0, I_1, \dots, I_{n-1})$  associated with  $\mathfrak{M}$  as in [Theorem 3.3](#) is related to the Grassmann necklace  $(J_1, \dots, J_n)$  of  $\mathfrak{M}$  by  $I_i = c^{-i}(J_{i+1})$ , for  $i = 0, \dots, n - 1$ .

The following result shows how to reconstruct a positroid  $\mathfrak{M}$  from its Grassmann necklace, cf. [Theorem 3.3](#). For  $I \in \binom{[n]}{k}$ , the *Schubert matroid* is  $\mathfrak{M}_I := \{J \in \binom{[n]}{k} \mid I \preceq J\}$ , where “ $\preceq$ ” is the Gale order.

**Theorem 3.9.** *Oh [2011, Theorem 6] For a Grassmann necklace  $\mathfrak{g} = (J_1, \dots, J_n)$ , the associated positroid  $\mathfrak{M}(\mathfrak{g}) = \mathfrak{M}$  is given by*

$$\mathfrak{M} = \bigcap_{i=0}^{n-1} c^i(\mathfrak{M}_{I_i}),$$

where  $I_i = c^{-i}(J_{i+1})$ .

Let us describe positroids in the language of convex geometry. The *hypersimplex*

$$\Delta_{kn} := \text{conv} \left\{ e_I \mid I \in \binom{[n]}{k} \right\}$$

is the convex hull of the  $\binom{n}{k}$  points  $e_I = \sum_{i \in I} e_i$ , for all  $I \in \binom{[n]}{k}$ . Here  $e_1, \dots, e_n$  is the standard basis in  $\mathbb{R}^n$ . For a subset  $\mathfrak{M} \subset \binom{[n]}{k}$ , let  $P_{\mathfrak{M}} := \text{conv}\{e_I \mid I \in \mathfrak{M}\}$  be the convex hull of vertices of  $\Delta_{kn}$  associated with elements of  $\mathfrak{M}$ .

By Gelfand, Goresky, MacPherson, and Serganova [1987],  $\mathfrak{M}$  is a matroid if and only if every edge of the polytope  $P_{\mathfrak{M}}$  has the form  $[e_I, e_J]$ , for  $I, J \in \binom{[n]}{k}$  with  $|I \cap J| = k - 1$ . Here is an analogous description of positroids, which is not hard to derive from Theorem 3.9.

**Theorem 3.10.** *T. Lam [2018] A nonempty subset  $\mathfrak{M} \subset \binom{[n]}{k}$  is a positroid if and only if*

1. *Every edge of  $P_{\mathfrak{M}}$  has the form  $[e_I, e_J]$ , for  $I, J \in \binom{[n]}{k}$  with  $|I \cap J| = k - 1$ .*
2. *Every facet of  $P_{\mathfrak{M}}$  is given by  $x_i + x_{i+1} + \dots + x_j = a_{ij}$  for some cyclic interval  $\{i, i+1, \dots, j\} \subset [n]$  and  $a_{ij} \in \mathbb{Z}$ .*

Many of the results on the positive Grassmannian are based on an explicit birational parametrization Postnikov [2006] of the positroid cells  $\Pi_{\mathfrak{M}}$  in terms of plabic graphs. In the next section we describe a more general class of Grassmannian graphs that includes plabic graphs.

## 4 Grassmannian graphs

**Definition 4.1.** A *Grassmannian graph* is a finite graph  $G = (V, E)$ , with vertex set  $V$  and edge set  $E$ , embedded into a disk (and considered up to homeomorphism) with  $n$  boundary vertices  $b_1, \dots, b_n \in V$  of degree 1 on the boundary of the disk (in the clockwise order), and possibly some *internal vertices*  $v$  in the interior of the disk equipped with integer parameters  $h(v) \in \{0, 1, \dots, \deg(v)\}$ , called *helicities* of vertices. Here  $\deg(v)$  is the degree of vertex  $v$ . We say that an internal vertex  $v$  is of *type*  $(h, d)$  if  $d = \deg(v)$  and  $h = h(v)$ .

The set of internal vertices of  $G$  is denoted by  $V_{\text{int}} = V \setminus \{b_1, \dots, b_n\}$ , and the set of *internal edges*, i.e., the edges which are not adjacent to the boundary vertices, is denoted by  $E_{\text{int}} \subset E$ . The *internal subgraph* is  $G_{\text{int}} = (V_{\text{int}}, E_{\text{int}})$ .

A *perfect orientation* of a Grassmannian graph  $G$  is a choice of directions for all edges  $e \in E$  of the graph  $G$  such that, for each internal vertex  $v \in V_{\text{int}}$  with helicity  $h(v)$ , exactly  $h(v)$  of the edges adjacent to  $v$  are directed towards  $v$  and the remaining  $\deg(v) - h(v)$

of adjacent edges are directed away from  $v$ . A Grassmannian graph is called *perfectly orientable* if it has a perfect orientation.

The *helicity* of a Grassmannian graph  $G$  with  $n$  boundary vertices is the number  $h(G)$  given by

$$h(G) - n/2 = \sum_{v \in V_{\text{int}}} (h(v) - \deg(v)/2).$$

For a perfect orientation  $\mathcal{O}$  of  $G$ , let  $I(\mathcal{O})$  be the set of indices  $i \in [n]$  such that the boundary edge adjacent to  $b_i$  is directed towards the interior of  $G$  in the orientation  $\mathcal{O}$ .

**Lemma 4.2.** *For a perfectly orientable Grassmannian graph  $G$  and any perfect orientation  $\mathcal{O}$  of  $G$ , we have  $|I(\mathcal{O})| = h(G)$ . In particular, in this case,  $h(G) \in \{0, 1, \dots, n\}$ .*

*Remark 4.3.* This lemma expresses the *Helicity Conservation Law*. We leave it as an exercise for the reader.

For a perfectly orientable Grassmannian graph  $G$  of helicity  $h(G) = k$ , let

$$\mathfrak{M}(G) = \{I(\mathcal{O}) \mid \mathcal{O} \text{ is a perfect orientation of } G\} \subset \binom{[n]}{k}.$$

Here is one result that links Grassmannian graphs with positroids.

**Theorem 4.4.** *For a perfectly orientable Grassmannian graph  $G$  with  $h(G) = k$ , the set  $\mathfrak{M}(G)$  is a positroid of rank  $k$ . All positroids have form  $\mathfrak{M}(G)$  for some  $G$ .*

**Definition 4.5.** A *strand*  $\alpha$  in a Grassmannian graph  $G$  is a directed walk along edges of  $G$  that either starts and ends at some boundary vertices, or is a closed walk in the internal subgraph  $G_{\text{int}}$ , satisfying the following *Rules of the Road*: For each internal vertex  $v \in V_{\text{int}}$  with adjacent edges labelled  $a_1, \dots, a_d$  in the clockwise order, where  $d = \deg(v)$ , if  $\alpha$  enters  $v$  through the edge  $a_i$ , it leaves  $v$  through the edge  $a_j$ , where  $j = i + h(v) \pmod{d}$ .

A Grassmannian graph  $G$  is *reduced* if

1. There are no strands which are closed loops in the internal subgraph  $G_{\text{int}}$ .
2. All strands in  $G$  are simple curves without self-intersections. The only exception is that we allow strands  $b_i \rightarrow v \rightarrow b_i$  where  $v \in V_{\text{int}}$  is a *boundary leaf*, that is a vertex of degree 1 connected with  $b_i$  by an edge.
3. Any two strands  $\alpha \neq \beta$  cannot have a *bad double crossing*, that is, a pair of vertices  $u \neq v$  such that both  $\alpha$  and  $\beta$  pass through  $u$  and  $v$  and both are directed from  $u$  to  $v$ . (We allow double crossings where  $\alpha$  goes from  $u$  to  $v$  and  $\beta$  goes from  $v$  to  $u$ .)

4. The graph  $G$  has no vertices of degree 2.

The *decorated strand permutation*  $w = w_G$  of a reduced Grassmannian graph  $G$  is the permutation  $w : [n] \rightarrow [n]$  with fixed points colored in colors 0 or 1 such that

1.  $w(i) = j$  if the strand that starts at the boundary vertex  $b_i$  ends at the boundary vertex  $b_j$ .
2. For a boundary leaf  $v$  connected to  $b_i$ , the decorated permutation  $w$  has fixed point  $w(i) = i$  colored in color  $h(v) \in \{0, 1\}$ .

A *complete* reduced Grassmannian graph  $G$  of type  $(k, n)$ , for  $0 \leq k \leq n$ , is a reduced Grassmannian graph whose decorated strand permutation is given by  $w(i) = i + k \pmod{n}$ . In addition, for  $k = 0$  (resp., for  $k = n$ ), we require that  $G$  only has  $n$  boundary leaves of helicity 0 (resp., of helicity 1) and no other internal vertices.

**Theorem 4.6.** *cf. Postnikov [2006, Corollaries 14.7 and 14.10] (1) For any permutation  $w : [n] \rightarrow [n]$  with fixed points colored in 0 or 1, there exists a reduced Grassmannian graph  $G$  whose decorated strand permutation  $w_G$  is  $w$ .*

*(2) Any reduced Grassmannian graph is perfectly orientable. Moreover, it has an acyclic perfect orientation.*

*(3) A reduced Grassmannian graph  $G$  is complete of type  $(k, n)$  if and only if its helicity equals  $h(G) = k$  and the number of internal faces (excluding  $n$  boundary faces) equals*

$$f(k, n) - \sum_{v \in V_{\text{int}}} f(h(v), \deg(v)).$$

*where  $f(k, n) = (k - 1)(n - k - 1)$ . A reduced Grassmannian graph is complete if and only if it is not a proper induced subgraph of a larger reduced Grassmannian graph.*

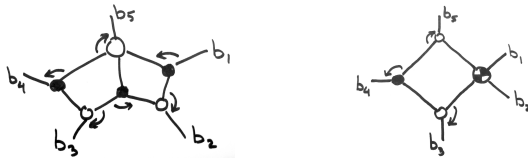


Figure 1: Two complete reduced Grassmannian graphs of type  $(2, 5)$  with 2 internal faces (left) and 1 internal face (right). The internal vertices of types  $(1, 3)$  and  $(1, 4)$  are colored in white, the type  $(2, 3)$  vertices colored in black, and the type  $(2, 4)$  vertex is “chessboard” colored.

Let us now describe a partial ordering and an equivalence relation on Grassmannian graphs.

**Definition 4.7.** For two Grassmannian graphs  $G$  and  $G'$ , we say that  $G$  *refines*  $G'$  (and that  $G'$  *coarsens*  $G$ ), if  $G$  can be obtained from  $G'$  by a sequence of the following operations: Replace an internal vertex of type  $(h, d)$  by a complete reduced Grassmannian graph of type  $(h, d)$ .

The *refinement order* on Grassmannian graphs is the partial order  $G \leq_{\text{ref}} G'$  if  $G$  refines  $G'$ . We say that  $G'$  *covers*  $G$ , if  $G'$  covers  $G$  in the refinement order.

Two Grassmannian graphs  $G$  and  $G'$  are *refinement-equivalent* if they are in the same connected component of the refinement order  $\leq_{\text{ref}}$ , that is, they can be obtained from each other by a sequence of refinements and coarsenings.

**Definition 4.8.** A Grassmannian graph is called a *plabic graph* if it is a minimal element in the refinement order.

The following is clear.

**Lemma 4.9.** *A Grassmannian graph is a plabic graph if and only if each internal vertex in the graph has type  $(1, 3)$ ,  $(2, 3)$ ,  $(0, 1)$ , or  $(1, 1)$ .*

In drawings of plabic and Grassmannian graphs, we color vertices of types  $(1, d)$  in white color, and vertices of types  $(d - 1, d)$  in black color.

Let us now describe almost minimal elements in the refinement order.

**Definition 4.10.** A Grassmannian graph  $G$  is called *almost plabic* if it covers a plabic graph (a minimal element) in the refinement order.

For example, the two graphs shown on [Figure 1](#) are almost plabic. The following lemma is also straightforward from the definitions.

**Lemma 4.11.** *Each almost plabic Grassmannian graph  $G$  has exactly one internal vertex (special vertex) of type  $(1, 4)$ ,  $(2, 4)$ ,  $(3, 4)$ ,  $(0, 2)$ ,  $(1, 2)$ , or  $(2, 2)$ , and all other internal vertices of types  $(1, 3)$ ,  $(2, 3)$ ,  $(0, 1)$ , or  $(1, 1)$ . An almost plabic graph with a special vertex of type  $(1, 4)$ ,  $(2, 4)$ , or  $(3, 4)$  covers exactly two plabic graphs. An almost plabic graph with a special vertex of type  $(0, 2)$ ,  $(1, 2)$ , or  $(2, 2)$  covers exactly one plabic graph.*

Note that a reduced Grassmannian graph cannot contain any vertices of degree 2. So each reduced almost plabic graph covers exactly two reduced plabic graphs.

**Definition 4.12.** Two plabic graphs are connected by a *move* of type  $(1, 4)$ ,  $(2, 4)$ , or  $(3, 4)$ , if they are both covered by an almost plabic graph with a special vertex of the corresponding type. Two plabic graphs  $G$  and  $G'$  are *move-equivalent* if they can be obtained from each other by a sequence of such moves.



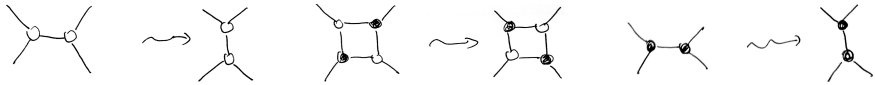


Figure 2: Three types of moves of plabic graphs: (1,4) contraction-uncontraction of white vertices, (2,4) square move, (3,4) contraction-uncontraction of black vertices.

Let us say that vertices of types  $(1,2)$ ,  $(0,d)$ ,  $(d,d)$  (except boundary leaves) are *extraneous*. A reduced graph cannot have a vertex of this form.

**Theorem 4.13.** (1) *For two reduced Grassmannian graphs  $G$  and  $G'$ , the graphs are refinement-equivalent if and only if they have the same decorated strand permutation  $w_G = w_{G'}$ .*

(2) cf. Postnikov [2006, Theorem 13.4] *For two reduced plabic graphs  $G$  and  $G'$ , the following are equivalent:*

- (a) *The graphs are move-equivalent.*
- (b) *The graphs are refinement-equivalent.*
- (c) *The graphs have the same decorated strand permutation  $w_G = w_{G'}$ .*

(3) *A Grassmannian graph is reduced if and only if it has no extraneous vertices and is not refinement-equivalent to a graph with a pair of parallel edges (two edges between the same vertices), or a loop-edge (an edge with both ends attached to the same vertex).*

(4) *A plabic graph is reduced if and only if it has no extraneous vertices and is not move-equivalent to a plabic graph with a pair of parallel edges or a loop-edge.*

**Remark 4.14.** Plabic graphs are similar to *wiring diagrams* that represent decompositions of permutations into products of adjacent transpositions. In fact, plabic graphs extend the notions of wiring diagrams and, more generally, *double wiring diagrams* of Fomin-Zelevinsky Fomin and A. Zelevinsky [1999], see Postnikov [2006, Remark 14.8, Figure 18.1]. Moves of plabic graphs are analogous to Coxeter moves of decompositions of permutations. Reduced plabic graphs extend the notion of reduced decompositions of permutations.

Let us now summarize the results about the relationship between positroids, Grassmannian and plabic graphs, decorated permutations, and Grassmann necklaces. For a decorated permutation  $w : [n] \rightarrow [n]$  (a permutation with fixed points colored 0 or 1), define  $\mathfrak{J}(w) := (J_1, \dots, J_n)$ , where

$$J_i = \{j \in [n] \mid c^{-i+1}w^{-1}(j) > c^{-i+1}(j)\} \cup \{j \in [n] \mid w(j) = j \text{ colored } 1\}.$$

The *helicity* of  $w$  is defined as  $h(w) := |J_1| = \cdots = |J_n|$ . Conversely, for a Grassmann necklace  $\mathfrak{g} = (J_1, \dots, J_n)$ , let

$$w(\mathfrak{g}) := w, \quad \text{where } w(i) = \begin{cases} j & \text{if } J_{i+1} = (J_i \setminus \{i\}) \cup \{j\}, \\ i \text{ (colored 0)} & \text{if } i \notin J_i = J_{i+1}, \\ i \text{ (colored 1)} & \text{if } i \in J_i = J_{i+1}. \end{cases}$$

**Theorem 4.15.** cf. [Postnikov \[ibid.\]](#) *The following sets are in one-to-one correspondence:*

1. *Positroids  $\mathfrak{M}$  of rank  $k$  on  $n$  elements.*
2. *Decorated permutation  $w$  of size  $n$  and helicity  $k$ .*
3. *Grassmann necklaces  $\mathfrak{g}$  of type  $(k, n)$ .*
4. *Move-equivalence classes of reduced plabic graphs  $G$  with  $n$  boundary vertices and helicity  $h(G) = k$ .*
5. *Refinement-equivalence classes of reduced Grassmannian graphs  $G'$  with  $n$  boundary vertices and helicity  $h(G') = k$ .*

*The following maps (described above in the paper) give explicit bijection between these sets and form a commutative diagram:*

1. *Reduced Grassmannian/plabic graphs to positroids:  $G \mapsto \mathfrak{M}(G)$ .*
2. *Reduced Grassmannian/plabic graphs to decorated permutations:  $G \mapsto w_G$ .*
3. *Positroids to Grassmann necklaces:  $\mathfrak{M} \mapsto \mathfrak{g}(\mathfrak{M})$ .*
4. *Grassmann necklaces to positroids:  $\mathfrak{g} \mapsto \mathfrak{M}(\mathfrak{g})$ ,*
5. *Grassmann necklaces to decorated permutations:  $\mathfrak{g} \mapsto w(\mathfrak{g})$ .*
6. *Decorated permutations to Grassmann necklaces:  $w \mapsto \mathfrak{g}(w)$ .*

*Proof of Theorems 4.4, 4.6, 4.13 and 4.15.* In case of plabic graphs, most of these results were proved in [Postnikov \[ibid.\]](#). The extension of results to Grassmannian graphs follows from a few easy observations.

Let  $G$  and  $G'$  be a pair of Grassmannian graphs such that  $G$  refines  $G'$ . Any perfect orientation of  $G$  induces a perfect orientation of  $G'$ . Conversely, any perfect orientation of  $G'$  can be extended (not uniquely, in general) to a perfect orientation of  $G$ . Thus  $G$  is perfectly orientable if and only if  $G'$  is perfectly orientable, and  $\mathfrak{M}(G) = \mathfrak{M}(G')$  and

$h(G) = h(G')$ . Any strand of  $G$  corresponds to a strand of  $G'$ . The graph  $G$  is reduced if and only if  $G'$  is reduced. If they are reduced, then they have the same decorated strand permutation  $w_G = w_{G'}$ . Finally, any Grassmannian graph can be refined to a plabic graph. So the results for plabic graphs imply the results for Grassmannian graphs.  $\square$

## 5 Weakly separated collections and cluster algebras

**Definition 5.1.** J. Scott [2005], cf. Leclerc and A. Zelevinsky [1998] Two subsets  $I, J \in \binom{[n]}{k}$  are *weakly separated* if there is no  $a < b < c < d$  such that  $a, c \in I \setminus J$  and  $b, d \in J \setminus I$ , or vice versa. A collection of subsets  $S \subset \binom{[n]}{k}$  is *weakly separated* if it is pairwise weakly separated.

This is a variation of Leclerc-Zelevinsky's notion of weak separation Leclerc and A. Zelevinsky [ibid.] given by J. Scott [2005]. It appeared in their study of quasi-commuting quantum minors.

**Definition 5.2.** The *face labelling* of a reduced Grassmannian graph  $G$  is the labelling of faces  $F$  of  $G$  by subsets  $I_F \subset [n]$  given by the condition: For each strand  $\alpha$  that goes from  $b_i$  to  $b_j$ , we have  $j \in I_F$  if and only if the face  $F$  lies to the left of the strand  $\alpha$  (with respect to the direction of the strand from  $b_i$  to  $b_j$ ).

Let us stay that two reduced plabic graphs are *contraction-equivalent* if they can be transformed to each other by the moves of type (1, 4) and (3, 4) (contraction-uncontraction moves) without using the move of type (2, 4) (square move).

**Theorem 5.3.** Oh, Postnikov, and D. E. Speyer [2015] (1) *Face labels of a reduced Grassmannian graph form a weakly separated collection in  $\binom{[n]}{k}$ , where  $k = h(G)$  is the helicity of  $G$ .*

(2) *Every maximal by inclusion weakly separated collection in  $\binom{[n]}{k}$  is the collection of face labels of a complete reduced plabic graph of type  $(k, n)$ .*

(3) *This gives a bijection between maximal by inclusion weakly separated collections in  $\binom{[n]}{k}$ , and contraction-equivalence classes of complete reduced plabic graphs of type  $(k, n)$ .*

**Remark 5.4.** Weakly separated collections are related to the *cluster algebra* structure Fomin and A. Zelevinsky [2002a, 2003], Berenstein, Fomin, and A. Zelevinsky [2005], and Fomin and A. Zelevinsky [2007] on the Grassmannian studied by J. S. Scott [2006]. In general, the cluster algebra on  $Gr(k, n)$  has infinitely many clusters. (See J. S. Scott [ibid.] for a classification of finite cases.) There is, however, a nicely behaved finite set of clusters, called the *Plücker clusters*, which are formed by subsets of the Plücker coordinates

$\Delta_I$ . According to [Oh, Postnikov, and D. E. Speyer \[2015, Theorem 1.6\]](#), the Plücker clusters for  $Gr(k, n)$  are exactly the sets  $\{\Delta_I\}_{I \in S}$  associated with maximal weakly-separated collections  $S \subset \binom{[n]}{k}$ . They are in bijection with contraction-equivalence classes of type  $(k, n)$  complete reduced plabic graphs, and are given by the  $k(n - k) + 1$  face labels of such graphs. Square moves of plabic graphs correspond to *mutations* of Plücker clusters in the cluster algebra.

[Theorem 5.3](#) implies an affirmative answer to the *purity conjecture* of [Leclerc and A. Zelevinsky \[1998\]](#). An independent solution of the purity conjecture was given by [Danilov, Karzanov, and Koshevoy \[2010\]](#) in terms of *generalized tilings*. The relationship between the parametrization a positroid cell given by a plabic graph  $G$  (see [Section 7](#) below) and the Plücker cluster  $\{\Delta_I\}_{I \in S}$  associated with the same graph  $G$  induces a nontrivial transformation, called the *twist map*, which was explicitly described by [Muller and D. E. Speyer \[2017\]](#). Weakly separated collections appeared in the study of *arrangements of equal minors* [Farber and Postnikov \[2016\]](#). In [Galashin and Postnikov \[2017\]](#) the notion of weakly separated collections was extended in the general framework of *oriented matroids* and *zonotopal tilings*.

## 6 Cyclically labelled Grassmannian

Let us reformulate the definition of the Grassmannian and its positive part in a more invariant form, which makes its cyclic symmetry manifest. In the next section, we will consider “little positive Grassmannians” associated with vertices  $v$  of a Grassmannian graph  $G$  whose ground sets correspond to the edges adjacent to  $v$ . There is no natural total ordering on such a set of edges, however there is the natural cyclic (clockwise) ordering.

We say that a *cyclic ordering* of a finite set  $C$  is a choice of closed directed cycle that visits each element of  $C$  exactly once. A total ordering of  $C$  is *compatible* with a cyclic ordering if it corresponds to a directed path on  $C$  obtained by removing an edge of the cycle. Clearly, there are  $|C|$  such total orderings.

**Definition 6.1.** Let  $C$  be a finite set of indices with a cyclic ordering of its elements, and let  $k$  be an integer between 0 and  $|C|$ . The *cyclically labelled Grassmannian*  $Gr(k, C)$  over  $\mathbb{R}$  is defined as the subvariety of the projective space  $\mathbb{P}^{\binom{|C|}{k}-1}$  with projective Plücker coordinates  $(\Delta_I)$  labelled by *unordered*  $k$ -element subsets  $I \subset C$  satisfying the Plücker relations written with respect to *any* total order “ $<$ ” on  $C$  compatible with the given cyclic ordering:

$$\sum_{i \in A \setminus B} (-1)^{|\{a \in A, a > i\}| + |\{b \in B, b < i\}|} \Delta_{A \setminus \{i\}} \Delta_{B \cup \{i\}} = 0,$$

where  $A$  and  $B$  are any  $(k + 1)$ -element and  $(k - 1)$ -element subsets of  $C$ , respectively. (More precisely,  $Gr(k, C)$  is the projective algebraic variety given by the radical of the ideal generated by the above Plücker relations.)

The *positive part*  $Gr^{>0}(k, C)$  is the subset of  $Gr(k, C)$  where the Plücker coordinates can be simultaneously rescaled so that  $\Delta_I > 0$ , for all  $k$ -element subsets  $I \subset C$ .

*Remark 6.2.* The Plücker relations (written as above) are invariant with respect to cyclic shifts of the ordering “ $<$ ”. Thus the definition of the cyclically labelled Grassmannian  $Gr(k, C)$  is independent of a choice of the total order on  $C$ . For example, for  $k = 2$  and  $C = \{1, 2, 3, 4\}$ ,  $Gr(2, C)$  is the subvariety of  $\mathbb{P}^{6-1}$  given by the Plücker relation:

$$\Delta_{\{1,3\}} \Delta_{\{2,4\}} = \Delta_{\{1,2\}} \Delta_{\{3,4\}} + \Delta_{\{1,4\}} \Delta_{\{2,3\}}.$$

Observe the cyclic symmetry of this relation! The ordering of indices  $2 < 3 < 4 < 1$  gives exactly the same  $Gr(2, C)$  with the same positive part  $Gr^{>0}(2, C)$ .

*Remark 6.3.* There is a subtle yet important difference between the cyclically labelled Grassmannian  $Gr(k, C)$  with the Plücker coordinates  $\Delta_I$  and the usual definition of the Grassmannian  $Gr(k, n)$ ,  $n = |C|$ , with the “usual Plücker coordinates” defined as the minors  $D_{(i_1, \dots, i_k)} = \det(A_{i_1, \dots, i_k})$  of submatrices  $A_{i_1, \dots, i_k}$  of a  $k \times n$  matrix  $A$ .

The  $D_{(i_1, \dots, i_k)}$  are labelled by *ordered* collections  $(i_1, \dots, i_k)$  of indices. They are *anti-symmetric* with respect to permutations of the indices  $i_1, \dots, i_k$ . On the other hand, the  $\Delta_{\{i_1, \dots, i_k\}}$  are labelled by *unordered* subsets  $I = \{i_1, \dots, i_k\}$ . So they are *symmetric* with respect to permutations of the indices  $i_1, \dots, i_k$ .

The “usual Plücker relations” for the  $D_{(i_1, \dots, i_k)}$  have the  $S_n$ -*symmetry* with respect to *all permutations* of the ground set. On the other hand, the above Plücker relations for the  $\Delta_{\{i_1, \dots, i_k\}}$  have only the  $\mathbb{Z}/n\mathbb{Z}$ -*symmetry* with respect to *cyclic shifts* of the ground set.

Of course, if we fix a total order of the ground set, we can rearrange the indices in  $D_{(i_1, \dots, i_k)}$  in the increasing order and identify  $D_{(i_1, \dots, i_k)}$ , for  $i_1 < \dots < i_k$ , with  $\Delta_{\{i_1, \dots, i_k\}}$ . This identifies the cyclically labelled Grassmannian  $Gr(k, C)$  with the usual Grassmannian  $Gr(k, n)$ . However, this isomorphism is not canonical because it depends on a choice of the total ordering of the index set. For even  $k$ , the isomorphism is *not invariant* under cyclic shifts of the index set.

## 7 Perfect orientation parametrization of positroid cells

Positroid cells were parametrized in Postnikov [2006] in terms of *boundary measurements* of perfect orientations of plabic graphs. Equivalent descriptions of this parametrization were given in terms of *network flows* by Talaska [2008] and in terms of *perfect matchings* Postnikov, D. Speyer, and L. Williams [2009] and Lam [2016]. Another interpretation of

this parametrization was motivated by physics [Arkani-Hamed, Bourjaily, Cachazo, Goncharov, Postnikov, and Trnka \[2016\]](#), where plabic graphs were viewed as *on-shell diagrams*, whose vertices represent little Grassmannians  $Gr(1, 3)$  and  $Gr(2, 3)$  and edges correspond to gluings, see also [Lam \[2016, Section 14\]](#) for a more mathematical description. Here we give a simple and invariant way to describe the parametrization in the general setting of Grassmannian graphs and their perfect orientations. It easily specializes to all the other descriptions. Yet it clarifies the idea of gluings of little Grassmannians.

Let  $G = (V, E)$  be a perfectly orientable Grassmannian graph with  $n$  boundary vertices and helicity  $h(G) = k$ , and let  $G_{\text{int}} = (V_{\text{int}}, E_{\text{int}})$  be its internal subgraph. Also let  $E_{\text{bnd}} = E \setminus E_{\text{int}}$  be the set of boundary edges of  $G$ .

Informally speaking, each internal vertex  $v \in V_{\text{int}}$  represents the “little Grassmannian”  $Gr(h, d)$ , where  $d$  is the degree of vertex  $v$  and  $h$  is its helicity. We “glue” these little Grassmannians along the internal edges  $e \in E_{\text{int}}$  of the graph  $G$  to form a subvariety in the “big Grassmannian”  $Gr(k, n)$ . Gluing along each edge kills one parameter. Let us give a more rigorous description of this construction.

For an internal vertex  $v \in V_{\text{int}}$ , let  $E(v) \subset E$  be the set of all adjacent edges to  $v$  (possibly including some boundary edges), which is cyclically ordered in the clockwise order (as we go around  $v$ ). Define the *positive vertex-Grassmannian*  $Gr^{>0}(v)$  as the positive part of the cyclically labelled Grassmannian

$$Gr^{>0}(v) := Gr^{>0}(h(v), E(v)).$$

Let  $(\Delta_J^{(v)})$  be the Plücker coordinates on  $Gr^{>0}(v)$ , where  $J$  ranges over the set  $\binom{E(v)}{h(v)}$  of all  $h(v)$ -element subsets in  $E(v)$ .

Let us define several positive tori (i.e., positive parts of complex tori). The *boundary positive torus* is  $T_{\text{bnd}}^{>0} := (\mathbb{R}_{>0})^{E_{\text{bnd}}} \simeq (\mathbb{R}_{>0})^n$ . The *internal positive torus* is  $T_{\text{int}}^{>0} := (\mathbb{R}_{>0})^{E_{\text{int}}}$ , and the *total positive torus*  $T_{\text{tot}}^{>0} := T_{\text{bnd}}^{>0} \times T_{\text{int}}^{>0}$ . The boundary/internal/total positive torus is the group of  $\mathbb{R}_{>0}$ -valued functions on boundary/internal/all edges of  $G$ .

These tori act on the positive vertex-Grassmannians  $Gr^{>0}(v)$  by rescaling the Plücker coordinates. For  $(t_e)_{e \in E} \in T_{\text{tot}}^{>0}$ ,

$$(t_e) : (\Delta_J^{(v)}) \mapsto \left( \left( \prod_{e \in J} t_e \right) \Delta_J^{(v)} \right).$$

The boundary torus  $T_{\text{bnd}}^{>0}$  also acts of the “big Grassmannian”  $Gr(k, n)$  as usual  $(t_1, \dots, t_n) : \Delta_I \mapsto (\prod_{i \in I} t_i) \Delta_I$ , for  $(t_1, \dots, t_n) \in T_{\text{bnd}}^{>0}$ .

Recall that, for a perfect orientation  $\mathcal{O}$  of  $G$ ,  $I(\mathcal{O})$  denotes the set of  $i \in [n]$  such that the boundary edge adjacent to  $b_i$  is directed towards the interior of  $G$  in  $\mathcal{O}$ . For an internal vertex  $v \in V_{\text{int}}$ , let  $J(v, \mathcal{O}) \subset E(v)$  be the subset of edges adjacent to  $v$  which are directed towards  $v$  in the orientation  $\mathcal{O}$ .

We are now ready to describe the *perfect orientation parametrization* of the positroid cells.

**Theorem 7.1.** *Let  $G$  be a perfectly orientable Grassmannian graph. Let  $\mu_G$  be the map defined on the direct product of the positive vertex-Grassmannians  $Gr^{>0}(v)$  and written in terms of the Plücker coordinates as*

$$\begin{aligned}\mu_G : \prod_{v \in V_{\text{int}}} Gr^{>0}(v) &\longrightarrow \mathbb{P}^{\binom{[n]}{k}-1} \\ \mu_G : \prod_{v \in V_{\text{int}}} (\Delta_J^{(v)})_{J \in \binom{E(v)}{h(v)}} &\longmapsto (\Delta_I)_{I \in \binom{[n]}{k}},\end{aligned}$$

where  $\Delta_I$  is given by the sum over all perfect orientations  $\mathcal{O}$  of the graph  $G$  such that  $I(\mathcal{O}) = I$ :

$$\Delta_I = \sum_{I(\mathcal{O})=I} \prod_{v \in V_{\text{int}}} \Delta_{J(v, \mathcal{O})}^{(v)}.$$

(1) The image of  $\mu_G$  is exactly the positroid cell  $\Pi_{\mathfrak{M}} \subset Gr^{\geq 0}(k, n) \subset \mathbb{P}^{\binom{[n]}{k}-1}$ , where  $\mathfrak{M} = \mathfrak{M}(G)$  is the positroid associated with  $G$ .

(2) The map  $\mu_G$  is  $T_{\text{int}}^{>0}$ -invariant and  $T_{\text{bnd}}^{>0}$ -equivariant, that is,  $\mu_G(t \cdot x) = \mu_G(x)$  for  $t \in T_{\text{int}}^{>0}$ , and  $\mu_G(t' \cdot x) = t' \cdot \mu_G(x)$  for  $t' \in T_{\text{bnd}}^{>0}$ .

(3) The map  $\mu_G$  induces the birational subtraction-free bijection  $\bar{\mu}_G$

$$\bar{\mu}_G : \left( \prod_{v \in V_{\text{int}}} Gr^{>0}(v) \right) / T_{\text{int}}^{>0} \longrightarrow \Pi_G$$

if and only if the Grassmannian graph  $G$  is a reduced.

*Remark 7.2.* The phrase “birational subtraction-free bijection” means that both  $\bar{\mu}_G$  and its inverse  $(\bar{\mu}_G)^{-1}$  can be expressed in terms of the Plücker coordinates by rational (or even polynomial) expressions written without using the “−” sign.

*Proof.* Part (2) is straightforward from the definitions. Let us first prove the remaining claims in the case when  $G$  is a plabic graph. In fact, in this case this construction gives exactly the *boundary measurement parametrization* of  $\Pi_G$  from Postnikov [2006, Section 5]. The Plücker coordinates for the boundary measurement parametrization were given in Postnikov [ibid., Proposition 5.3] and expressed by Talaska [2008, Theorem 1.1] in terms of network flows on the graph  $G$ . The construction of the boundary measurement parametrization (and Talaska’s formula) depends on a choice of a reference perfect orientation  $\mathcal{O}_0$ . One observes that any other perfect orientation  $\mathcal{O}$  of the plabic graph  $G$  is obtained from  $\mathcal{O}_0$  by reversing the edges along a network flow, which gives a bijection

between network flows and perfect orientations of  $G$ . This shows that the above expression for  $\Delta_I$  is equivalent to Talaska's formula, which proves the equivalence of the above perfect orientation parametrization and the boundary measurement parametrization from Postnikov [2006]. Parts (1) and (3) now follow from results of Postnikov [ibid.].

For an arbitrary Grassmannian graph  $G'$ , let  $G$  be a plabic graph that refines  $G'$ . We already know that each “little plabic graph”  $G_v$ , i.e., the subgraph of  $G$  that refines a vertex  $v$  of  $G'$ , parametrizes each positive vertex-Grassmannian  $Gr^{>0}(v)$  by a birational subtraction-free bijection  $\bar{\mu}_v := \bar{\mu}_{G_v}$ . We also know the map  $\bar{\mu}_G$  for the plabic graph  $G$  parametrizes the cell  $\Pi_{\mathfrak{m}(G)}$  if  $G$  reduced, or maps surjectively but not bijectively onto  $\Pi_{\mathfrak{m}(G)}$  if  $G$  is not reduced. Then the map  $\bar{\mu}_{G'}$  is given by the composition  $\bar{\mu}_G \circ (\times_{v \in V_{\text{int}}} (\bar{\mu}_v)^{-1})$  and the needed result follows.  $\square$

The above construction can be thought of as a gluing of the “big Grassmannian” out of “little Grassmannians.” This is similar to a tiling of a big geometric object (polytope) by smaller pieces (smaller polytopes). As we will see, this construction literally corresponds to certain subdivisions of polytopes.

## 8 Polyhedral subdivisions: Baues poset and fiber polytopes

In this section we discuss Billera-Sturmfels' theory Billera and Sturmfels [1992] of *fiber polytopes*, the *generalized Baues problem* Billera, Kapranov, and Sturmfels [1994], and *flip-connectivity*, see also Reiner [1999], Rambau and Santos [2000], Athanasiadis, Rambau, and Santos [1999], Athanasiadis [2001], and Athanasiadis and Santos [2002] for more details.

**8.1 The Baues poset of  $\pi$ -induced subdivisions.** Let  $\pi : P \rightarrow Q$  be an affine projection from one convex polytope  $P$  to another convex polytope  $Q = \pi(P)$ .

Informally, a  $\pi$ -induced polyhedral subdivision is a collection of faces of the polytope  $P$  that projects to a polyhedral subdivision of the polytope  $Q$ .

Here is a rigorous definition, see Billera and Sturmfels [1992]. Let  $A$  be the multiset of projections  $\pi(v)$  of vertices  $v$  of  $P$ . Each element  $\pi(v)$  of  $A$  is labelled by the vertex  $v$ . For  $\sigma \subset A$ , let  $\text{conv}(\sigma)$  denotes the convex hull of  $\sigma$ . We say that  $\sigma' \subset \sigma$  is a *face* of  $\sigma$  if  $\sigma'$  consists of all elements of  $\sigma$  that belong to a face of the polytope  $\text{conv}(\sigma)$ .

A  $\pi$ -induced subdivision is a finite collection  $S$  of subsets  $\sigma \subset A$ , called *cells*, such that

1. Each  $\sigma \in S$  is the projection under  $\pi$  of the vertex set of a face of  $P$ .
2. For each  $\sigma \in S$ ,  $\dim(\text{conv}(\sigma)) = \dim(Q)$ .



3. For any  $\sigma_1, \sigma_2 \in S$ ,  $\text{conv}(\sigma_1) \cap \text{conv}(\sigma_2) = \text{conv}(\sigma_1 \cap \sigma_2)$ .
4. For any  $\sigma_1, \sigma_2 \in S$ ,  $\sigma_1 \cap \sigma_2$  is either empty or a face of both  $\sigma_1$  and  $\sigma_2$ .
5.  $\bigcup_{\sigma \in S} \text{conv}(\sigma) = Q$ .

The *Baues poset*  $\omega(P \xrightarrow{\pi} Q)$  is the poset of all  $\pi$ -induced subdivisions partially ordered by *refinement*, namely,  $S \leq T$  means that, for every cell  $\sigma \in S$ , there exists a cell  $\tau \in T$  such that  $\sigma \subset \tau$ . This poset has a unique maximal element  $\hat{1}$ , called the *trivial subdivision*, that consists of a single cell  $\sigma = A$ . All other elements are called *proper subdivisions*. Let  $\hat{\omega}(P \xrightarrow{\pi} Q) := \omega(P \xrightarrow{\pi} Q) - \hat{1}$  be the poset of proper  $\pi$ -induced subdivisions obtained by removing the maximal element  $\hat{1}$ . The minimal elements of the Baues poset are called *tight*  $\pi$ -induced subdivisions.

Among all  $\pi$ -induced subdivisions, there is a subset of *coherent* subdivisions that come from linear *height functions*  $h : P \rightarrow \mathbb{R}$  as follows. For each  $q \in Q$ , let  $\bar{F}_q$  be the face of the fiber  $\pi^{-1}(q) \cap P$  where the height function  $h$  reaches its maximal value. The face  $\bar{F}_q$  lies in the relative interior of some face  $F_q$  of  $P$ . The collection of faces  $\{\bar{F}_q\}_{q \in Q}$  projects to a  $\pi$ -induced subdivision of  $Q$ . Let  $\omega_{\text{coh}}(P \xrightarrow{\pi} Q) \subseteq \omega(P \xrightarrow{\pi} Q)$  be the subposet of the Baues poset formed by the coherent  $\pi$ -induced subdivisions. This coherent part of the Baues poset is isomorphic to the face lattice of the convex polytope  $\Sigma(P \xrightarrow{\pi} Q)$ , called the *fiber polytope*, defined as the Minkowskii integral of fibers of  $\pi$  (the limit Minkowskii sums):

$$\Sigma(P \xrightarrow{\pi} Q) := \int_{q \in Q} (\pi^{-1}(q) \cap P) dq$$

In general, the whole Baues poset  $\omega(P \xrightarrow{\pi} Q)$  may not be polytopal.

**8.2 The generalized Baues problem and flip-connectivity.** For a finite poset  $\omega$ , the *order complex*  $\Delta\omega$  is the simplicial complex of all chains in  $\omega$ . The “topology of a poset  $\omega$ ” means the topology of the simplicial complex  $\Delta\omega$ . For example, if  $\omega$  is the face poset of a regular cell complex  $\Delta$ , then  $\Delta\omega$  is the barycentric subdivision of the cell complex  $\Delta$ ; and, in particular,  $\Delta\omega$  is homeomorphic to  $\Delta$ .

Clearly, the subposet  $\hat{\omega}_{\text{coh}}(P \xrightarrow{\pi} Q)$  of proper coherent  $\pi$ -induced subdivisions homotopy equivalent to a  $(\dim(P) - \dim(Q) - 1)$ -sphere, because it is the face lattice of a convex polytope of dimension  $\dim(P) - \dim(Q)$ , namely, the fiber polytope  $\Sigma(P \xrightarrow{\pi} Q)$ .

The *generalized Baues problem* (GBP) posed by [Billera](#), [Kapranov](#), and [Sturmfels \[1994\]](#) asks whether the same is true about the poset of all proper  $\pi$ -induced subdivisions. Is it true that  $\hat{\omega}(P \xrightarrow{\pi} Q)$  is homotopy equivalent to a  $(\dim(P) - \dim(Q) - 1)$ -sphere? In

general, the GBP is a hard question. Examples of Baues posets with disconnected topology were constructed by [Rambau and Ziegler \[1996\]](#) and more recently by [Liu \[2017\]](#). There are, however, several general classes of projections of polytopes, where the GBP has an affirmative answer, see the next section.

Another related question is about connectivity by flips. For a projection  $\pi : P \rightarrow Q$ , the *flip graph* is the restriction of the Hasse diagram of the Baues poset  $\omega(P \xrightarrow{\pi} Q)$  to elements of rank 0 (tight subdivisions) and rank 1 (subdivisions that cover a tight subdivision). The elements of rank 1 in the flip graph are called *flips*. The flip-connectivity problem asks whether the flip graph is connected. The coherent part of the flip graph is obviously connected, because it is the 1-skeleton of the fiber polytope  $\Sigma(P \xrightarrow{\pi} Q)$ .

The GBP and the flip-connectivity problem are related to each other, but, strictly speaking, neither of them implies the other, see [Reiner \[1999, Section 3\]](#) for more details.

**8.3 Triangulations and zonotopal tilings.** There are two cases of the above general setting that attracted a special attention in the literature.

The first case is when the polytope  $P$  is the  $(n - 1)$ -dimensional *simplex*  $\Delta^{n-1}$ . The multiset  $A$  of projections of vertices of the simplex can be an arbitrary multiset of  $n$  points, and  $Q = \text{conv}(A)$  can be an arbitrary convex polytope. In this case, the Baues poset  $\omega(\Delta^{n-1} \xrightarrow{\pi} Q)$  is the poset of all polyhedral subdivisions of  $Q$  (with vertices at  $A$ ); tight  $\pi$ -induced subdivisions are *triangulations* of  $Q$ ; and the fiber polytope  $\Sigma(\Delta^{n-1} \xrightarrow{\pi} Q)$  is exactly is the *secondary polytope* of Gelfand-Kapranov-Zelevinsky [Gelfand, Kapranov, and A. V. Zelevinsky \[1994\]](#), which appeared in the study of discriminants.

In particular, for a projection of the simplex  $\Delta^{n-1}$  to an  $n$ -gon  $Q$ ,  $\pi$ -induced subdivisions are exactly the subdivisions of the  $n$ -gon by noncrossing chords. All of them are coherent. Tight subdivisions are triangulations of the  $n$ -gon. There are the Catalan number  $C_{n-2} = \frac{1}{n-1} \binom{2n-4}{n-2}$  of triangulations of the  $n$ -gon. The fiber polytope (or the secondary polytope) in this case is the Stasheff *associahedron*.

Another special case is related to projections  $\pi : P \rightarrow Q$  of the *hypercube*  $P = \square_n := [0, 1]^n$ . The projections  $Q = \pi(\square_n)$  of the hypercube form a special class of polytopes, called *zonotopes*. In this case,  $\pi$ -induced subdivisions are *zonotopal tilings* of zonotopes  $Q$ . According to Bohne-Dress theorem [Bohne \[1992\]](#), zonotopal tilings of  $Q$  are in bijection with *1-element extensions* of the oriented matroid associated with the zonotope  $Q$ .

For a projection of the  $n$ -hypercube  $\square_n$  to a 1-dimensional line segment, the fiber polytope is the *permutohedron*. For a projection  $\pi : \square_n \rightarrow Q$  of the  $n$ -hypercube  $\square_n$  to a  $2n$ -gon  $Q$ , *fine zonotopal tilings* (i.e., tight  $\pi$ -induced subdivisions) are known as *rhombus tilings* of the  $2n$ -gon. They correspond to commutation classes of *reduced redecompositions* of the longest permutation  $w_o \in S_n$ .

## 9 Cyclic polytopes and cyclic zonotopes

Fix two integers  $n$  and  $0 \leq d \leq n - 1$ .

**Definition 9.1.** A *cyclic projection* is a linear map

$$\pi : \mathbb{R}^n \rightarrow \mathbb{R}^{d+1}, \quad \pi : x \mapsto Mx$$

given by a  $(d + 1) \times n$  matrix  $M = (u_1, \dots, u_n)$  (the  $u_i$  are the column vectors) with all positive maximal  $(d + 1) \times (d + 1)$  minors and such that  $f(u_1) = \dots = f(u_n) = 1$  for some linear form  $f : \mathbb{R}^{d+1} \rightarrow \mathbb{R}$ . In other words,  $M$  represents a point of the positive Grassmannian  $Gr^{>0}(d + 1, n)$  with columns  $u_i$  rescaled so that they all lie on the same affine hyperplane  $H_1 = \{y \in \mathbb{R}^{d+1} \mid f(y) = 1\}$ .

The *cyclic polytope* is the image under a cyclic projection  $\pi$  of the standard  $(n - 1)$ -dimensional simplex  $\Delta^{n-1} := \text{conv}(e_1, \dots, e_n)$

$$C(n, d) := \pi(\Delta^{n-1}) \subset H_1.$$

The *cyclic zonotope* is the image of the standard  $n$ -hypercube  $\boxplus_n := [0, 1]^n \subset \mathbb{R}^n$

$$Z(n, d + 1) := \pi(\boxplus_n) \subset \mathbb{R}^{d+1}.$$

Remark that, for each  $n$  and  $d$ , there are many combinatorially (but not linearly) isomorphic cyclic polytopes  $C(n, d)$  and cyclic zonotopes  $Z(n, d + 1)$  that depend on a choice of the cyclic projection  $\pi$ . Clearly,  $C(n, d) = Z(n, d + 1) \cap H_1$ .

[Ziegler \[1993\]](#) identified *fine zonotopal tilings* of the cyclic zonotope  $Z(n, d + 1)$ , i.e., the minimal elements of the Baues poset  $\omega(\boxplus_n \xrightarrow{\pi} Z(n, d + 1))$ , with elements of Manin-Shehtman's *higher Bruhat order* [Manin and Shehtman \[1986\]](#), also studied by [Voevodskii and Kapranov \[1991\]](#). According to results of [Sturmfels and Ziegler \[1993\]](#), [Ziegler \[1993\]](#), [Rambau \[1997\]](#), and [Rambau and Santos \[2000\]](#), the GBP and flip-connectivity have affirmative answers in these cases.

**Theorem 9.2.** (1) [Sturmfels and Ziegler \[1993\]](#) For  $\pi : \boxplus_n \rightarrow Z(n, d + 1)$ , the poset of proper zonotopal tilings of the cyclic zonotope  $Z(n, d + 1)$  is homotopy equivalent to an  $(n - d - 2)$ -dimensional sphere. The set of fine zonotopal tilings of  $Z(n, d + 1)$  is connected by flips.

(2) [Rambau and Santos \[2000\]](#) For  $\pi : \Delta^{n-1} \rightarrow C(n, d)$ , the poset of proper subdivisions of the cyclic polytope  $C(n, d)$  is homotopy equivalent to an  $(n - d - 2)$ -dimensional sphere. [Rambau \[1997\]](#) The set of triangulations of the cyclic polytope  $C(n, d)$  is connected by flips.

## 10 Cyclic projections of the hypersimplex

Fix three integers  $0 \leq k \leq n$  and  $0 \leq d \leq n - 1$ . The *hypersimplex*  $\Delta_{kn} := \text{conv} \{e_I \mid I \in \binom{[n]}{k}\}$  is the  $k$ -th section of the  $n$ -hypercube  $\boxplus_n \subset \mathbb{R}^n$

$$\Delta_{kn} = \boxplus_n \cap \{x_1 + \cdots + x_n = k\}.$$

Let  $\pi : \mathbb{R}^n \rightarrow \mathbb{R}^{d+1}$  be a cyclic projection as above. Define the polytope

$$Q(k, n, d) := \pi(\Delta_{kn}) = Z(n, d + 1) \cap H_k,$$

where  $H_k$  is the affine hyperplane  $H_k := \{y \in \mathbb{R}^{d+1} \mid f(y) = k\}$ . Clearly, for  $k = 1$ , the polytope  $Q(1, n, d)$  is the cyclic polytope  $C(n, d)$ .

Let  $\omega(k, n, d)$  be the Baues poset of  $\pi$ -induced subdivisions for a cyclic projection  $\pi : \Delta_{kn} \rightarrow Q(k, n, d)$ :

$$\omega(k, n, d) := \omega(\Delta_{kn} \xrightarrow{\pi} Q(k, n, d)).$$

Let  $\omega_{\text{coh}}^{\pi}(k, n, d) := \omega_{\text{coh}}(\Delta_{kn} \xrightarrow{\pi} Q(k, n, d)) \subseteq \omega(k, n, d)$  be its coherent part. Note that the coherent part  $\omega_{\text{coh}}^{\pi}(k, n, d)$  depends on a choice of the cyclic projection  $\pi$ , but the whole poset  $\omega(k, n, d)$  is independent of any choices. The coherent part  $\omega_{\text{coh}}^{\pi}(k, n, d)$  may not be equal  $\omega(k, n, d)$ . For example they are not equal for  $(k, n, d) = (3, 6, 2)$ .

The poset  $\omega(k, n, d)$  is a generalization of the Baues poset of subdivisions of the cyclic polytope  $C(n, d)$ , and is related to the Baues poset of zonotopal tilings of the cyclic zonotope  $Z(n, d + 1)$  in an obvious manner. For  $k = 1$ ,  $\omega(1, n, d) = \omega(\Delta^{n-1} \xrightarrow{\pi} C(n, d))$ . For any  $k$ , there is the order preserving  $k$ -th section map

$$\text{Section}_k : \omega(\boxplus_n \xrightarrow{\pi} Z(n, d + 1)) \rightarrow \omega(k, n, d)$$

that send a zonotopal tiling of  $Z(n, d + 1)$  to its section by the hyperplane  $H_k$ .

Let  $\omega_{\text{lift}}(k, n, d) \subseteq \omega(k, n, d)$  be the image of the map  $\text{Section}_k$ . We call the elements of  $\omega_{\text{lift}}(k, n, d)$  the *lifting*  $\pi$ -induced subdivisions. They form the subset of  $\pi$ -induced subdivisions from  $\omega(k, n, d)$  that can be lifted to a zonotopal tiling of the cyclic zonotope  $Z(n, d + 1)$ . Clearly, we have

$$\omega_{\text{coh}}^{\pi}(k, n, d) \subseteq \omega_{\text{lift}}(k, n, d) \subseteq \omega(k, n, d).$$

The equality of the sets of minimal elements of  $\omega(k, n, d)$  and  $\omega_{\text{lift}}(k, n, d)$  was proved in the case  $k = 1$  by [Rambau and Santos \[2000\]](#), who showed that all triangulations of the cyclic polytope  $C(n, d)$  are lifting triangulations. For  $d = 2$ , the equality follows from the result of [Galashin \[2016\]](#) ([Theorem 11.7](#) below) about plabic graphs, as we will explain in the next section.

**Theorem 10.1.** *The minimal elements of the posets  $\omega_{\text{lift}}(k, n, d)$  and  $\omega(k, n, d)$  are the same in the following cases: (1)  $k = 1$  and any  $n, d$ ; (2)  $d = 2$  and any  $k, n$ .*

Flip-connectivity [Sturmfels and Ziegler \[1993\]](#) and [Ziegler \[1993\]](#) of zonotopal tilings of  $Z(n, d + 1)$  easily implies the following claim.

**Lemma 10.2.** *The minimal elements of  $\omega_{\text{lift}}(k, n, d)$  are connected by flips.*

Indeed, for any pair of fine zonotopal tilings  $T$  and  $T'$  of  $Z(n, d + 1)$  connected by a flip, their  $k$ -sections  $\text{Section}_k(T)$  and  $\text{Section}_k(T')$  are either equal to each other or connected by a flip.

The Baues posets of the form  $\omega(k, n, d)$  are good candidates for a general class of projections of polytopes where the GBP and flip-connectivity problem might have affirmative answers.

**Problem 10.3.** *Is the poset  $\omega(k, n, d) - \hat{1}$  homotopy equivalent of a sphere? Can its minimal elements be connected by flips? Is it true that  $\omega_{\text{lift}}(k, n, d) = \omega(k, n, d)$ ?*

*Example 10.4.* For  $d = 1$ , the Baues poset  $\omega(k, n, 1)$  is already interesting. Its minimal elements correspond to *monotone paths* on the hypersimplex  $\Delta_{kn}$ , which are increasing paths that go along the edges of the hypersimplex  $\Delta_{kn}$ . Such paths are the subject of the original (non-generalized) Baues problem [Baues \[1980\]](#), which was proved by [Billera, Kapranov, and Sturmfels \[1994\]](#) (for any 1-dimensional projection of a polytope). More specifically, monotone paths on  $\Delta_{kn}$  correspond to directed paths from  $[1, k]$  to  $[n - k + 1, n]$  in the directed graph on  $\binom{[n]}{k}$  with edges  $I \rightarrow J$  if  $J = (I \setminus \{i\}) \cup \{j\}$  for  $i < j$ .

It is not hard to see that, for  $k = 1, n - 1$ , there are  $2^{n-2}$  monotone paths, and the posets  $\omega(1, n, 1)$  and  $\omega(n - 1, n, 1)$  are isomorphic to the Boolean lattice  $B_{n-2}$ , i.e., the face poset of the hypercube  $\square_{n-2}$ . For  $n = 2, 3, 4, 5$ , the Baues poset  $\omega(2, n, 1)$  has 1, 2, 10, 62 minimal elements.

Monotone paths on  $\Delta_{kn}$  might have different lengths. The *longest* monotone paths are in an easy bijection with *standard Young tableaux* of the rectangular shape  $k \times (n - k)$ . By the hook-length formula, their number is  $(k(n - k))! \prod_{i=0}^{n-k-1} \frac{i!}{(k+i)!}$ .

Note, however,  $\omega(k, n, d) \neq \omega_{\text{lift}}(k, n, d)$  for  $(k, n, d) = (2, 5, 1)$ . Indeed, Galashin pointed out that the monotone path  $\{1, 2\} \rightarrow \{1, 3\} \rightarrow \{1, 4\} \rightarrow \{2, 4\} \rightarrow \{3, 4\}$  *cannot* be lifted to a rhombus tiling of the  $2n$ -gon  $Z(n, 2)$ , because it is not weakly separated.

## 11 Grassmannian graphs as duals of polyhedral subdivisions induced by projections of hypersimplices

Let us now discuss the connection between the positive Grassmannian and combinatorics of polyhedral subdivisions. In fact, the positive Grassmannian is directly related to the setup of the previous section for  $d = 2$ .

**Theorem 11.1.** *The poset of complete reduced Grassmannian graphs of type  $(k, n)$  ordered by refinement is canonically isomorphic to the Baues poset  $\omega(k, n, 2)$  of  $\pi$ -induced subdivisions for a 2-dimensional cyclic projection  $\pi$  of the hypersimplex  $\Delta_{kn}$ . Under this isomorphism, plabic graphs correspond to tight  $\pi$ -induced subdivisions and moves of plabic graphs correspond to flips between tight  $\pi$ -induced subdivisions.*

[Theorem 4.13\(2\)](#) ([Postnikov \[2006, Theorem 13.4\]](#)) immediately implies flip-connectivity.

**Corollary 11.2.** *The minimal elements of Baues poset  $\omega(k, n, 2)$  are connected by flips.*

*Example 11.3.* The Baues poset  $\omega(1, n, 2)$  is the poset of subdivisions of an  $n$ -gon by non-crossing chords, i.e., it is the Stasheff's *associahedron*. Its minimal elements correspond to the Catalan number  $\frac{1}{n-1} \binom{2n-4}{n-2}$  triangulations of the  $n$ -gon.

We can think of the Baues posets  $\omega(k, n, 2)$  as some kind of “generalized associahedra.” In general, they are not polytopal. But they share some nice features with the associahedron. It is well-known that every face of the associahedron is a direct product of smaller associahedra. The same is true for all  $\omega(k, n, 2)$ .

**Proposition 11.4.** *For any element  $S$  in  $\omega(k, n, 2)$ , the lower order interval  $\{S' \mid S' \leq S\}$  in the Baues poset  $\omega(k, n, 2)$  is a direct product of Baues posets of the same form  $\omega(k', n', 2)$ .*

*Proof.* This is easy to see in terms of complete reduced Grassmannian graphs  $G$ . Indeed, for any  $G$ , all refinements of  $G'$  are obtained by refining all vertices of  $G$  independently from each other.  $\square$

This property is related to the fact that every face of the hypersimplex  $\Delta_{kn}$  is a smaller hypersimplex, as we discuss below.

*Remark 11.5.* Among all reduced Grassmannian/plabic graphs, there is a subset of *coherent* (or *regular*) graphs, namely the ones that correspond to the coherent  $\pi$ -induced subdivisions from  $\omega_{\text{coh}}(k, n, 2)$ . Each of these graphs can be explicitly constructed in terms of a *height function*. This subclass depends on a choice of the cyclic projection  $\pi$ . Regular plabic graphs are related to the study of *soliton solutions* of Kadomtsev-Petviashvili (KP)

equation, see [Kodama and L. K. Williams \[2011\]](#) and [Kodama and L. Williams \[2014\]](#). We will investigate the class of regular plabic graphs in [Galashin, Postnikov, and L. Williams \[n.d.\]](#).

Let us now give more details on the correspondence between Grassmannian graphs and subdivisions. A cyclic projection  $\pi : \Delta_{kn} \rightarrow Q(k, n, 2)$  is the linear map given by a  $3 \times n$  matrix  $M = (u_1, \dots, u_n)$  such that  $[u_1, \dots, u_n] \in Gr^{>0}(3, n)$  and  $u_1, \dots, u_n$  all lie on the same affine plane  $H_1 \subset \mathbb{R}^3$ . Without loss of generality, assume that  $H_1 = \{(x, y, z) \mid z = 1\}$ . The positivity condition means that the points  $\pi(u_1), \dots, \pi(u_n)$  form a convex  $n$ -gon with vertices arranged in the counterclockwise order.

The polytope  $Q := Q(k, n, 2) = \pi(\Delta_{kn})$  is the convex  $n$ -gon in the affine plane  $H_k = \{(x, y, k)\} \subset \mathbb{R}^3$  with the vertices  $\pi(e_{[1,k]}), \pi(e_{[2,k+1]}), \dots, \pi(e_{[n,k-1]})$  (in the counterclockwise order) corresponding to all consecutive cyclic intervals of size  $k$  in  $[n]$ .

Notice that each face  $\gamma$  of the hypersimplex  $\Delta_{kn}$  is itself a smaller hypersimplex of the form

$$\gamma_{I_0, I_1} := \{(x_1, \dots, x_n) \in \Delta_{kn} \mid x_i = 0 \text{ for } i \in I_0, x_j = 1 \text{ for } j \in I_1\}$$

where  $I_0$  and  $I_1$  are disjoint subsets of  $[n]$ . So  $\gamma \simeq \Delta_{hm}$ , where  $h = k - |I_0|$  and  $m = n - |I_0| - |I_1|$ . The projection  $\pi$  maps the face  $\gamma$  to the  $m$ -gon  $\pi(\gamma)$  that carries an additional parameter  $h$ .

Thus the  $\pi$ -induced subdivisions  $S$  are in bijective correspondence with the tilings of the  $n$ -gon  $Q$  by smaller convex polygons such that:

1. Each vertex has the form  $\pi(e_I)$  for  $I \in \binom{[n]}{k}$ .
2. Each edge has the form  $[\pi(e_I), \pi(e_J)]$  for two  $k$ -element subsets  $I$  and  $J$  such that  $|I \cap J| = k - 1$ .
3. Each face is an  $m$ -gon of the form  $\pi(\gamma_{I_0, I_1})$ , as above.

Let  $S^*$  be the planar dual of such a tiling  $S$ . The graph  $S^*$  has exactly  $n$  boundary vertices  $b_i$  corresponding to the sides  $[\pi(e_{[i, i+k-1]}), \pi(e_{[i+1, i+k]})]$  of the  $n$ -gon  $Q$ . The internal vertices  $v$  of  $S^*$  (corresponding to faces  $\gamma$  of  $S$ ) are equipped with the parameter  $h = h(v) \in \{0, \dots, \deg(v)\}$ . Thus  $S^*$  has the structure of a Grassmannian graph. Moreover, each face  $F$  of  $S^*$  (corresponding a vertex  $\pi(e_I)$  of  $S$ ) is labelled by a subset  $I \in \binom{[n]}{k}$ . We can now make the previous theorem more precise.

**Theorem 11.6.** *The map  $S \mapsto S^*$  is an isomorphism between the Baues poset  $\omega(k, n, 2)$  and complete reduced Grassmannian graphs  $G$  of type  $(k, n)$ . For each face  $F$  of  $G = S^*$  corresponding to a vertex  $\pi(e_I)$  of  $S$ , the subset  $I \subset \binom{[n]}{k}$  is exactly the face label  $I_F$  (see [Definition 5.2](#)).*

*Proof.* Let us first show that *tight*  $\pi$ -induced subdivisions  $S$  are in bijection with complete reduced *plabic* graphs of type  $(k, n)$ . That means that, in addition to the conditions (1), (2), (3) above, we require the tiling  $S$  of the  $n$ -gon  $Q$  has all triangular faces. So  $S$  is a triangulation of the  $n$ -gon  $Q$  of a special kind, which we call a *plabic triangulation*.

Such plabic triangulations of the  $n$ -gon are closely related to *plabic tilings* from [Oh, Postnikov, and D. E. Speyer \[2015\]](#). The only difference between plabic triangulations and plabic tilings is that the latter correspond not to (3-valent) plabic graphs (as defined in the current paper) but to *bipartite* plabic graphs. A bipartite plabic graph  $G$  is exactly a Grassmannian graph such that each internal vertex either has type  $(1, d)$  (white vertex) or type  $(d - 1, d)$  (black vertex), and every edge of  $G$  connects vertices of different colors. Each reduced 3-valent plabic graph  $G'$  can be easily converted into a bipartite plabic graph  $G$  by contrating edges connecting vertices of the same color. It was shown in [Oh, Postnikov, and D. E. Speyer \[ibid., Theorem 9.12\]](#) that the planar dual graph of any reduced bipartite plabic graph  $G$  can be embedded inside an  $n$ -gon as a plabic tiling with black and white regions and all vertices of the form  $\pi(e_I)$ . If we now subdivide the black and white regions of such plabic tiling by chords into triangles, we can get back the plabic triangulation associated with a (3-valent) plabic graph  $G'$ . This shows that any complete reduced (3-valent) plabic graph is indeed the planar dual of a tight  $\pi$ -induced subdivision.

On the other hand, for each plabic triangulation  $S$  we can construct the plabic graph by taking its planar dual  $G = S^*$  as described above. It is easy to check from the definitions that the decorated strand permutation  $w$  of  $G$  is exactly  $w(i) = i + k \pmod{n}$ . It remains to show that this plabic graph  $G$  is reduced. Suppose that  $G$  is not reduced. Then by [Theorem 4.13\(5\)](#), after possibly applying a sequence of moves  $(1, 4)$ ,  $(2, 4)$ , and/or  $(3, 4)$ , we get a plabic graph with a pair of parallel edges or with a loop-edge. It is straightforward to check that applying the moves  $(1, 4)$ ,  $(2, 4)$ ,  $(3, 4)$  corresponds to local transformations of the plabic triangulation  $S$ , and transforms it into another plabic triangulations  $S'$ . However, it is clear that if a plabic graph  $G$  contains parallel edges or a loop-edge, then the dual graph is *not* a plabic triangulation. So we get a contradiction, which proves the result for plabic graphs and tight subdivisions.

Now let  $G'$  be any complete reduced Grassmannian graph of type  $(k, n)$ , and let  $G$  be its plabic refinement. We showed that we can embed the planar dual graph  $G^*$  as a plabic triangulation  $S$  into the  $n$ -gon. The union of triangles in  $S$  that correspond to a single vertex  $v$  of  $G'$  covers a region inside  $Q$ . We already know that this region is a convex  $m$ -gon (because we already proved the correspondence for plabic graphs). Thus, for each vertex of  $G'$ , we get a convex polygon in  $Q$  and all these polygons form  $\pi$ -induced subdivision. So we proved that the planar dual of  $G'$  can be embedded as a polyhedral subdivision of  $Q$ . The inverse map is  $S' \mapsto G' = (S')^*$ .  $\square$

Let us mention a related result of [Galashin \[2016\]](#).



**Theorem 11.7.** *Galashin [2016] Complete reduced plabic graphs of type  $(k, n)$  are exactly the dual graphs of sections of fine zonotopal tilings of the 3-dimensional cyclic zonotope  $Z(n, 3)$  by the hyperplane  $H_k$ .*

In view of the discussion above, this result means that any tight  $\pi$ -induced subdivision in  $\omega(k, n, 2)$  can be lifted to a fine zonotopal tiling of the cyclic zonotope  $Z(n, 3)$ . In other words, the posets  $\omega(k, n, 2)$  and  $\omega_{\text{lift}}(k, n, 2)$  have the same sets of minimal elements. A natural question to ask: Is the same true for all (not necessarily minimal) elements of  $\omega(k, n, 2)$ ?

## 12 Membranes and discrete Plateau's problem

Membranes from a project with T. Lam [2018] provide another related interpretation of plabic graphs. Let  $\Phi = \{e_i - e_j \mid i \neq j\} \in \mathbb{R}^n$ , where  $e_1, \dots, e_n$  are the standard coordinate vectors.

**Definition 12.1.** T. Lam [ibid.] A loop  $L$  is a closed piecewise-linear curve in  $\mathbb{R}^n$  formed by line segments  $[a, b]$  such that  $a, b \in \mathbb{Z}^n$  and  $a - b \in \Phi$ .

A membrane  $M$  with boundary loop  $L$  is an embedding of a 2-dimensional disk into  $\mathbb{R}^n$  such that  $L$  is the boundary of  $M$ , and  $M$  is made out of triangles  $\text{conv}(a, b, c)$ , where  $a, b, c \in \mathbb{Z}$  and  $a - b, b - c, a - c \in \Phi$ .

A minimal membrane  $M$  is a membrane that has minimal possible area (the number of triangles) among all membranes with the same boundary loop  $L$ .

The problem about finding a minimal membrane  $M$  with a given boundary loop  $L$  is a discrete version *Plateau's problem* about minimal surface. Informally speaking, membranes correspond to (the duals of) plabic graphs, and minimal membranes correspond to reduced plabic graphs. Here is a more careful claim.

**Theorem 12.2.** T. Lam [ibid.] Let  $w \in S_n$  be a permutation without fixed points with helicity  $h(w) = k$ . Let  $L_w$  the closed loop inside the hypersimplex  $\Delta_{kn}$  formed by the line segments  $[a_1, a_2], [a_2, a_3], \dots, [a_{n-1}, a_n], [a_n, a_1]$  such that  $a_{i+1} - a_i = e_{w(i)} - e_i$ , for  $i = 1, \dots, n$ , with indices taken modulo  $n$ .

Then minimal membranes  $M$  with boundary loop  $L_w$  are in bijection with reduced plabic graphs  $G$  with strand permutation  $w$ . Explicitly, the correspondence is given as follows. Faces  $F$  of  $G$  with face labels  $I = I_F$  correspond to vertices  $e_I$  of the membrane  $M$ . Vertices of  $G$  with 3 adjacent faces labeled by  $I_1, I_2, I_3$  correspond to triangles  $\text{conv}(e_{I_1}, e_{I_2}, e_{I_3})$  in  $M$ .

Moves of plabic graphs correspond to local area-preserving transformations of membranes. Any two minimal membranes with the same boundary loop  $L_w$  can be obtained from each other by these local transformations.

### 13 Higher positive Grassmannians and Amplituhedra

The relation between the positive Grassmannian  $Gr^{>0}(k, n)$  and the Baues poset  $\omega(k, n, 2)$  raises a natural question: What is the geometric counterpart of the Baues poset  $\omega(k, n, d)$  for any  $d$ ? These “higher positive Grassmannians” should generalize  $Gr^{>}(k, n)$  in the same sense as Manin-Shekhtman’s higher Bruhat orders generalize the weak Bruhat order. The first guess is that they might be related to amplituhedra.

Arkani-Hamed and Trnka [2014] motivated by the study of scattering amplitudes in  $\mathfrak{N} = 4$  supersymmetric Yang-Mills (SYM) theory, defined the *amplituhedron*  $A_{n,k,m} = A_{n,k,m}(Z)$  as the image of the nonnegative Grassmannian  $Gr^{\geq}(k, n)$  under the “linear projection”

$$\tilde{Z} : Gr^{\geq 0}(k, n) \rightarrow Gr(k, k + m), \quad [A] \mapsto [A Z^T]$$

induced by a totally positive  $(k + m) \times n$  matrix  $Z$ , for  $0 \leq m \leq n - k$ . The case  $m = 4$  is of importance for physics.

In general, the amplituhedron  $A_{n,k,m}$  has quite mysterious geometric and combinatorial structure. Here are few special cases where its structure was understood better. For  $m = n - k$ ,  $A_{n,k,n-k}$  is isomorphic to the nonnegative Grassmannian  $Gr^{\geq 0}(k, n)$ . For  $k = 1$ ,  $A_{n,1,m}$  is (the projectivization of) the cyclic polytope  $C(n, m)$ . For  $m = 1$ , Karp and L. K. Williams [2017] showed that the structure of the amplituhedron  $A_{n,k,1}$  is equivalent to the complex of bounded regions of the cyclic hyperplane arrangement. In general, the relationship between the amplituhedron  $A_{n,k,m}$  and polyhedral subdivisions is yet to be clarified.

### References

- N. Arkani-Hamed and J. Trnka (2014). “The amplituhedron”. *J. High Energy Phys.* 10 (33) (cit. on pp. 3202, 3227).
- Nima Arkani-Hamed, Jacob Bourjaily, Freddy Cachazo, Alexander Goncharov, Alexander Postnikov, and Jaroslav Trnka (2016). *Grassmannian geometry of scattering amplitudes*. Cambridge University Press, Cambridge, pp. ix+194. MR: 3467729 (cit. on pp. 3201, 3215).
- Christos A. Athanasiadis (2001). “Zonotopal subdivisions of cyclic zonotopes”. *Geom. Dedicata* 86.1-3, pp. 37–57. MR: 1856417 (cit. on p. 3217).
- Christos A. Athanasiadis, Jörg Rambau, and Francisco Santos (1999). “The generalized Baues problem for cyclic polytopes. II”. *Publ. Inst. Math. (Beograd) (N.S.)* 66(80). Geometric combinatorics (Kotor, 1998), pp. 3–15. MR: 1765037 (cit. on p. 3217).

- Christos A. Athanasiadis and Francisco Santos (2002). “On the topology of the Baues poset of polyhedral subdivisions”. *Topology* 41.3, pp. 423–433. MR: [1910043](#) (cit. on p. [3217](#)).
- H. J. Baues (1980). “Geometry of loop spaces and the cobar construction”. *Mem. Amer. Math. Soc.* 25.230, pp. ix+171. MR: [567799](#) (cit. on p. [3222](#)).
- Arkady Berenstein, Sergey Fomin, and Andrei Zelevinsky (2005). “Cluster algebras. III. Upper bounds and double Bruhat cells”. *Duke Math. J.* 126.1, pp. 1–52. MR: [2110627](#) (cit. on pp. [3200](#), [3212](#)).
- L. J. Billera, M. M. Kapranov, and B. Sturmfels (1994). “Cellular strings on polytopes”. *Proc. Amer. Math. Soc.* 122.2, pp. 549–555. MR: [1205482](#) (cit. on pp. [3201](#), [3217](#), [3218](#), [3222](#)).
- Louis J. Billera and Bernd Sturmfels (1992). “Fiber polytopes”. *Ann. of Math. (2)* 135.3, pp. 527–549. MR: [1166643](#) (cit. on pp. [3201](#), [3217](#)).
- J. Bohne (1992). “Eine kombinatorische analyse zonotopaler raumaufteilungen”. PhD thesis. Univ. Bielefeld (cit. on p. [3219](#)).
- S. Chakravarty and Y. Kodama (2009). “Soliton solutions of the KP equation and application to shallow water waves”. *Stud. Appl. Math.* 123.1, pp. 83–151. MR: [2538287](#) (cit. on p. [3201](#)).
- Sylvie Corteel and Lauren K. Williams (2007). “Tableaux combinatorics for the asymmetric exclusion process”. *Adv. in Appl. Math.* 39.3, pp. 293–310. MR: [2352041](#) (cit. on p. [3201](#)).
- Vladimir I. Danilov, Alexander V. Karzanov, and Gleb A. Koshevoy (2010). “On maximal weakly separated set-systems”. *J. Algebraic Combin.* 32.4, pp. 497–531. MR: [2728757](#) (cit. on p. [3213](#)).
- Miriam Farber and Alexander Postnikov (2016). “Arrangements of equal minors in the positive Grassmannian”. *Adv. Math.* 300, pp. 788–834. MR: [3534845](#) (cit. on p. [3213](#)).
- Sergey Fomin and Andrei Zelevinsky (1999). “Double Bruhat cells and total positivity”. *J. Amer. Math. Soc.* 12.2, pp. 335–380. MR: [1652878](#) (cit. on pp. [3205](#), [3210](#)).
- (2002a). “Cluster algebras. I. Foundations”. *J. Amer. Math. Soc.* 15.2, pp. 497–529. MR: [1887642](#) (cit. on pp. [3200](#), [3212](#)).
- (2002b). “The Laurent phenomenon”. *Adv. in Appl. Math.* 28.2, pp. 119–144. MR: [1888840](#) (cit. on p. [3200](#)).
- (2003). “Cluster algebras. II. Finite type classification”. *Invent. Math.* 154.1, pp. 63–121. MR: [2004457](#) (cit. on pp. [3200](#), [3212](#)).
- (2007). “Cluster algebras. IV. Coefficients”. *Compos. Math.* 143.1, pp. 112–164. MR: [2295199](#) (cit. on pp. [3200](#), [3212](#)).
- P. Galashin, A. Postnikov, and L. Williams (n.d.). “Regular plabic graphs”. In preparation (cit. on p. [3224](#)).

- Pavel Galashin (Nov. 2016). “[Plabic graphs and zonotopal tilings](#)”. arXiv: [1611.00492](#) (cit. on pp. [3201](#), [3221](#), [3225](#), [3226](#)).
- Pavel Galashin, Steven N. Karp, and Thomas Lam (July 2017). “[The totally nonnegative Grassmannian is a ball](#)”. arXiv: [1707.02010](#) (cit. on p. [3205](#)).
- Pavel Galashin and Alexander Postnikov (Aug. 2017). “[Purity and separation for oriented matroids](#)”. arXiv: [1708.01329](#) (cit. on p. [3213](#)).
- F. R. Gantmacher and M. G. Krein (1935). “Sur les matrices oscillatoires”. *C. R. Acad. Sci. Paris* 201, pp. 577–579 (cit. on p. [3200](#)).
- I. M. Gelfand, R. M. Goresky, R. D. MacPherson, and V. V. Serganova (1987). “[Combinatorial geometries, convex polyhedra, and Schubert cells](#)”. *Adv. in Math.* 63.3, pp. 301–316. MR: [877789](#) (cit. on pp. [3199](#), [3200](#), [3203](#), [3206](#)).
- I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky (1994). *[Discriminants, resultants, and multidimensional determinants](#)*. Mathematics: Theory & Applications. Birkhäuser Boston, Inc., Boston, MA, pp. x+523. MR: [1264417](#) (cit. on p. [3219](#)).
- Steven N. Karp and Lauren K. Williams (2017). “The  $m = 1$  amplituhedron and cyclic hyperplane arrangements”. *Sém. Lothar. Combin.* 78B, Art. 20, 12. MR: [3678602](#) (cit. on p. [3227](#)).
- Allen Knutson, Thomas Lam, and David E. Speyer (2013). “[Positroid varieties: juggling and geometry](#)”. *Compos. Math.* 149.10, pp. 1710–1752. MR: [3123307](#) (cit. on p. [3201](#)).
- Yuji Kodama and Lauren Williams (2014). “[KP solitons and total positivity for the Grassmannian](#)”. *Invent. Math.* 198.3, pp. 637–699. MR: [3279534](#) (cit. on pp. [3201](#), [3224](#)).
- Yuji Kodama and Lauren K. Williams (2011). “[KP solitons, total positivity, and cluster algebras](#)”. *Proc. Natl. Acad. Sci. USA* 108.22, pp. 8984–8989. MR: [2813307](#) (cit. on pp. [3201](#), [3224](#)).
- Thomas Lam (2016). “[Totally nonnegative Grassmannian and Grassmann polytopes](#)”. In: *Current developments in mathematics 2014*. Int. Press, Somerville, MA, pp. 51–152. arXiv: [1506.00603](#). MR: [3468251](#) (cit. on pp. [3214](#), [3215](#)).
- Bernard Leclerc and Andrei Zelevinsky (1998). “[Quasicommuting families of quantum Plücker coordinates](#)”. In: *Kirillov’s seminar on representation theory*. Vol. 181. Amer. Math. Soc. Transl. Ser. 2. Amer. Math. Soc., Providence, RI, pp. 85–108. MR: [1618743](#) (cit. on pp. [3212](#), [3213](#)).
- Gaku Liu (2017). “[A counterexample to the extension space conjecture for realizable oriented matroids](#)”. *Sém. Lothar. Combin.* 78B, Art. 31, 7. arXiv: [1606.05033](#). MR: [3678613](#) (cit. on p. [3219](#)).
- G. Lusztig (1990). “[Canonical bases arising from quantized enveloping algebras](#)”. *J. Amer. Math. Soc.* 3.2, pp. 447–498. MR: [1035415](#) (cit. on p. [3200](#)).
- (1992). “[Canonical bases in tensor products](#)”. *Proc. Nat. Acad. Sci. U.S.A.* 89.17, pp. 8177–8179. MR: [1180036](#) (cit. on p. [3200](#)).

- G. Lusztig (1994). “Total positivity in reductive groups”. In: *Lie theory and geometry*. Vol. 123. Progr. Math. Birkhäuser Boston, Boston, MA, pp. 531–568. MR: [1327548](#) (cit. on p. [3200](#)).
- (1998a). “Total positivity in partial flag manifolds”. *Represent. Theory* 2, pp. 70–78. MR: [1606402](#) (cit. on pp. [3200](#), [3203](#)).
- George Lusztig (1993). *Introduction to quantum groups*. Vol. 110. Progress in Mathematics. Birkhäuser Boston, Inc., Boston, MA, pp. xii+341. MR: [1227098](#) (cit. on p. [3200](#)).
- (1998b). “Introduction to total positivity”. In: *Positivity in Lie theory: open problems*. Vol. 26. De Gruyter Exp. Math. de Gruyter, Berlin, pp. 133–145. MR: [1648700](#) (cit. on p. [3200](#)).
- Yu. I. Manin and V. V. Shekhtman (1986). “Higher Bruhat orderings connected with the symmetric group”. *Funktsional. Anal. i Prilozhen.* 20.2, pp. 74–75. MR: [847150](#) (cit. on p. [3220](#)).
- N. E. Mnëv (1988). “The universality theorems on the classification problem of configuration varieties and convex polytopes varieties”. In: *Topology and geometry—Rohlin Seminar*. Vol. 1346. Lecture Notes in Math. Springer, Berlin, pp. 527–543. MR: [970093](#) (cit. on p. [3199](#)).
- Greg Muller and David E. Speyer (2017). “The twist for positroid varieties”. *Proc. Lond. Math. Soc.* (3) 115.5, pp. 1014–1071. arXiv: [1606.08383](#). MR: [3733558](#) (cit. on p. [3213](#)).
- Suho Oh (2011). “Positroids and Schubert matroids”. *J. Combin. Theory Ser. A* 118.8, pp. 2426–2435. MR: [2834184](#) (cit. on p. [3205](#)).
- Suho Oh, Alexander Postnikov, and David E. Speyer (2015). “Weak separation and plabic graphs”. *Proc. Lond. Math. Soc.* (3) 110.3, pp. 721–754. MR: [3342103](#) (cit. on pp. [3201](#), [3212](#), [3213](#), [3225](#)).
- Alexander Postnikov (Sept. 2006). “Total positivity, Grassmannians, and networks”. arXiv: [math/0609764](#) (cit. on pp. [3200](#), [3201](#), [3203–3206](#), [3208](#), [3210](#), [3211](#), [3214](#), [3216](#), [3217](#), [3223](#)).
- Alexander Postnikov, David Speyer, and Lauren Williams (2009). “Matching polytopes, toric geometry, and the totally non-negative Grassmannian”. *J. Algebraic Combin.* 30.2, pp. 173–191. MR: [2525057](#) (cit. on pp. [3201](#), [3204](#), [3214](#)).
- J. Rambau and G. M. Ziegler (1996). “Projections of polytopes and the generalized Baues conjecture”. *Discrete Comput. Geom.* 16.3, pp. 215–237. MR: [1410159](#) (cit. on p. [3219](#)).
- Jörg Rambau (1997). “Triangulations of cyclic polytopes and higher Bruhat orders”. *Mathematika* 44.1, pp. 162–194. MR: [1464385](#) (cit. on p. [3220](#)).
- Jörg Rambau and Francisco Santos (2000). “The generalized Baues problem for cyclic polytopes. I”. *European J. Combin.* 21.1. Combinatorics of polytopes, pp. 65–83. MR: [1737328](#) (cit. on pp. [3217](#), [3220](#), [3221](#)).

- Victor Reiner (1999). “The generalized Baues problem”. In: *New perspectives in algebraic combinatorics (Berkeley, CA, 1996–97)*. Vol. 38. Math. Sci. Res. Inst. Publ. Cambridge Univ. Press, Cambridge, pp. 293–336. MR: [1731820](#) (cit. on pp. [3217](#), [3219](#)).
- Konstanze Rietsch (1999). “An algebraic cell decomposition of the nonnegative part of a flag variety”. *J. Algebra* 213.1, pp. 144–154. MR: [1674668](#) (cit. on p. [3200](#)).
- Konstanze Christina Rietsch (1998). *Total positivity and real flag varieties*. Thesis (Ph.D.)–Massachusetts Institute of Technology. ProQuest LLC, Ann Arbor, MI, (no paging). MR: [2716793](#) (cit. on p. [3200](#)).
- Konstanze Rietsch and Lauren Williams (2010). “Discrete Morse theory for totally non-negative flag varieties”. *Adv. Math.* 223.6, pp. 1855–1884. MR: [2601003](#) (cit. on p. [3205](#)).
- Isac Schoenberg (1930). “Über variationsvermindernde lineare Transformationen”. *Math. Z.* 32.1, pp. 321–328. MR: [1545169](#) (cit. on p. [3200](#)).
- Josh Scott (2005). “Quasi-commuting families of quantum minors”. *J. Algebra* 290.1, pp. 204–220. MR: [2154990](#) (cit. on pp. [3201](#), [3212](#)).
- Joshua S. Scott (2006). “Grassmannians and cluster algebras”. *Proc. London Math. Soc.* (3) 92.2, pp. 345–380. MR: [2205721](#) (cit. on pp. [3201](#), [3212](#)).
- Bernd Sturmfels and Günter M. Ziegler (1993). “Extension spaces of oriented matroids”. *Discrete Comput. Geom.* 10.1, pp. 23–45. MR: [1215321](#) (cit. on pp. [3220](#), [3222](#)).
- A. Postnikov T. Lam (2018). “Polypositroids I”. In preparation (cit. on pp. [3202](#), [3206](#), [3226](#)).
- Kelli Talaska (2008). “A formula for Plücker coordinates associated with a planar network”. *Int. Math. Res. Not. IMRN*, Art. ID rnn 081, 19. MR: [2439562](#) (cit. on pp. [3214](#), [3216](#)).
- V. A. Voevodskiĭ and M. M. Kapranov (1991). “The free  $n$ -category generated by a cube, oriented matroids and higher Bruhat orders”. *Funktsional. Anal. i Prilozhen.* 25.1, pp. 62–65. MR: [1113124](#) (cit. on p. [3220](#)).
- Günter M. Ziegler (1993). “Higher Bruhat orders and cyclic hyperplane arrangements”. *Topology* 32.2, pp. 259–279. MR: [1217068](#) (cit. on pp. [3220](#), [3222](#)).

Received 2018-02-28.

ALEXANDER POSTNIKOV  
DEPARTMENT OF MATHEMATICS  
MASSACHUSETTS INSTITUTE OF TECHNOLOGY  
77 MASSACHUSETTS AVENUE  
CAMBRIDGE MA 02139  
USA  
[apost@math.mit.edu](mailto:apost@math.mit.edu)

# FROM GRAPH LIMITS TO HIGHER ORDER FOURIER ANALYSIS

BALÁZS SZEGEDY

## Abstract

The so-called graph limit theory is an emerging diverse subject at the meeting point of many different areas of mathematics. It enables us to view finite graphs as approximations of often more perfect infinite objects. In this survey paper we tell the story of some of the fundamental ideas in structural limit theories and how these ideas led to a general algebraic approach (the nilspace approach) to higher order Fourier analysis.

## 1 Introduction

Finite objects are often imperfect approximations of much nicer infinite objects. For example the equations of fluid dynamics or thermodynamics are much simpler if we replace discrete particles by continuous mass. If the particle system is large enough then the continuous model behaves sufficiently similarly to the discrete one in many practical applications. This connection between finite and infinite structures is useful in both directions. Passing to infinite limits can greatly simplify messy calculations with finite objects. Various small quantities (epsilon's), that appear as errors in calculations often disappear in the limit. Beyond getting rid of epsilon's there is a deeper advantage of limit theories. Certain algebraic structures, that are present only in approximate forms in finite structures, appear in a precise form when going to the limit. One of the most surprising discoveries in higher order Fourier analysis is that functions on finite Abelian groups can behave as approximations of functions on inherently non-commutative, topological structures such as nilmanifolds.

The goal of this paper is to take the reader to a journey that starts with a general introduction to structural limits and their applications. We use ergodic theory and graph limit theory to demonstrate a number of fundamental concepts including sampling, quasi-randomness, uniformity norms, convergence, limit space and topologization. We devote



a separate chapter to the non-standard approach which is a powerful tool in limit theories. Finally we turn to higher order Fourier analysis where we explain how nilmanifolds and other even more exotic structures come into play when we look at finite additive structures in the limit.

## 2 History and basic concepts

The history of structural limits can be traced back all the way to ancient Greeks. Archimedes (287-212 BC) used polygon approximations of the circle to compute its area. Structural limit theories are routinely used in physics. Continuous limits are essential in thermodynamics and fluid dynamics where large but finite particle systems are investigated. On the other hand discrete approximations of continuous objects such as lattice gauge theory play also an important role in physics.

Many of the above limit theories are based on very simple correspondences between finite objects and continuous limit objects. Most of the time the finite approximation is directly related to a continuous space through a prescribed geometric connection. By somewhat abusing the term, we call such limits *scaling limits*. Much more mysterious and surprising limit theories emerged more recently where simple and very general structures are considered such as 0-1 sequences or graphs. In these theories there is no "prescribed" geometry to be approximated. The geometry emerges from the internal "logic" of the structure and thus a great variety of geometric, topological and algebraic structures can appear in the limit. Many of these limit theories are based on taking small random samples from large structures. We call such limit theories *local* limit theories. Some other limit theories are based on observable, large scale properties and we call them *global* limit theories. Furthermore there are hybrid theories such as the local-global convergence of bounded degree graphs [Hatami, Lovász, and Szegedy \[2014\]](#).

**Scaling limits of 0 – 1 sequences:** As an illustration we start with a rather simple (warm up) limit theory for 0 – 1 sequences. Later we will see a different and much more complicated theory for the same objects. For  $k \in \mathbb{N}$  let  $[k] := \{1, 2, \dots, k\}$ . A 0 – 1 sequence of length  $k$  is a function  $f : [k] \rightarrow \{0, 1\}$ . Assume that we are given a growing sequence  $\{f_n\}_{n=1}^\infty$  of 0 – 1 sequences. *In what sense can we say that these sequences converge?* A simple and natural approach would be to regard the set  $[k]$  as a discretization of the  $(0, 1]$  interval. This way, for a 0 – 1 sequence  $s$  of length  $k$  we can define the function  $\tilde{s} : [0, 1] \rightarrow \{0, 1\}$  by  $\tilde{s}(x) := s(\lceil kx \rceil)$  (and  $\tilde{s}(0) := 0$ ). Now we can replace the functions  $f_n$  by  $\tilde{f}_n$  and use one of the readily available convergence notions for functions on  $[0, 1]$  such as  $L^2$  or  $L^1$  convergence. Note that they are equivalent for 0 – 1 valued functions. The limit object in  $L^2$  is a Lebesgue measurable function  $f : [0, 1] \rightarrow \{0, 1\}$  with the property that the measure of  $f^{-1}(1) \Delta \tilde{f}_n^{-1}(1)$  converges to 0 as  $n$  goes to infinity. A

much more interesting and flexible limit concept is given by the weak convergence in  $L^2([0, 1])$ . For  $0 - 1$  valued functions this is equivalent with the fact that for every interval  $I = [a, b] \subseteq [0, 1]$  the measure of  $I \cap \tilde{f}_n^{-1}(1)$  converges to some quantity  $\mu(I)$  as  $n$  goes to infinity. The limit object is a measurable function  $f : [0, 1] \rightarrow [0, 1]$  with the property that  $\mu(I) = \int_I f d\lambda$  where  $\lambda$  is the Lebesgue measure. If  $\tilde{f}_n$  is  $L^2$  convergent then its weak limit is the same as the  $L^2$  limit. However many more sequences satisfy weak convergence.

*Let  $\{f_n\}_{n=1}^\infty$  be a sequence of  $0 - 1$  sequences. We say that  $f_n$  is **scaling convergent** if  $\{\tilde{f}_n\}_{n=1}^\infty$  is a weakly convergent sequence of functions in  $L^2([0, 1])$ . The limit object (scaling limit) is a measurable function of the form  $f : [0, 1] \rightarrow [0, 1]$ .*

Although scaling convergence is a rather simplistic limit notion we can use it as a toy example to illustrate some of the fundamental concepts that appear in other, more interesting limit theories.

- **Compactness:** Every sequence of  $0 - 1$  sequences has a scaling convergent subsequence
- **Uniformity norm:** Scaling convergence can be metrized through norms. An example for such a norm is the "intervall norm" defined by  $\|f\|_{\text{in}} := \sup_I |\int_I f d\lambda|$ , where  $I$  runs through all intervals in  $(0, 1]$ . The distance of two  $0 - 1$  sequences  $f_1$  and  $f_2$  (not necessarily of equal length) is defined as  $\|\tilde{f}_1 - \tilde{f}_2\|_{\text{in}}$ .
- **Quasi randomness:** A  $0 - 1$  sequence  $f$  is  $\epsilon$ -quasi random with density  $p \in [0, 1]$  if  $\|\tilde{f} - p\|_{\text{in}} \leq \epsilon$ . Note that if  $f_n$  is a sequence of  $0 - 1$  sequences such that  $f_n$  is  $\epsilon_n$  quasi random with density  $p$  and  $\epsilon_n$  goes to 0 then  $f_n$  converges to the constant  $p$  function.
- **Random objects are quasi random:** Let  $f_n$  be a random  $0 - 1$  sequence of length  $n$  in which the probability of 1 is  $p$ . For an arbitrary  $\epsilon > 0$  we have that if  $n$  is large enough then with probability arbitrarily close to 1 the function  $f_n$  is  $\epsilon$  quasi random.
- **Low complexity approximation (regularization):** For every  $\epsilon > 0$  there is some natural number  $N_\epsilon$  such that for every  $0 - 1$  sequence  $f$  there is a function  $g : [N_\epsilon] \rightarrow [0, 1]$  such that  $\|\tilde{f} - \tilde{g}\|_{\text{in}} \leq \epsilon$ . (Note  $\tilde{g}$  is defined by the same formula as for  $0 - 1$  sequences and  $\tilde{g}$  is a step function on  $[0, 1]$  with  $N_\epsilon$  steps.)

**Local limits of  $0 - 1$  sequences:** The main problem with scaling convergence is that highly structured sequences such as periodic sequences like  $0, 1, 0, 1, 0, 1, \dots$  are viewed

as quasi random. The above limit concept is based on a prescribed geometric correspondence between integer intervals and the continuous  $[0, 1]$  interval. A different and much more useful limit concept does not assume any prescribed geometry. It is based on the local statistical properties of  $0 - 1$  sequences. For any given  $0 - 1$  sequence  $h$  of length  $k$  and  $f$  of length  $n \geq k$  we define  $t(h, f)$  to be the probability that randomly chosen  $k$  consecutive bits in  $f$  are identical to the sequence  $h$  (if  $n < k$  then we simply define  $t(h, f)$  to be 0).

A sequence  $\{f_n\}_{n=1}^{\infty}$  of growing  $0 - 1$  sequences is called **locally convergent** if for every  $0 - 1$  sequence  $h$  we have that  $\lim_{n \rightarrow \infty} t(h, f_n)$  exists.

This definition was first used by [Furstenberg \[1977\]](#) in his famous correspondence principle stated in the 70's, a major inspiration for all modern limit theories. In Furstenberg's approach finite  $0 - 1$  sequences are regarded as approximations of subsets in certain dynamical systems called measure preserving systems. A measure preserving system is a probability space  $(\Omega, \mathfrak{B}, \mu)$  together with a measurable transformation  $T : \Omega \rightarrow \Omega$  with the property that  $\mu(T^{-1}(A)) = \mu(A)$  for every  $A \in \mathfrak{B}$ .

**Furstenberg's correspondence principle for  $\mathbb{Z}$ :** Let  $f_n$  be a locally convergent sequence of  $0 - 1$  sequences. Then there is a measure preserving system  $(\Omega, \mathfrak{B}, \mu, T)$  and a measurable set  $S \subseteq \Omega$  such that for every  $0 - 1$  sequence  $h : [k] \rightarrow \{0, 1\}$  the quantity  $\lim_{n \rightarrow \infty} t(h, f_n)$  is equal to the probability that  $(1_S(x), 1_S(x^T), \dots, 1_S(x^{T^{k-1}})) = h$  for a random element  $x \in \Omega$ .

Note that originally the correspondence principle was stated in a different and more general form for amenable groups. If the group is  $\mathbb{Z}$  then it is basically equivalent with the above statement. A measure preserving system is called **ergodic** if there is no set  $A \in \mathfrak{Q}$  such that  $0 < \mu(A) < 1$  and  $\mu(A \Delta T^{-1}(A)) = 0$ . Every measure preserving system is the combination of ergodic ones and thus ergodic measure preserving systems are the building blocks of this theory.

We give two examples for convergent  $0 - 1$  sequences and their limits. Let  $\alpha$  be a fixed irrational number. Then, as  $n$  tends to infinity, the sequences  $1_{[0, 1/2]}(\{\alpha i\}), i = 1, 2, \dots, n$  (where  $\{x\}$  denotes the fractional part of  $x$ ) approximate the semicircle in a dynamical system where the circle is rotated by  $2\pi\alpha$  degrees. Both the circle and the semicircle appears in the limit. A much more surprising example (in a slightly different form) is given by [Host and Kra \[2008a\]](#). Let us take two  $\mathbb{Q}$ -independent irrational numbers  $\alpha, \beta$  and let  $a_i := 1_{[0, 1/2]}(\{[i\beta]\alpha - i(i-1)\alpha\beta/2\})$  where  $[x]$  denotes the integer part of  $x$ . In this case the limiting dynamical system is defined on a three dimensional compact manifold called *Heisenberg nilmanifold*.

**Topologization and algebraization:** At this point it is important to mention that Furstenberg's correspondence principle does not immediately give a "natural" topological representation of the limiting measure preserving system. In fact the proof yields a system in

which the ground space is the compact set  $\{0, 1\}^{\mathbb{Z}}$  with the Borel  $\sigma$ -algebra,  $T$  is the shift of coordinates by one and  $\mu$  is some shift invariant measure. The notion of isomorphism between systems allows us to switch  $\{0, 1\}^{\mathbb{Z}}$  to any other standard Borel space. However in certain classes of systems it is possible to define a "nicest" or "most natural" topology. An old example for such a topologization is given by Kronecker systems [Furstenberg \[1981\]](#). Assume that the measure preserving map  $T$  is ergodic and it has the property that  $L^2(\Omega)$  is generated by the eigenvectors of the induced action of  $T$  on  $L^2(\Omega)$ . It turns out that such systems can be represented as rotations in compact abelian groups (called Kronecker systems). The problem of topologization is a recurring topic in limit theories. It often comes together with some form of "algebraization" in the frame of which the unique nicest topology is used to identify an underlying algebraic structure that is intimately tied to the dynamics. Again this can be demonstrated on Kronecker systems where finding the right topology helps in identifying the Abelian group structure. Note that there is a highly successful and beautiful story of topologization and algebraization in ergodic theory in which certain factor-systems of arbitrary measure preserving systems (called characteristic factors) are identified as inverse limits of geometric objects (called nilmanifolds) arising from nilpotent Lie groups [Host and Kra \[2005\]](#), [Ziegler \[2007\]](#). As this breakthrough was also crucial in the development of higher order Fourier analysis we will give more details in the next paragraph. In many limit theories the following general scheme appears.

**discrete objects  $\rightarrow$  measurable objects  $\rightarrow$  topological objects  $\rightarrow$  algebraic objects**

*The first arrow denotes the limit theory, the second arrow denotes topologization and the third arrow is the algebraization.*

**Factors:** Factor systems play a crucial role in ergodic theory. A factor of a measure preserving system  $(\Omega, \mathfrak{B}, \mu, T)$  is a sub  $\sigma$ -algebra  $\mathfrak{F}$  in  $\mathfrak{B}$  that is  $T$  invariant (if  $B \in \mathfrak{F}$  then  $T^{-1}(B) \in \mathfrak{F}$ ). Note that if  $\mathfrak{F}$  is a factor then  $(\Omega, \mathfrak{F}, \mu, T)$  is also a measure preserving system. Often there is a duality between a system of "observable quantities" defined through averages and certain factors, called *characteristic factors*. For example the averages

$$t(f) := \lim_{n \rightarrow \infty} n^{-1} \sum_{i=1}^n \int_{\Omega} f(x) f(T^i(x)) f(T^{2i}(x)) d\mu$$

defined for bounded measurable functions satisfy that  $t(f) = t(\mathbb{E}(f|\mathcal{K}))$  where  $\mathcal{K}$  is the Kronecker factor of the system (the unique largest factor that is a Kronecker system) and  $\mathbb{E}(f|\mathcal{K})$  is the conditional expectation with respect to  $\mathcal{K}$ . Since conditional expectation is an elementary operation, this means that properties of  $t(f)$  can be completely described in terms of Kronecker systems. The ergodic theoretic proof [Furstenberg \[1977\]](#) of Roth theorem [Roth \[1953\]](#) on 3-term arithmetic progressions is based on this fact and a limiting

argument using Furstenberg's correspondence principle. It turns out that in every ergodic measure preserving system there is a sequence of increasing, uniquely defined factors  $\mathcal{K}_1 \leq \mathcal{K}_2 \leq \dots$  which starts with the Kronecker factor. Similarly to Roth's theorem, the study of  $k$ -term arithmetic progressions can be reduced to  $\mathcal{K}_{k-2}$ . The results in [Host and Kra \[2005\]](#) and [Ziegler \[2007\]](#) give a complete geometric description for these factors in terms of nilsystems. Let  $G$  be a  $k$ -step nilpotent Lie group and  $\Lambda \leq G$  be a co-compact subgroup. The space  $N = \{g\Lambda : g \in G\}$  of left cosets of  $\Lambda$  is a finite dimensional compact manifold on which  $G$  acts by left multiplication. It is known that there is a unique  $G$  invariant probability measure  $\mu$  on  $N$ . We have that  $\{N, \mathfrak{B}, \mu, g\}$  is a measure preserving system for every  $g \in G$  (where  $\mathfrak{B}$  is the Borel  $\sigma$ -algebra). If  $g$  acts in an ergodic way then it is called a  $k$ -step nilsystem. It was proved in [Host and Kra \[2005\]](#) and [Ziegler \[2007\]](#) that for every  $k$  the factor  $\mathcal{K}_k$  of an ergodic system is the inverse limits of  $k$ -step nilsystems.

**Local and global limits of graphs:** Although Furstenberg's correspondence principle gives the first example for a local limit theory, a systematic study of similar structural limit theories started much later. The general program of studying structures in the limit became popular in the early 2000's when graph limit theory was born [Benjamini and Schramm \[2001\]](#), [Lovász and Szegedy \[2006\]](#), [Lovász and Szegedy \[2007\]](#), [Borgs, J. Chayes, Lovász, Sós, Szegedy, and Vesztergombi \[2006\]](#), [Borgs, J. T. Chayes, Lovász, Sós, and Vesztergombi \[2008\]](#), [Borgs, J. T. Chayes, Lovász, Sós, and Vesztergombi \[2012\]](#). The motivation to develop an analytic theory for large networks came partially from applied mathematics. The growing access to large networks such as social networks, internet graphs and biological networks like the brain generated a demand for new mathematical tools to understand their approximate structure. Another motivation came from extremal combinatorics where inequalities between subgraph densities are extensively studied. An analytic view of graphs enables the use of powerful methods such as differential calculus to solve extremal problems. Similarly to ergodic theory certain graph sequences approximate infinite structures which can not be perfectly represented by finite objects. It turns out that there are simple extremal problems for graphs which have no precise finite solutions but a nice exact solution appears in the limit. This is somewhat similar to the situation with the inequality  $(x^2 - 2)^2 \geq 0$  which has no precise solution in  $\mathbb{Q}$  but it has two solutions in  $\mathbb{R}$ .

Similarly to 0 – 1 sequences graph convergence can be defined through converging sample distributions and thus the convergence notion will depend on the sampling method. Quite surprisingly there are two different natural sampling methods. The first one works well if the graph has a non negligible edge density (such graphs are called *dense*) and the second one is defined only for bounded degree graphs. Note that on  $n$  vertices a dense graph has  $cn^2$  edges for some non negligible  $c > 0$  whereas a bounded degree graph has

$cn$  edges for some bounded  $c$ . This means that dense and bounded degree graphs are at the two opposite ends of the density spectrum. If a graph is neither dense nor bounded degree then we call it *intermediate*.

Let  $G = (V, E)$  be a finite graph. In the first sampling method we choose  $k$  vertices  $v_1, v_2, \dots, v_k$  independently and uniformly from  $V$  and take the graph  $G_k$  spanned on these vertices. We regard  $G_k$  as a random graph on  $[k]$ . For a graph  $H$  on the vertex set  $[k]$  let  $t^0(H, G)$  denote the probability that  $G_k = H$ . In dense graph limit theory, a graph sequence  $\{G_n\}_{n=1}^\infty$  is called convergent if for every fixed graph  $H$  the limit  $\lim_{i \rightarrow \infty} t^0(H, G)$  exists. Another equivalent approach is to define  $t(H, G)$  as the probability that a random map from  $V(H)$  to  $V(G)$  is a *graph homomorphism* i.e. it takes every edge of  $H$  to an edge of  $G$ . This number is called the homomorphism density of  $H$  in  $G$ . In a sequence  $\{G_n\}_{n=1}^\infty$ , the convergence of  $t(H, G_n)$  for all graphs  $H$  is equivalent with the convergence of  $t^0(H, G_n)$  for all graphs  $H$ . The advantage of using homomorphism densities is that they have nicer algebraic properties such as multiplicativity and reflection positivity Lovász and Szegedy [2006].

For the second sampling method let  $\mathcal{G}_d$  denote the set of finite graphs with maximum degree at most  $d$ . Let furthermore  $\mathcal{G}_d^r$  denote the set of graphs of maximum degree at most  $d$  with a distinguished vertex  $o$  called the *root* such that every other vertex is of distance at most  $r$  from  $o$ . Now if  $G = (V, E)$  is in  $\mathcal{G}_d$  then let  $v$  be a uniform random vertex in  $V$ . Let  $N_r(v)$  denote the  $v$ -rooted isomorphism class of the radius  $r$ -neighborhood of  $v$  in  $G$ . We have that  $N_r(v)$  is an element in  $\mathcal{G}_d^r$  and thus the random choice of  $v$  imposes a probability distribution  $\mu(r, G)$  on  $\mathcal{G}_d^r$ . A graph sequence  $\{G_n\}_{n=1}^\infty$  is called Benjamini-Schramm convergent if  $\mu(r, G_n)$  is convergent in distribution for every fixed  $r$ . The convergence notion was introduced in the paper Benjamini and Schramm [2001] to study random walks on planar graphs. Colored and directed versions of this convergence notion can be also introduced in a similar way. Benjamini-Schramm convergence provides a rather general framework for many different problems. Note that it generalizes the local convergence of  $0-1$  sequences because one can represent finite  $0-1$  sequences by directed paths with 0 and 1 labels on the nodes. Limit objects for Benjamini-Schramm convergent sequences are probability distributions on infinite rooted graphs with a certain measure preserving property that generalizes the concept of measure preserving system. Note that Benjamini-Schramm convergence is closely related to group theory. A finitely presented group is called Sofic if its Cayley graph is the limit of finite graphs in which the edges are directed and labeled by the generators of the group. Sofic groups are much better understood than general abstract groups. The study of sofic groups is a fruitful interplay between graph limit theory and group theory.

Global aspects of graph limit theory arise both in the dense and the bounded degree frameworks. In case of dense graph limit theory the local point of view is often not strong

enough. Although it turns out that one can represent convergent sequences by so-called graphons [Lovász and Szegedy \[2006\]](#) i.e. symmetric measurable functions of the form  $W : [0, 1]^2 \rightarrow [0, 1]$  a stronger theorem that connects the convergence with Szemerédi's regularity lemma is more useful. Szemerédi's famous regularity lemma is a structure theorem describing the large scale structure of graphs in terms of quasi random parts. A basic compactness result in dense graph limit theory [Lovász and Szegedy \[2007\]](#) (see also [Theorem 1](#)) connects the local and global point of views. This is used in many applications including property testing [Lovász and Szegedy \[2010b\]](#) and large deviation principles [Chatterjee and Varadhan \[2011\]](#).

The Benjamini-Schramm convergence is inherently a local convergence notion and thus it is not strong enough for many applications. For example random  $d$ -regular graphs are locally tree-like but they have a highly non-trivial global structure that has not been completely described. To formalize this problem one needs a refinement of Benjamini-Schramm convergence called local-global convergence [Hatami, Lovász, and Szegedy \[2014\]](#). The concept of local-global convergence was successfully used in the study of eigenvectors of random regular graphs. It was proved by complicated analytic, information theoretic and graph limit methods in [Backhausz and Szegedy \[2016\]](#) that almost eigenvectors of random regular graphs have a near Gaussian entry distribution. This serves as an illustrative example for the fact that deep results in graph theory can be obtained through the limit approach.

We have to mention that the branch of graph limit theory that deals with intermediate graphs (between dense and bounded degree) is rather underdeveloped. There are numerous competing candidates for an intermediate limit theory [Borgs, J. T. Chayes, Cohn, and Zhao \[n.d.\]](#), [Borgs, J. T. Chayes, Cohn, and Zhao \[2018\]](#), [Szegedy \[n.d.\(c\)\]](#), [Kunszenti-Kovács, Lovász, and Szegedy \[2016\]](#), [Nesetril and Mendez \[2013\]](#), [Frenkel \[2018\]](#) but they have very few applications so far. The hope is that at least one of these approaches will become a useful tool to study real life networks such as connections in the brain or social networks. These networks are typically of intermediate type.

**Limits in additive combinatorics and higher order Fourier analysis:** Let  $A$  be a finite Abelian group and  $S$  be a subset in  $A$ . Many questions in additive combinatorics deal with the approximate structure of  $S$ . For example Szemerédi's theorem can be interpreted as a result about the density of arithmetic progressions of subsets in cyclic groups. It turns out that limit approaches are natural in this subject. Let  $M \in \mathbb{Z}^{m \times n}$  be an integer matrix such that each element in  $M$  is coprime to the order of  $A$ . Then we can define the density of  $M$  in the pair  $(A, S)$  as the probability that  $\sum M_{i,j} x_j \in S$  holds for every  $i$  with random uniform independent choice of elements  $x_1, x_2, \dots, x_n \in A$ . For example the density of 3 term arithmetic progressions in  $S$  is the density of the matrix  $((1, 0), (1, 1), (1, 2))$  in  $S$ . We say that a sequence  $\{(A_i, S_i)\}_{i=1}^\infty$  is convergence if the density of all coprime

matrix  $M$  in the elements of the sequence converges. This type of convergence was first investigated in [Szegedy \[n.d.\(b\)\]](#) and limit objects were also constructed. The subject is deeply connected to Gowers norms and the subject of higher order Fourier analysis. We give more details in chapter 5.

### 3 Dense graph limit theory

A *graphon* is a measurable function of the form  $W : [0, 1]^2 \rightarrow [0, 1]$  with the property  $W(x, y) = W(y, x)$  for every  $x, y \in [0, 1]$ . Let  $\mathcal{W}$  denote the set of all graphons. If  $G$  is a finite graph on the vertex set  $[n]$  then its graphon representation  $W_G$  is defined by the formula

$$W(x, y) = 1_{E(G)}(\lceil nx \rceil, \lceil ny \rceil).$$

For a graph  $H$  on the vertex set  $[k]$  let

$$t(H, W) := \int_{x_1, x_2, \dots, x_k \in [0, 1]} \prod_{(i, j) \in E(H)} W(x_i, x_j) dx_1 dx_2 \dots dx_k.$$

The quantity  $t(H, W)$  is an analytic generalization of the so called homomorphism density defined for finite graphs. This is justified by the easy observation that  $t(H, G) = t(H, W_G)$ . We will need the so-called cut norm  $\|\cdot\|_{\square}$  on  $L^\infty([0, 1]^2)$ . Let  $F : [0, 1]^2 \rightarrow \mathbb{R}$  be a bounded measurable function. Then

$$\|F\|_{\square} := \sup_{A, B \subseteq [0, 1]} \left| \int_{A \times B} F(x, y) dx dy \right|$$

where  $A$  and  $B$  run through all measurable sets in  $[0, 1]$ . Using this norm we can introduce a measure for "similarity" of two graphons  $U$  and  $W$  by  $\|U - W\|_{\square}$ . However this is not the similarity notion that we use for convergence. We need to factor out by graphon isomorphisms. If  $\psi : [0, 1] \rightarrow [0, 1]$  is a measure preserving transformation the we define  $W^\psi(x, y) := W(\psi(x), \psi(y))$ . It is easy to check that this transformation on graphons preserves the homomorphism densities:  $t(H, W) = t(H, W^\psi)$  holds for every finite graph  $H$ . The next distance was introduced in [Lovász and Szegedy \[2007\]](#) :

$$\delta_{\square}(U, W) := \inf_{\phi, \psi : [0, 1] \rightarrow [0, 1]} \|U^\phi - W^\psi\|_{\square}$$

where  $\phi$  and  $\psi$  are measure preserving transformations. It is easy to check that  $\delta_{\square}$  is a pseudometrics i.e. it satisfies all axioms except that  $d(x, y) = 0$  does not necessarily imply that  $x = y$ . In order to get an actual metrics we have to factor out by the equivalence relation  $\sim_{\delta_{\square}}$  defined by  $x \sim_{\delta_{\square}} y \Leftrightarrow d(x, y) = 0$ . Let  $\mathfrak{X} := \mathcal{W} / \sim_{\delta_{\square}}$ . Since



$\delta_{\square}(U, W) = 0$  implies that  $t(H, U) = t(H, W)$  holds for every graph  $H$  we have that  $t(H, -)$  is well defined on  $\mathfrak{X}$ . The following result [Lovász and Szegedy \[2006\]](#), [Lovász and Szegedy \[2007\]](#) in graph limit theory is fundamental in many applications.

**Theorem 1.** *We have the following statements for the metric space  $(\mathfrak{X}, \delta_{\square})$ .*

1. *The metric  $\delta_{\square}$  defines a compact, Hausdorff, second countable topology on  $\mathfrak{X}$ .*
2. *The function  $X \rightarrow t(H, X)$  is a continuous function on  $\mathfrak{X}$  for every finite graph  $H$ .*

Two important corollaries are the following.

**Corollary 3.1.** *Assume that  $\{G_i\}_{i=1}^{\infty}$  is a sequence of graphs such that  $f(H) := \lim_{i \rightarrow \infty} t(H, G_i)$  exists for every finite graph  $H$ . Then there is a graphon  $W \in \mathfrak{W}$  such that  $f(H) = t(H, W)$  holds for every  $H$ .*

**Corollary 3.2.** *Szemerédi's regularity lemma [Szemerédi \[1978\]](#) (even in stronger forms) follows from [Theorem 1](#).*

Note that, although [Corollary 3.1](#) may be deduced from earlier results on exchangeability [Aldous \[1985\]](#), [Theorem 1](#) combines both the local and global aspects of convergence and so it is a stronger statement. In some sense it can be regarded as a common generalization of both Szemerédi's regularity lemma [Szemerédi \[1978\]](#) and a result on exchangeability [Aldous \[1985\]](#).

**Topologization of graph limit theory:** In the definition of a graphon  $W : [0, 1]^2 \rightarrow [0, 1]$  the  $[0, 1]$  interval on the left hand side is replaceable by any standard probability space  $(\Omega, \mu)$ . In general we need that  $(\Omega, \mu)$  is atomless but for certain special graphons even atoms maybe allowed. Note that the values of  $W$  represent probabilities and so the  $[0, 1]$  interval is crucial on the right hand side. Thus the general form of a graphon is a symmetric measurable function  $W : \Omega \times \Omega \rightarrow [0, 1]$ . Homomorphism densities  $t(H, W)$  are defined for all such general graphons and two of them are equivalent if all homomorphism densities are the same. The following question arises: *Given a graphon  $W$ . Is there a most natural topological space  $X$  and Borel measure  $\mu$  on  $X$  such that  $W$  is equivalent with a graphon of the form  $W' : X^2 \rightarrow [0, 1]$ ?* An answer to this question was given in [Lovász and Szegedy \[2010a\]](#). For a general graphon  $W : \Omega \times \Omega \rightarrow [0, 1]$  there is a unique purified version of  $W$  on some Polish space  $X$  with various useful properties. The language of topologization induced a line of exciting research in extremal combinatorics. Here we give a brief overview on applications of graphons in extremal graph theory.

**Extremal graphs and graphons:** The study of inequalities between subgraph densities and the structure of extremal graphs is an old topic in extremal combinatorics. A classical example is Mantel's theorem which implies that a triangle free graph  $H$  on  $2n$  vertices

maximizes the number of edges if  $H$  is the complete bipartite graph with equal color classes. Another example is given by the Chung-Graham-Wilson theorem [Chung, Graham, and Wilson \[1989\]](#). If we wish to minimize the density of the four cycle in a graph  $H$  with edge density  $1/2$  then  $H$  has to be sufficiently quasi random. However the perfect minimum of the problem (that is  $1/16$ ) can not be attained by any finite graph but one can get arbitrarily close to it. Both statements can be conveniently formulated in the framework of dense graph limit theory. In the first one we maximize  $t(e, G)$  in a graph  $G$  with the restriction that  $t(C_3, G) = 0$  (where  $e$  is the edge and  $C_3$  is the triangle). In the second one we fix  $t(e, G)$  to be  $1/2$  and we minimize  $t(C_4, G)$ . Since the graphon space is the completion of the space of graphs it is very natural to investigate these problems in a way that we replace  $G$  by a graphon  $W$ . If we fix finite graphs  $H_1, H_2, \dots, H_k$  then all possible inequalities between  $t(H_1, W), t(H_2, W), \dots, t(H_k, W)$  are encoded in the  $k$ -dimensional point set

$$\mathcal{L}(H_1, H_2, \dots, H_k) := \{(t(H_1, W), t(H_2, W), \dots, t(H_k, W)) : W \in \mathcal{W}\}.$$

Note that this is a closed subset in  $[0, 1]^k$ . As an example let  $e$  be a single edge and let  $P_2$  denote the path with two edges. It is easy to prove that  $t(P_2, W) \geq t(e, W)^2$ . This inequality is encoded in  $\mathcal{L}(e, P_2)$  in the form that  $\mathcal{L}(e, P_2) \subseteq \{(x, y) : y \geq x^2\}$ . We have however that  $\mathcal{L}(e, P_2)$  carries much more information. The shape of  $\mathcal{L}(H_1, H_2, \dots, H_k)$  is known in very few instances. It took decades of research to completely describe the two dimensional shape  $\mathcal{L}(e, C_3)$  which gives all possible inequalities between  $t(e, W)$  and  $t(C_3, W)$ . The characterization of  $\mathcal{L}(e, C_3)$  was completed by [Razborov \[2008\]](#) partially using limit methods (a certain differentiation on the graph limit space). Another direction of research investigates the structure of a graphon  $W$  with given subgraph densities. A graphon  $W$  is called *finitely forcible* [Lovász and Szegedy \[2011\]](#) if there are finitely many graphs  $H_1, H_2, \dots, H_k$  such that if  $t(H_i, W') = t(H_i, W)$  holds for  $i = 1, 2, \dots, k$  for some  $W' \in \mathcal{W}$  then  $W'$  is equivalent with  $W$ . The motivation to study finitely forcible graphons is that they represent a large family of extremal problems with unique solution. It is very natural to ask how complicated can extremal graph theory get at the structural level. Originally it was conjectured that finitely forcible graphons admit a step function structure which is equivalent with the fact that the topologization of the graphon is a finite space. This was disproved in [Lovász and Szegedy \[ibid.\]](#) and various examples were given with more interesting underlying topology. However the topology in all of these examples is compact and finite dimensional. It was asked in [Lovász and Szegedy \[ibid.\]](#) whether this is always the case. Quite surprisingly both conjectures turned out to be false. Extremal problems with strikingly complicated topologies were constructed in [Glebov, Klimosova, and Kral \[2014\]](#), [Cooper, Kaiser, Noel, et al. \[2015\]](#). This gives a very strong justification of graph limit theory in extremal combinatorics by showing that complicated infinite structures are somehow encoded into finite looking problems. The marriage between extremal

graph theory and graph limit theory has turned into a growing subject with surprising results. It brought topology and analysis into graph theory and gave a deep insight into the nature and complexity of extremal structures.

## 4 The ultra limit method

The use of ultra products in structural limit theory Elek and Szegedy [2012], Szegedy [n.d.(b)] was partially motivated by the great complexity of the proofs in hypergraph regularity theory Nagle, Rödl, and Schacht [2006], Rödl and Schacht [2007], Rödl and Skokan [2004] and later in higher order Fourier analysis. For example the hypergraph removal lemma is a simple to state and beautiful theorem, but its combinatorial proofs are extremely complicated. This may not be surprising in the light of the fact that it implies Szemerédi's famous theorem Szemerédi [1975] on arithmetic progressions even in a multi-dimensional form Solymosi [2004]. However it was observed in Elek and Szegedy [2012] that great simplification can be made to these proofs if one works in a limiting setting. The limit theory which is particularly useful here is based on ultra products of measure spaces. Without going to technical details we give an overview of a scheme that was successfully used in hypergraph theory Elek and Szegedy [ibid.] and additive combinatorics Szegedy [n.d.(b)]. *This scheme is based on the philosophy that if there is any "reasonable" limit of a sequence of structures  $S_1, S_2, \dots$  then it has to appear somehow on the ultra product space  $\mathbf{S} := \prod_{\omega} S_i$  where  $\omega$  is a non-principal ultra filter. Usually the limit object appears as a factor space of  $\mathbf{S}$  endowed with some structure obtained from  $\mathbf{S}$ .* In the followings we give a strategy that unifies some of the applications of the ultra limit method without aiming for full generality.

### Introducing a limit theory:

1. **Structures:** Let  $\mathcal{F}$  be a family of structures (for example finite graphs, hypergraphs or subsets in finite or more generally compact Abelian groups).
2. **Function representation:** Represent each element  $F \in \mathcal{F}$  as a function on some simpler structure  $Q \in \mathcal{Q}$ . We assume that each structure  $Q \in \mathcal{Q}$  is equipped with a probability measure  $\mu_Q$ . Let  $\mathcal{R}$  denote the representation function  $\mathcal{R} : \mathcal{F} \mapsto \bigcup_{Q \in \mathcal{Q}} L^{\infty}(Q, \mu_Q)$ .

For example, in case of graphs,  $\mathcal{Q}$  is the family of finite product sets of the form  $V \times V$  and  $\mu_{V \times V}$  is the uniform measure. The representation function for a graph  $G = (V, E)$  is given by  $\mathcal{R}(G) := 1_E$  on the product set  $V \times V$ . In other words  $\mathcal{R}(G)$  is the adjacency matrix of  $G$ . For  $k$  uniform hypergraphs  $\mathcal{Q}$  is the set of power sets of the form  $V^k$ . If  $\mathcal{F}$  is the set of measurable subsets in compact Abelian groups then  $\mathcal{Q}$  is the set of compact Abelian groups with the Haar measure.

3. **Moments:** Define a set of moments  $\mathfrak{M}$  such that for each  $m \in \mathfrak{M}$  and  $Q \in \mathcal{Q}$  there is a functional  $m : L^\infty(Q, \mu_Q) \rightarrow \mathbb{C}$ .

Note that the name "moment" refers to the fact that elements in  $L^\infty(Q, \mu)$  are random variables. Here we use generalizations of classical moments which make use of the underlying structure of  $Q$ . For example if  $\mathcal{F}$  is the family of finite graphs, then  $\mathfrak{M}$  is the set of finite directed graphs. Each finite directed graph  $H = ([k], F)$  defines a moment by  $t(H, W) := \mathbb{E}_{x_1, x_2, \dots, x_k \in V} (\prod_{(i,j) \in F} W(x_i, x_j))$  for an arbitrary function  $W : V \times V \rightarrow \mathbb{C}$ . (If  $W$  is symmetric then  $t(H, W)$  does not depend on the direction on  $H$ .) Note that if we allow multiple edges in  $H$  then the  $n$ -fold single edge corresponds to the  $n$ -th classical moment  $\mathbb{E}(W^n)$ . For functions on Abelian groups, moments are densities of additive patterns. For example if  $f : A \rightarrow \mathbb{C}$  then the 3 term arithmetic progression density in  $f$  is defined by  $\mathbb{E}_{x,t} f(x)f(x+t)f(x+2t)$ . Similarly the parallelogram density is defined by  $\mathbb{E}_{x,t_1,t_2} f(x)f(x+t_1)f(x+t_2)f(x+t_1+t_2)$ .

4. **Convergence:** Define a limit notion in  $\mathcal{F}$  in the following way. A sequence  $\{F_i\}_{i=1}^\infty$  in  $\mathcal{F}$  is called convergent if for every  $m \in \mathfrak{M}$  we have that  $\lim_{i \rightarrow \infty} m(\mathcal{R}(F_i))$  exists.

Note that this convergence notion naturally extends to functions on structures in  $\mathcal{Q}$ . This allows us to define convergent sequences of matrices, multidimensional arrays (functions on product sets) or functions on Abelian groups.

5. **Quasi randomness and similarity** Define a norm  $\| \cdot \|_U$  on each function space  $L^\infty(Q, \mu_Q)$  that measures quasi randomness such that if  $\|f\|_U$  is close to 0 then  $f$  is considered to be quasi random. We need the property that for every  $m \in \mathfrak{M}$  and  $\epsilon > 0$  there is  $\delta > 0$  such that if  $f, g \in L^\infty(Q, \mu_Q)$  satisfy  $|f|, |g| \leq 1$  and  $\|f - g\|_U \leq \delta$  then  $|m(f) - m(g)| \leq \epsilon$ .

In case of graphs we can use the four cycle norm  $\|f\|_U := t(C_4, f)^{1/4}$  or an appropriately normalized version of the cut norm. For hypergraphs we can use the so-called octahedral norms. On Abelian groups we typically work with one of the Gowers norms [Gowers \[2001\]](#), [Gowers \[1998\]](#) depending on the set of moments we need to control.

## The ultra limit method:

1. **The ultra limit space** Let  $\mathbf{Q}$  denote the ultra product of some sequence  $\{Q_i\}_{i=1}^\infty$  in  $\mathcal{Q}$ . There is an ultra product  $\sigma$ -algebra  $\mathcal{Q}$  and an ultra product measure  $\mu$  on  $\mathbf{Q}$  that comes from the measure space structures on  $Q_i$  by a known construction. Each uniformly bounded function system  $\{f_i \in L^\infty(Q_i)\}_{i=1}^\infty$  has an ultra limit function  $f \in L^\infty(\mathbf{Q}, \mathcal{Q}, \mu)$  and each function  $g \in L^\infty(\mathbf{Q}, \mathcal{Q}, \mu)$  arises this way

(up to 0 measure change). We can also lift the moments in  $\mathfrak{M}$  and the norm  $\|\cdot\|_U$  to  $L^\infty(\mathbf{Q}, \mathfrak{Q}, \mu)$ . Note that  $\|\cdot\|_U$  usually becomes a seminorm on  $L^\infty(\mathbf{Q}, \mathfrak{Q}, \mu)$  and thus it can take 0 value.

2. **Characteristic factors of the ultra limit space:** Similarly to ergodic theory our goal here is to identify a sub  $\sigma$ -algebra  $\mathfrak{Q}_U$  in  $\mathfrak{Q}$  such that  $\|\cdot\|_U$  is a norm on  $L^\infty(\mathbf{Q}, \mathfrak{Q}_U, \mu)$  and  $\|f - \mathbb{E}(f|\mathfrak{Q}_U)\|_U = 0$  holds for every  $f \in L^\infty(\mathbf{Q}, \mathfrak{Q}, \mu)$ . With this property we also obtain that  $m(f) = m(\mathbb{E}(f|\mathfrak{Q}_U))$  and  $m(f - \mathbb{E}(f|\mathfrak{Q}_U)) = 0$  holds.

These equations imply that the information on the values of the moments is completely encoded in the projection to  $\mathfrak{Q}_U$ . Once we identify this  $\sigma$ -algebra, the goal is to understand its structure. Note that  $\mathfrak{Q}_U$  is a huge non-separable  $\sigma$ -algebra. The next step is to reduce it to a separable factor.

3. **Separable realization:** Let us fix a function in  $f \in L^\infty(\mathbf{Q}, \mathfrak{Q}_U, \mu)$ . Our goal here is to find a separable (countable based) sub  $\sigma$ -algebra  $\mathfrak{Q}_f$  in  $\mathfrak{Q}_U$  which respects certain operations that come from the algebraic structure of  $\mathbf{Q}$  but at the same time  $f \in L^\infty(\mathbf{Q}, \mathfrak{Q}_f, \mu)$ .

Note that  $f$  itself generates a separable sub  $\sigma$ -algebra in  $\mathfrak{Q}$ . However this  $\sigma$ -algebra does not automatically respects the algebraic structure on  $\mathbf{Q}$ . For example in case of graphs  $\mathbf{Q} = \mathbf{V} \times \mathbf{V}$  where  $Q_i = V_i \times V_i$  and  $\mathbf{V}$  is the ultraproduct of  $\{V_i\}_{i=1}^\infty$ . Here we look for a separable  $\sigma$ -algebra that respects this product structure i.e. it is the "square" of some  $\sigma$ -algebra on  $\mathbf{V}$ .

4. **Topologization and algebraization (the separable model):** The set  $\mathbf{Q}$  is naturally endowed with a  $\sigma$ -topology [Szegedy \[n.d.\(b\)\]](#). Our goal here is to find a compact, Hausdorff, separable factor topology with factor map  $\pi : \mathbf{Q} \rightarrow X$  such that the  $\sigma$  algebra generated by  $\pi$  is  $\mathfrak{Q}_f$ . We also wish to construct an algebraic structure on  $X$  such that  $\pi$  is a morphism in an appropriate category. This way we can find a Borel measurable function  $f' : X \rightarrow \mathbb{C}$  such that  $f' \circ \pi = f$  holds almost everywhere. Now we regard  $(X, f')$  as a separable model for the non standard object  $\mathbf{Q}$  together with  $f$ .

This is the part of the method where we came back from the non standard universe to the world of reasonable, constructible structures. Note however that many times the algebraic structure on  $X$  is (has to be) more general than the structures in  $\mathbf{Q}$ . It is in a class  $\mathfrak{Q}$  containing  $\mathbf{Q}$ . In other words the non standard framework "teaches" us how to extend the class  $\mathbf{Q}$  to get a limit closed theory. This was very beneficial in case of higher order Fourier analysis where the non standard framework "suggested" the class of nilspaces [Camarena and Szegedy \[2010\]](#).

5. **Using the separable model for limit theory and regularization:** *There are two main application of the separable model of a function on  $\mathbb{Q}$ . The first one is that if  $\{f_i : \mathbb{Q}_i \rightarrow \mathbb{R}\}_{i=1}^\infty$  is a convergent sequence of uniformly bounded functions then the separable model for their ultra limit is an appropriate limit object for the sequence. The second application is to prove regularity lemmas in  $\mathcal{F}$  or more generally for functions on elements in  $\mathbb{Q}$ .*

Let  $Q \in \mathbb{Q}$  and  $f \in L^\infty(Q, \mu_Q)$  with  $|f| \leq 1$ . A regularity lemma is a decomposition theorem of  $f$  into a structured part and a quasi random part.

## 5 Higher order Fourier analysis: limit theory and nilspaces

Fourier analysis is a very powerful tool to study the structure of functions on finite or more generally compact Abelian groups [Rudin \[1990\]](#). If  $f : A \rightarrow \mathbb{C}$  is a measurable function in  $L^2(A, \mu_A)$  (where  $\mu_A$  is the Haar measure on  $A$ ) then there is a unique decomposition of the form  $f = \sum_{\chi \in \hat{A}} c_\chi \chi$  converging in  $L^2$  where  $\hat{A}$  is the set of linear characters of  $A$  and the numbers  $c_\chi \in \mathbb{C}$  are the Fourier coefficients. Note that for finite  $A$  the Haar measure is the uniform probability measure on  $A$ . The uniqueness of the decomposition follows from the fact that  $\hat{A}$  is a basis in the Hilbert space  $L^2(A, \mu_A)$ . It is also an important fact that the characters them self form a commutative group, called *dual group*, with respect to point-wise multiplication.

In 1953 Roth used Fourier analysis to prove a lower bound for the number of 3 term arithmetic progressions in subsets of cyclic groups [Roth \[1953\]](#). In particular it implies that positive upper density sets in  $\mathbb{Z}$  contain non trivial 3-term arithmetic progressions. The same problem for  $k$ -term arithmetic progressions was conjectured by Erdős and Turán in 1936 and solved by [Szemerédi \[1975\]](#) in 1974. Szemerédi's solution is completely combinatorial. It is quite remarkable that despite of the strength of Fourier analysis it is less useful for higher than 3 term arithmetic progressions (although it was extended for 4 term progressions in 1972 [Roth \[1972\]](#)). A deep reason for this phenomenon was discovered by [Gowers \[2001\]](#), [Gowers \[1998\]](#) in 1998. His results gave a new insight into how densities of additive patterns behave in subsets of Abelian groups by revealing a hierarchy of structural complexity classes governed by the so-called Gowers norms. Roughly speaking, at the bottom of the hierarchy there is the universe of structures, or observable quantities that can be detected by the dominant terms in Fourier decompositions. In particular the density of 3-term arithmetic progressions belongs to this part of the hierarchy. However it turns out that Fourier analysis does not go deep enough into the structure of a function (or characteristic function of a set) to clearly detect 4 or higher term arithmetic progressions: this information may be "dissolved" into many small Fourier terms. The Gowers norms  $U_2, U_3, \dots$  provide an increasingly fine way of comparing functions from a structural

point of view. The  $U_2$  norm is closely connected to Fourier analysis. Gowers formulated the following very far reaching hypothesis: *for every natural number  $k$  there is a  $k$ -th order version of Fourier analysis that is connected to the  $U_{k+1}$  norm.* In particular  $k$ -th order Fourier analysis should be the appropriate theory to study  $k + 2$  term arithmetic progressions.

Gowers coined the term “*higher order Fourier analysis*” and he developed a version of it that was enough to improve the bounds in Szemerédi’s theorem. However the following question was left open: *Is there some structural decomposition theorem in  $k$ -th order Fourier analysis that relates the  $U_k$  norm to some algebraically defined functions similar to characters?* The intuitive meaning behind these norms is that  $\|f\|_{U_k}$  is small if and only if  $f$  is quasi random in  $k - 1$ -th order Fourier analysis. A way of posing the previous question is the following: For each  $k$  let us find a set of nice enough functions (called structured functions) such that for  $|f| \leq 1$  we have that  $\|f\|_{U_k}$  is non negligible if and only if  $f$  has a non negligible correlation with one of these functions. (In case of the  $U_2$  norm the set of linear characters satisfy this property.) Such a statement is called an *inverse theorem for the  $U_k$  norm*. Despite of that fact that an inverse theorem is seemingly weaker than a complete decomposition theorem, known techniques can be used to turn them into Szemerédi type regularity lemmas.

There are several reasons why higher order Fourier analysis can’t be as exact and rigid as ordinary Fourier analysis. One obvious reason is that linear characters span the full Hilbert space  $L^2(A, \mu_A)$  and thus there is no room left for other basis elements. Quite surprisingly this obstacle disappears in the limit. If we have an increasing sequence of finite Abelian groups  $A_i$ , then there are very many function sequences  $f_i : A_i \rightarrow \mathbb{C}$  such that  $\|f_i\|_2 = 1$  and  $f_i$  is more and more orthogonal to every character  $\chi$  i.e.  $\|\hat{f}_i\|_\infty$  goes to 0. On the ultra limit group  $\mathbf{A}$  we find that ultra limits of linear characters generate only a small part of the Hilbert space  $L^2(\mathbf{A}, \mu)$ . This leaves more than enough room for higher order terms. In the rest of this chapter we give a short introduction to Gowers norms and explain how they lead to exact higher order Fourier decompositions in the limit. Then we explain an even deeper theory describing the algebraic meaning of these decompositions in terms of nilspace theory [Host and Kra \[2008b\]](#), [Camarena and Szegedy \[2010\]](#). This leads to general inverse theorems and regularity lemmas for the Gowers norms on arbitrary compact abelian groups [Szegedy \[n.d.\(b\)\]](#). Note that another but not equivalent approach to inverse theorems was developed by [Green, Tao, and Ziegler \[2012\]](#), [Tao and Ziegler \[2012\]](#) for various classes on abelian groups. They were particularly interested in inverse theorems from Gowers norms for integer sequences [Green, Tao, and Ziegler \[2012\]](#) since it leads to spectacular number theoretic applications developed by and [Green and Tao \[2010\]](#). It is important to mention that in Ergodic theory, the Host-Kra seminorms [Host and Kra \[2005\]](#) play a similar role in measure preserving systems as Gowers norms do on compact Abelian groups. Thus a tremendous amount of great ideas were transported from ergodic

theory (especially from the works of [Host and Kra \[ibid.\]](#), [Host and Kra \[2008b\]](#), [Host and Kra \[2008a\]](#) and [Ziegler \[2007\]](#)) to higher order Fourier analysis.

Let  $\Delta_t : L^\infty(A, \mu_A) \rightarrow L^\infty(A, \mu_A)$  denote the multiplicative "differential operator" defined by  $(\Delta_t f)(x) := f(x)f(x+t)$ . Since  $\Delta_{t_1}(\Delta_{t_2}(f)) = \Delta_{t_2}(\Delta_{t_1}(f))$  we can simply use multi indices  $\Delta_{t_1, t_2, \dots, t_k}$ . The  $U_k$  norm of a function  $f$  is defined by

$$\|f\|_{U_k} := \left( \mathbb{E}_{x, t_1, t_2, \dots, t_k} (\Delta_{t_1, t_2, \dots, t_k} f)(x) \right)^{1/2^k}.$$

Note that  $U_k$  is only a norm if  $k \geq 2$ . If  $k = 1$  then  $\|f\|_{U_1} = |\mathbb{E}(f)|$  and thus it is a seminorm. It was observed by Gowers that  $\|f\|_{U_2}$  is the  $l_4$  norm of the Fourier transform of  $f$  showing the connection between the  $U_2$  norm and Fourier analysis.

We say that  $\{A_i\}_{i=1}^\infty$  is a growing sequence of compact Abelian groups if their sizes tend to infinity. (If the size of a group  $A$  is already infinite then the constant sequence  $A_i = A$  satisfies this.) Let  $\mathbf{A}$  be the ultra product of a growing sequence of Abelian groups. The Gowers norms are also defined for functions in  $L^\infty(\mathbf{A}, \mathcal{Q}, \mu)$ . Quite surprisingly, all the Gowers norms become seminorms in this non-standard framework. For each  $U_k$  the set  $W_k = \{f : \|f\|_{U_k} = 0\}$  is a linear subspace in  $L^\infty(\mathbf{A}, \mathcal{Q}, \mu)$ . It turns out that the orthogonal space of  $W_k$  in  $L^2$  is equal to  $L^2(\mathbf{A}, \mathcal{F}_{k-1}, \mu)$  for some sub  $\sigma$ -algebra  $\mathcal{F}_{k-1}$  in  $\mathcal{Q}$ . Intuitively,  $\mathcal{F}_k$  is the  $\sigma$ -algebra of the  $k$ -th order structured sets. The next theorem from [Szegedy \[n.d.\(a\)\]](#) (and proved with different methods in [Szegedy \[n.d.\(b\)\]](#)) uses these  $\sigma$ -algebras to define higher order Fourier decompositions.

**Theorem 2.** *Let  $\mathbf{A}$  be as above. Then*

1. *For each natural number  $k$  there is a unique  $\sigma$ -algebra  $\mathcal{F}_k \subset \mathcal{Q}$  such that  $U_{k+1}$  is a norm on  $(\mathbf{A}, \mathcal{F}_k, \mu)$  and  $\|f - \mathbb{E}(f|\mathcal{F}_k)\|_{U_{k+1}} = 0$  for every  $f \in L^\infty(\mathbf{A}, \mathcal{Q}, \mu)$ .*
2. *We have that  $L^2(\mathbf{A}, \mathcal{F}_k, \mu) = \bigoplus_{W \in \hat{\mathbf{A}}_k} W$  where  $\hat{\mathbf{A}}_k$  is the set of shift invariant rank one modules in  $L^2(\mathbf{A}, \mathcal{F}_k, \mu)$  over the algebra  $L^\infty(\mathbf{A}, \mathcal{F}_{k-1}, \mu)$  and the sum is an orthogonal sum.*
3. *Every function  $f \in L^\infty(\mathbf{A}, \mathcal{Q}, \mu)$  has a unique decomposition in the form*

$$f = f - \mathbb{E}(f|\mathcal{F}_k) + \sum_{W \in \hat{\mathbf{A}}_k} P_W(f)$$

*converging in  $L^2$  where  $P_W$  is the projection of  $f$  to the rank one module  $\hat{\mathbf{A}}_k$ . Note that only countably many terms in the sum are non-zero.*

4. *The set  $\hat{\mathbf{A}}_k$  is an Abelian group (called  $k$ -th order dual group) with respect to point wise multiplication (using bounded representatives chosen from the modules).*



Linear characters are in shift invariant one dimensional subspaces on  $L^2(A, \mu_A)$  and each one dimensional subspace is a module over  $L^\infty$  of the trivial  $\sigma$ -algebra. This way the above theorem is a direct generalization of ordinary Fourier decomposition. The fact that higher order generalizations of the dual group appear in the limit already shows the algebraic benefits of the limit framework however, as we will see soon, this is just the surface of an even deep algebraic theory. [Theorem 2](#) can also be turned back to statements on usual compact Abelian groups using standard methods, however the exact nature disappears and various errors appear.

[Theorem 2](#) does not give a full explanation of the algebraic nature of higher order Fourier analysis. It does not provide a structural description of how the rank one modules look like. To obtain a complete algebraic characterization the new theory of nilspaces was needed which generalizes to notion of Abelian groups. This theory was developed in [Camarena and Szegedy \[2010\]](#) but it was initiated in a different form in [Host and Kra \[2008b\]](#). More detailed lecture notes on [Camarena and Szegedy \[2010\]](#) is [Candela \[2017b\]](#), [Candela \[2017a\]](#). Here we give a brief description of nilspace theory and explain how it appears in higher order Fourier analysis.

A combinatorial cube of dimension  $n$  is the product set  $\{0, 1\}^n$ . A morphism between two combinatorial cubes is a map  $f : \{0, 1\}^n \rightarrow \{0, 1\}^m$  such that it extends to an affine homomorphism (a homomorphism and a shift) from  $\mathbb{Z}^n$  to  $\{0, 1\}^m$ . A combinatorial description of morphisms is the following: each coordinate of  $f(x_1, x_2, \dots, x_n)$  is one of  $1, 0, x_i$  and  $1 - x_i$  for some  $i \in [n]$ . For example  $f(x_1, x_2, x_3, x_4) := (1, x_1, x_1, 1 - x_1, x_2, 0)$  is morphism from  $\{0, 1\}^4$  to  $\{0, 1\}^6$ . An abstract nilspace is a set  $N$  together with maps (also called morphisms) from cubes to  $N$  satisfying three simple axioms. For each  $k$  we denote by  $C^k(N) \subseteq N^{\{0, 1\}^n}$  the set of morphisms from  $\{0, 1\}^n$  to  $N$ .

1. **Composition axiom:** If  $f : \{0, 1\}^n \rightarrow \{0, 1\}^m$  is a morphism and  $g \in C^m(N)$  then  $g \circ f \in C^n(N)$ .
2. **Ergodicity axiom:**  $C^1(N) = N^{\{0, 1\}}$  (Every map  $f : \{0, 1\} \rightarrow N$  is a morphism.)
3. **Completion axiom:** Let  $f : \{0, 1\}^n \setminus \{1^n\} \rightarrow N$  be a function such that its restriction to every  $n - 1$  dimensional face of  $\{0, 1\}^n$  is a morphism. Then  $f$  can be extended to a function  $\tilde{f} : \{0, 1\}^n \rightarrow N$  such that  $\tilde{f}$  is a morphism.

If the completion in the last axiom is unique for some  $n$  then  $N$  is called an  $n - 1$  step nilspace. One step nilspaces are Abelian groups such that  $f : \{0, 1\}^n \rightarrow A$  is a morphism if and only if it can be extended to a map  $\mathbb{Z} \rightarrow A$  which is an affine morphism. We give a general family of examples for nilspaces.

**The group construction:** Let  $G$  be an at most  $k$ -nilpotent group. Let  $\{G_i\}_{i=1}^{k+1}$  be a central series with  $G_{k+1} = \{1\}$ ,  $G_1 = G$  and  $[G_i, G_j] \subseteq G_{i+j}$ . We define a cubic structure on

$G$  which depends on the given central series. The set of  $n$  morphisms  $f : \{0, 1\}^n \rightarrow G$  is the smallest set satisfying the following properties.

1. *The constant 1 map is a cube,*
2. *If  $f : \{0, 1\}^n \rightarrow G$  is a cube and  $g \in G_i$  then the function  $f'$  obtained from  $f$  by multiplying the values on some  $(n - i)$ -dimensional face from the left by  $g$  is a cube.*

Let  $\Lambda \leq G$  be a subgroup in  $G$  and let  $N = \{g\Lambda : g \in G\}$  denote the set of left cosets of  $\Lambda$  in  $G$ . We define  $C^n(N)$  to be the set of morphisms  $f : \{0, 1\}^n \rightarrow G$  composed with the map  $g \mapsto g\Lambda$ . It can be verified that  $N$  with this cubic structure is a nilspace.

**Example 1.) higher degree abelian groups:** In the previous construction if  $G$  is abelian,  $G_1, G_2, \dots, G_k = G$ ,  $G_{k+1} = \{1\}$  and  $\Lambda = \{1\}$  then  $G$  with the above cubic structure is called a  $k$ -degree Abelian group. It is true that 1-degree Abelian groups are exactly the one step nilspaces. Every  $k$ -degree Abelian group is a  $k$ -step nilspace. However there are many more  $k$ -step nilspaces for  $k \geq 2$ .

**Example 2.) nilmanifolds:** Let  $G$  be a connected nilpotent Lie group with some filtration and assume that  $\Lambda$  is a discrete co-compact subgroup. Then the left coset space of  $\Lambda$  is a compact manifold. The above construction produces a nilspace structure on  $N$ .

We can talk about topological or compact nilspaces. Assume that  $N$  is a topological space and  $C^n(N) \subseteq N^{\{0,1\}^n}$  is closed in the product topology. Then we say that  $N$  is a **topological nilspace**. If  $N$  is a compact (Hausdorff and second countable) topological nilspace then we say that  $N$  is a **compact nilspace**. Nilspaces form a category. A morphism between two nilspaces is a function  $f : N \rightarrow M$  such that for every  $n$  and  $g \in C^n(N)$  we have that  $f \circ g$  is in  $C^n(M)$ . In the category of compact nilspaces we assume that morphisms are continuous. The next theorem Szegedy [n.d.(b)] gives an algebraic description of the  $\sigma$ -algebras on  $\mathbf{A}$ . Note that the Abelian group  $\mathbf{A}$  is a one step nilspace.

**Theorem 3.** *Let  $f \in L^\infty(\mathbf{A}, \mathcal{Q}, \mu)$ . Then  $f$  is measurable in  $\mathfrak{F}_k$  if and only if there is a morphism  $\gamma : \mathbf{A} \rightarrow N$  into a  $k$ -step compact nilspace such that*

1.  *$\gamma$  is continuous in the  $\sigma$ -topology on  $\mathbf{A}$ .*
2. *There is a Borel function  $g : N \rightarrow \mathbb{C}$  such that  $f = g \circ \gamma$  almost surely.*

The above theorem shows that in the limit, the  $k$  degree structured functions are exactly those that factor through  $k$ -step compact nilspaces. This statement also implies inverse theorems for the Gowers norms on compact Abelian groups, however they are more technical. The reason for the difficulty is that the clean qualitative separation between complexity classes that we detected in the limit on  $\mathbf{A}$  becomes a more quantitative issue for concrete

compact Abelian groups. For this reason we need to involve notions such as the complexity of a nilspace and a function on it. We mention that the structured functions that appear in the inverse theorem for the  $U_{k+1}$  norm (see [Szegedy \[n.d.\(b\)\]](#)) have the form  $g \circ \gamma$  where  $\gamma$  is a morphism from  $A$  to a bounded complexity, finite dimensional nilspace  $N$  and  $g$  is a continuous function with bounded Lipschitz constant.

## References

- David J. Aldous (1985). “[Exchangeability and related topics](#)”. In: *École d’été de probabilités de Saint-Flour, XIII—1983*. Vol. 1117. Lecture Notes in Math. Springer, Berlin, pp. 1–198. MR: [883646](#) (cit. on p. [3240](#)).
- Ágnes Backhausz and Balázs Szegedy (2016). “[On the almost eigenvectors of random regular graphs](#)”. arXiv: [1607.04785](#) (cit. on p. [3238](#)).
- Itai Benjamini and Oded Schramm (2001). “[Recurrence of distributional limits of finite planar graphs](#)”. *Electron. J. Probab.* 6, no. 23, 13. MR: [1873300](#) (cit. on pp. [3236](#), [3237](#)).
- C. Borgs, J. T. Chayes, L. Lovász, V. T. Sós, and K. Vesztergombi (2008). “[Convergent sequences of dense graphs. I. Subgraph frequencies, metric properties and testing](#)”. *Adv. Math.* 219.6, pp. 1801–1851. MR: [2455626](#) (cit. on p. [3236](#)).
- (2012). “[Convergent sequences of dense graphs II. Multiway cuts and statistical physics](#)”. *Ann. of Math. (2)* 176.1, pp. 151–219. MR: [2925382](#) (cit. on p. [3236](#)).
- Christian Borgs, Jennifer T. Chayes, Henry Cohn, and Yufei Zhao (n.d.). “[An  \$L^p\$  theory of sparse graph convergence I: limits, sparse random graph models, and power law distributions](#)”. arXiv: [1401.2906](#) (cit. on p. [3238](#)).
- (2018). “[An  \$L^p\$  theory of sparse graph convergence II: LD convergence, quotients and right convergence](#)”. *Ann. Probab.* 46.1, pp. 337–396. arXiv: [1408.0744](#). MR: [3758733](#) (cit. on p. [3238](#)).
- Christian Borgs, Jennifer Chayes, László Lovász, Vera T. Sós, Balázs Szegedy, and Katalin Vesztergombi (2006). “[Graph limits and parameter testing](#)”. In: *STOC’06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing*. ACM, New York, pp. 261–270. MR: [2277152](#) (cit. on p. [3236](#)).
- Omar Antolin Camarena and Balázs Szegedy (2010). “[Nilspaces, nilmanifolds and their morphisms](#)”. arXiv: [1009.3825](#) (cit. on pp. [3244](#), [3246](#), [3248](#)).
- Pablo Candela (2017a). “Notes on compact nilspaces”. *Discrete Anal.* Paper No. 16, 57. MR: [3695479](#) (cit. on p. [3248](#)).
- (2017b). “Notes on nilspaces: algebraic aspects”. *Discrete Anal.* Paper No. 15, 59. MR: [3695478](#) (cit. on p. [3248](#)).

- Sourav Chatterjee and S. R. S. Varadhan (2011). “The large deviation principle for the Erdős-Rényi random graph”. *European J. Combin.* 32.7, pp. 1000–1017. MR: 2825532 (cit. on p. 3238).
- F. R. K. Chung, R. L. Graham, and R. M. Wilson (1989). “Quasi-random graphs”. *Combinatorica* 9.4, pp. 345–362. MR: 1054011 (cit. on p. 3241).
- Jacob W Cooper, Tomáš Kaiser, Jonathan A Noel, et al. (2015). “Weak regularity and finitely forcible graph limits”. *Electronic Notes in Discrete Mathematics* 49, pp. 139–143 (cit. on p. 3241).
- Manfred Einsiedler and Thomas Ward (2011). *Ergodic theory with a view towards number theory*. Vol. 259. Graduate Texts in Mathematics. Springer-Verlag London, Ltd., London, pp. xviii+481. MR: 2723325.
- Gábor Elek and Balázs Szegedy (2012). “A measure-theoretic approach to the theory of dense hypergraphs”. *Adv. Math.* 231.3–4, pp. 1731–1772. MR: 2964622 (cit. on p. 3242).
- Péter E. Frenkel (2018). “Convergence of graphs with intermediate density”. *Trans. Amer. Math. Soc.* 370.5, pp. 3363–3404. arXiv: 1602.05937. MR: 3766852 (cit. on p. 3238).
- H. Furstenberg (1981). *Recurrence in ergodic theory and combinatorial number theory*. M. B. Porter Lectures. Princeton University Press, Princeton, N.J., pp. xi+203. MR: 603625 (cit. on p. 3235).
- Harry Furstenberg (1977). “Ergodic behavior of diagonal measures and a theorem of Szemerédi on arithmetic progressions”. *J. Analyse Math.* 31, pp. 204–256. MR: 0498471 (cit. on pp. 3234, 3235).
- Roman Glebov, Tereza Klimosova, and Daniel Kral (2014). “Infinite dimensional finitely forcible graphon”. arXiv: 1404.2743 (cit. on p. 3241).
- W. T. Gowers (1998). “Fourier analysis and Szemerédi’s theorem”. In: *Proceedings of the International Congress of Mathematicians, Vol. I (Berlin, 1998)*. Extra Vol. I, pp. 617–629. MR: 1660658 (cit. on pp. 3243, 3245).
- (2001). “A new proof of Szemerédi’s theorem”. *Geom. Funct. Anal.* 11.3, pp. 465–588. MR: 1844079 (cit. on pp. 3243, 3245).
- Ben Green, Terence Tao, and Tamar Ziegler (2012). “An inverse theorem for the Gowers  $U^{s+1}[N]$ -norm”. *Ann. of Math. (2)* 176.2, pp. 1231–1372. arXiv: 1009.3998. MR: 2950773 (cit. on p. 3246).
- Benjamin Green and Terence Tao (2010). “Linear equations in primes”. *Ann. of Math. (2)* 171.3, pp. 1753–1850. MR: 2680398 (cit. on p. 3246).
- Hamed Hatami, László Lovász, and Balázs Szegedy (2014). “Limits of locally-globally convergent graph sequences”. *Geom. Funct. Anal.* 24.1, pp. 269–296. MR: 3177383 (cit. on pp. 3232, 3238).

- Bernard Host and Bryna Kra (2005). “Nonconventional ergodic averages and nilmanifolds”. *Ann. of Math. (2)* 161.1, pp. 397–488. MR: [2150389](#) (cit. on pp. [3235](#), [3236](#), [3246](#), [3247](#)).
- (2008a). “Analysis of two step nilsequences”. *Ann. Inst. Fourier (Grenoble)* 58.5, pp. 1407–1453. MR: [2445824](#) (cit. on pp. [3234](#), [3247](#)).
  - (2008b). “Parallelepipeds, nilpotent groups and Gowers norms”. *Bull. Soc. Math. France* 136.3, pp. 405–437. MR: [2415348](#) (cit. on pp. [3246](#)–[3248](#)).
- Dávid Kunszenti-Kovács, László Lovász, and Balázs Szegedy (2016). “Measures on the square as sparse graph limits”. arXiv: [1610.05719](#) (cit. on p. [3238](#)).
- L. Lovász and B. Szegedy (2011). “Finitely forcible graphons”. *J. Combin. Theory Ser. B* 101.5, pp. 269–301. MR: [2802882](#) (cit. on p. [3241](#)).
- László Lovász (2012). *Large networks and graph limits*. Vol. 60. American Mathematical Society Colloquium Publications. American Mathematical Society, Providence, RI, pp. xiv+475. MR: [3012035](#).
- László Lovász and Balázs Szegedy (2006). “Limits of dense graph sequences”. *J. Combin. Theory Ser. B* 96.6, pp. 933–957. MR: [2274085](#) (cit. on pp. [3236](#)–[3238](#), [3240](#)).
- (2007). “Szemerédi’s lemma for the analyst”. *Geom. Funct. Anal.* 17.1, pp. 252–270. MR: [2306658](#) (cit. on pp. [3236](#), [3238](#)–[3240](#)).
  - (2010a). “Regularity partitions and the topology of graphons”. In: *An irregular mind*. Vol. 21. Bolyai Soc. Math. Stud. János Bolyai Math. Soc., Budapest, pp. 415–446. MR: [2815610](#) (cit. on p. [3240](#)).
  - (2010b). “Testing properties of graphs and functions”. *Israel J. Math.* 178, pp. 113–156. MR: [2733066](#) (cit. on p. [3238](#)).
  - (2015). “The automorphism group of a graphon”. *J. Algebra* 421, pp. 136–166. MR: [3272377](#).
- Brendan Nagle, Vojtěch Rödl, and Mathias Schacht (2006). “The counting lemma for regular  $k$ -uniform hypergraphs”. *Random Structures Algorithms* 28.2, pp. 113–179. MR: [2198495](#) (cit. on p. [3242](#)).
- Jaroslav Nešetřil and Patrice Ossona De Mendez (Mar. 2013). “A unified approach to structural limits, and limits of graphs with bounded tree-depth”. arXiv: [1303.6471](#) (cit. on p. [3238](#)).
- Alexander A. Razborov (2008). “On the minimal density of triangles in graphs”. *Combin. Probab. Comput.* 17.4, pp. 603–618. MR: [2433944](#) (cit. on p. [3241](#)).
- Vojtěch Rödl and Mathias Schacht (2007). “Regular partitions of hypergraphs: regularity lemmas”. *Combin. Probab. Comput.* 16.6, pp. 833–885. MR: [2351688](#) (cit. on p. [3242](#)).
- Vojtěch Rödl and Jozef Skokan (2004). “Regularity lemma for  $k$ -uniform hypergraphs”. *Random Structures Algorithms* 25.1, pp. 1–42. MR: [2069663](#) (cit. on p. [3242](#)).
- K. F. Roth (1953). “On certain sets of integers”. *J. London Math. Soc.* 28, pp. 104–109. MR: [0051853](#) (cit. on pp. [3235](#), [3245](#)).

- (1972). “Irregularities of sequences relative to arithmetic progressions. IV”. *Period. Math. Hungar.* 2. Collection of articles dedicated to the memory of Alfréd Rényi, I, pp. 301–326. MR: [0369311](#) (cit. on p. [3245](#)).
- Walter Rudin (1990). *Fourier analysis on groups*. Wiley Classics Library. Reprint of the 1962 original, A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, pp. x+285. MR: [1038803](#) (cit. on p. [3245](#)).
- J. Solymosi (2004). “A note on a question of Erdős and Graham”. *Combin. Probab. Comput.* 13.2, pp. 263–267. MR: [2047239](#) (cit. on p. [3242](#)).
- Balazs Szegedy (n.d.[a]). “Higher order Fourier analysis as an algebraic theory I”. arXiv: [0903.0897](#) (cit. on p. [3247](#)).
- (n.d.[b]). “On higher order Fourier analysis”. Preprint. arXiv: [1203.2260](#) (cit. on pp. [3239](#), [3242](#), [3244](#), [3246](#), [3247](#), [3249](#), [3250](#)).
- (n.d.[c]). “Sparse graph limits, entropy maximization and transitive graphs”. arXiv: [1504.00858](#) (cit. on p. [3238](#)).
- Balázs Szegedy (2011). “Limits of kernel operators and the spectral regularity lemma”. *European J. Combin.* 32.7, pp. 1156–1167. MR: [2825541](#).
- E. Szemerédi (1975). “On sets of integers containing no  $k$  elements in arithmetic progression”. *Acta Arith.* 27. Collection of articles in memory of JuriĭVladimirovič Linnik, pp. 199–245. MR: [0369312](#) (cit. on pp. [3242](#), [3245](#)).
- Endre Szemerédi (1978). “Regular partitions of graphs”. In: *Problèmes combinatoires et théorie des graphes (Colloq. Internat. CNRS, Univ. Orsay, Orsay, 1976)*. Vol. 260. Colloq. Internat. CNRS. CNRS, Paris, pp. 399–401. MR: [540024](#) (cit. on p. [3240](#)).
- Terence Tao (2012). *Higher order Fourier analysis*. Vol. 142. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, pp. x+187. MR: [2931680](#).
- Terence Tao and Tamar Ziegler (2012). “The inverse conjecture for the Gowers norm over finite fields in low characteristic”. *Ann. Comb.* 16.1, pp. 121–188. MR: [2948765](#) (cit. on p. [3246](#)).
- Evan Warner (2012). “Ultraproducts and the Foundations of Higher Order Fourier Analysis”. PhD thesis. Bachelor thesis, Princeton University.
- Tamar Ziegler (2007). “Universal characteristic factors and Furstenberg averages”. *J. Amer. Math. Soc.* 20.1, pp. 53–97. MR: [2257397](#) (cit. on pp. [3235](#), [3236](#), [3247](#)).

Received 2017-12-04.



# EXTREMAL THEORY OF ORDERED GRAPHS

GÁBOR TARDOS

## Abstract

We call simple graphs with a linear order on the vertices *ordered graphs*. Turán-type extremal graph theory naturally extends to ordered graphs. This is a survey on the ongoing research in the extremal theory of ordered graphs with an emphasis on open problems.

## 1 Definitions

An *ordered graph* is a simple graph with linear order on the vertices. Formally, an ordered graph is triple  $(V, E, <)$ , where  $V$  is the vertex set,  $E \subseteq \binom{V}{2}$  is the edge set and  $<$  is a linear order relation on  $V$ . In this survey we assume that  $V$  is finite. We say that  $(V, E)$  is the simple graph *underlying* the ordered graph  $(V, E, <)$  and that the ordered graph  $(V, E, <)$  is an *ordering* of the simple graph  $(V, E)$ . The notion of subgraph and isomorphism naturally extend to ordered graphs: the ordered graphs  $(V, E, <)$  and  $(V', E', <')$  are isomorphic if there is an order preserving isomorphism between the graphs  $(V, E)$  and  $(V', E')$ . The ordered graph  $(V', E', <')$  is an ordered subgraph of  $(V, E, <)$  if  $V' \subseteq V$ ,  $E' \subseteq E$  and  $<'$  is the restriction of  $<$  to  $V'$ .

Armed with this definition we can extend some classic areas of graph theory to ordered graphs. Here we do this for Turán-type extremal graph theory. It asks for the maximal number of edges in a simple graph of given size that *avoids* (i.e., does not contain as a subgraph) a specified pattern or all members of a given family of patterns. In particular, we are interested in the maximal number,  $\text{ex}(\mathcal{P}, n)$ , of edges in an  $n$ -vertex simple graph that has no subgraph isomorphic to any member of the family  $\mathcal{P}$ . Note that we must require that  $\mathcal{P}$  does not contain empty graphs in order for this definition to make sense. If the forbidden pattern is a singleton we write  $\text{ex}(P, n)$  to denote  $\text{ex}(\{P\}, n)$ . We call  $\text{ex}(\mathcal{P}, n)$  the *extremal function* of the family  $\mathcal{P}$  and will concentrate on its asymptotic

---

Supported by the Cryptography “Lendület” project of the Hungarian Academy of Sciences and by the National Research, Development and Innovation Office, NKFIH projects K-116769 and SNN-117879.



behavior. Accordingly, all the asymptotic notations like  $O(\cdot)$ ,  $o(\cdot)$  should be interpreted for a fixed family  $\mathcal{P}$  and, in particular, the implied constants in  $O(\cdot)$  may depend on this family.

For a natural extension of this theory to ordered graphs, we consider a family  $\mathcal{P}$  of ordered graphs and we are looking for the largest number  $\text{ex}_{<}(\mathcal{P}, n)$  of edges in an  $n$ -vertex ordered graph with no ordered subgraph isomorphic to any member of  $\mathcal{P}$ . As before, we require that each member of  $\mathcal{P}$  has at least one edge and simplify the notation for singleton families by writing  $\text{ex}_{<}(P, n)$  to denote  $\text{ex}_{<}(\{P\}, n)$ . Our remark on the asymptotic notation also applies here.

Let us first observe that the extremal theory of ordered graph is strictly richer than classical extremal graph theory in the sense that the classical questions can be equivalently asked in this setting, but we can also ask new questions. In particular, for any family  $\mathcal{P}$  of simple graphs one can form the family  $\mathcal{P}_{<}$  consisting all orderings of the patterns in  $\mathcal{P}$  and then we trivially have:

$$\text{ex}(\mathcal{P}, n) = \text{ex}_{<}(\mathcal{P}_{<}, n).$$

On the other hand, if we forbid, say, a single ordered graph  $P$ , the corresponding extremal function  $\text{ex}_{<}(P, n)$  has no direct analogue in the classical theory. We naturally have  $\text{ex}_{<}(P, n) \geq \text{ex}(\overline{P}, n)$ , where  $\overline{P}$  is the simple graph underlying  $P$ , but this lower bound is typically very weak, since avoiding  $\overline{P}$  in a particular order is often much easier than avoiding it in all possible orders.

Extensions of Ramsey theory to ordered graph is also studied extensively, see [Balko, Cibulka, Král, and Kynčl \[2015\]](#) and [Conlon, Fox, Lee, and Sudakov \[2017\]](#).

## 2 Basic results

Any survey about extremal graph theory should start with the following classical theorem of [Turán \[1941\]](#), of which the  $r = 2$  special case (the maximal number of edges in a triangle-free graph) was proved by Mantel in 1907. The result gives the exact extremal function when the forbidden graph is a complete graph. Further, for the  $(r + 1)$ -vertex complete graph  $K_{r+1}$  the theorem states that the unique (up to isomorphism)  $n$ -vertex graph with the maximum number of edges avoiding  $K_{r+1}$  is the *Turán graph*  $T(n, r)$  formed by partitioning the vertices into  $r$  almost equal parts and letting a pair of vertices form an edge if and only if they are from distinct parts. Note that the number of edges of the Turán graph  $T(n, r)$  is  $(1 - 1/r)n^2/2 - O(1)$ , where the  $O(1)$  error term comes from unequal parts and can go as high as  $\lfloor r/8 \rfloor$ . As a consequence, we have:

**Theorem 1** ([Turán \[ibid.\]](#)). *For every  $r \geq 1$  we have*

$$\text{ex}(K_{r+1}, n) = \left(1 - \frac{1}{r}\right) \frac{n^2}{2} - O(1).$$

A trivial generalization of this result to ordered graphs involves the ordered clique, the unique ordering of the complete graph. Let  $K_{r+1,<}$  stand for the  $(r+1)$ -vertex ordered clique and we trivially have  $\text{ex}_{<}(K_{r+1,<}, n) = \text{ex}(K_{r+1}, n)$ . A more revealing generalization is about the ordered path  $P_{r+1,<}$  obtained from the  $(r+1)$ -vertex path  $P_{r+1}$  with the natural order on the vertices where edges connect neighboring vertices in the order. We have  $\text{ex}_{<}(P_{r+1,<}, n) = \text{ex}(K_{r+1}, n)$ . Here the direction  $\leq$  follows from the fact that  $P_{r+1,<}$  is an ordered subgraph of  $K_{r+1,<}$  and  $\geq$  follows from the fact that if we order the vertices of  $T(n, r)$  in a way that the  $r$  parts become intervals in the ordering, then the resulting ordered graph does not contain  $P_{r+1,<}$  as an ordered subgraph. Note, however, that in the case  $r$  does not divide  $n$ , this process may yield several non-isomorphic extremal ordered graphs. Note also that the path  $P_{r+1}$  has several non-isomorphic orderings for  $r > 1$ , and by [Theorem 3](#) below, all other orderings have smaller extremal functions.

The most general result in Turán-type extremal graph theory is the following consequence of the Erdős–Stone theorem, [Erdős and Simonovits \[1966\]](#). It basically states that the extremal function of *any* simple graph is close to the extremal function of the complete graph with the same *chromatic number*.

**Theorem 2** ([Erdős and Stone \[1946\]](#) and [Erdős and Simonovits \[1966\]](#)). *Let  $\mathcal{P}$  be a family of simple graphs and  $r+1 = \min_{P \in \mathcal{P}} \chi(P)$  be the smallest chromatic number of a member of this family. We have*

$$\text{ex}(\mathcal{P}, n) = \left(1 - \frac{1}{r}\right) \frac{n^2}{2} + o(n^2).$$

[Pach and Tardos \[2006\]](#) gave a generalization of this result for ordered graphs. It is based on finding the “correct” version of the chromatic number for ordered graph.

The *interval coloring* of an ordered graph is a proper coloring of the underlying simple graph in which each color class is an interval of the linear order. The *interval chromatic number* of an ordered graph  $P$  is the smallest number of colors in an interval coloring of  $P$ . We write  $\chi_{<}(P)$  to denote the interval chromatic number of  $P$ .

Note that the interval chromatic number is much simpler to compute than the chromatic number because a greedy strategy suffices. Indeed, we can form the first color class by taking longest initial segment of the vertices that form an independent set and proceed similarly for subsequent color classes. The process yields an interval coloring with the fewest possible colors. Using this definition, the generalization of the Erdős–Stone–Simonovits theorem is rather straightforward:

**Theorem 3** (Erdős–Stone–Simonovits Theorem for ordered graphs [Pach and Tardos \[ibid.\]](#)). *Let  $\mathcal{P}$  be a family of ordered graphs and  $r+1 = \min_{P \in \mathcal{P}} \chi_{<}(P)$  be the smallest interval chromatic number of a member of this family. We have*

$$\text{ex}_{<}(\mathcal{P}, n) = \left(1 - \frac{1}{r}\right) \frac{n^2}{2} + o(n^2).$$

Just as the classic version of this theorem, it gives exact asymptotics for the extremal function of ordered graphs unless the ordered graph is *ordered bipartite* (i.e., has interval chromatic number 2). We will therefore concentrate on ordered bipartite graphs. Containment between ordered bipartite graphs can also be visualized using the language of containment in 0-1 matrices. This connection is explored in the next section.

### 3 Connection to 0-1 matrices

A 0-1 matrix is simply a matrix with all entries being 0 or 1. The *weight* of such a matrix is the number of its 1-entries. A 0-1 matrix  $A$  is said to *contain* another 0-1 matrix  $P$  if  $P$  is a submatrix of  $A$  or  $P$  is obtained from a submatrix of  $A$  by replacing some 1-entries with 0-entries. Note that permuting rows or columns is not allowed. If  $A$  does not contain  $P$ , we say it *avoids*  $P$ . The extremal problem for 0-1 matrix containment can be formulated as computing (or estimating) the following extremal function for families  $\mathcal{P}$  of 0-1 matrices:  $\text{Ex}(\mathcal{P}, n)$  is the maximal weight of an  $n$ -by- $n$  0-1 matrix that avoids all matrices in  $\mathcal{P}$ . We require that all matrices in  $\mathcal{P}$  have positive weights. We write  $\text{Ex}(P, n)$  to denote  $\text{Ex}(\{P\}, n)$ .

For a 0-1 matrix  $P$ , let  $G_P$  stand for the ordered bipartite graph whose vertices correspond to the rows and columns of  $P$ , the order of the vertices agrees with the order of rows and columns in  $P$  with all row-vertices preceding all column vertices, and with an edge between a row-vertex and a column-vertex if and only if the corresponding entry in  $P$  is 1. This makes  $P$  the bipartite adjacency matrix of  $G_P$  and turns the weight of  $P$  into the number of edges in  $G_P$ . The close connection between the extremal theory of ordered bipartite graphs and 0-1 matrices follows from the trivial observation that if a 0-1 matrix  $A$  contains another 0-1 matrix  $P$ , then the ordered graph  $G_A$  also contains  $G_P$ . The converse is also true if the homomorphism of  $G_P$  to  $G_A$  maps row-vertices to row-vertices and column-vertices to column-vertices. This extra condition is automatically satisfied if both the last row and first column of  $P$  contain at least one 1-entry, so in this case we have  $\text{Ex}(P, n) \leq \text{ex}_{<}(G_P, 2n)$ . There is no equality in general, because  $\text{ex}_{<}(G_P, 2n)$  is the maximum number of edges among all ordered graphs on  $2n$  vertices avoiding  $G_P$  and the extremal ones may not be ordered bipartite. Still, the two extremal functions are really close to each other as shown by the following observation:

**Theorem 4** (Pach and Tardos [2006]). *For a 0-1 matrix  $P$  and the corresponding ordered bipartite graph  $G_P$  we have*

$$\text{Ex}(n, P) \leq \text{ex}_{<}(2n, G_P) = O(\text{Ex}(n, P) \log n).$$

The logarithmic term in the bound above is needed even for some small matrices, e.g., for the matrix

$$P = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

For this matrix, we have  $\text{Ex}(n, P) = 2n - 1$ , but for the corresponding ordered graph  $G_P$  one has  $\text{ex}_{<}(n, G_P) = n \log n + O(n)$ , where  $\log$  stands for the binary logarithm. A construction showing the lower bound for this estimate is an ordered graph whose vertices are adjacent if and only if their distance in the ordering is a power of 2.

The extremal theory of 0-1 matrices predates the related theory of ordered graphs. Füredi [1990] established the extremal function for a specific 2-by-3 0-1 matrix and used this result for a problem in combinatorial geometry: he bounded the number of diagonals of equal length in a convex  $n$ -gon. Independently, Bienstock and Györi [1991] found the extremal function of few small 0-1 matrices. Later Füredi and Hajnal [1992] started a systematic study of the extremal theory of 0-1 matrices. This latter paper not only contained many nice results, but was also rich in conjectures and had a significant effect on future research. As we will see, some of these conjectures have since been proved, others disproved and some are still open.

## 4 Relation between ordered and unordered extremal functions

A (too) general conjecture that appeared in Füredi and Hajnal [ibid.] can be informally stated as

**Conjecture 1.** *For all 0-1 matrix  $P$  of positive weight we have*

$$\text{Ex}(P, n) \approx \text{ex}(\overline{G}_P, n),$$

where  $\overline{G}_P$  is the simple graph underlying the ordered graph  $G_P$ .

This conjecture connects ordered extremal theory to the classical unordered one. We clearly have an inequality in one direction:

$$\text{Ex}(n, P) \leq \text{ex}_{<}(2n, G_P) \leq \text{ex}(2n, \overline{G}_P) = O(\text{ex}(n, \overline{G}_P)).$$

By Theorem 4, the first inequality is almost tight for any pattern, so we concentrate on the second inequality and ask how large the ratio between the two sides can be:

**Question 1.** *How high can the ratio  $\frac{\text{ex}_{<}(n, P)}{\text{ex}(n, \overline{P})}$  be for an ordered bipartite graph  $P$  with more than two vertices and at least one edge and its underlying simple graph  $\overline{P}$ ?*

The paper Pach and Tardos [2006] gives an ordering  $P_k$  of the cycle  $C_{2k}$  with  $\text{ex}_{<}(n, P_k) = \Omega(n^{4/3})$ . Using the Bondy-Simonovits theorem on the extremal function of cycles Bondy

and Simonovits [1974], one obtains that the ratio in Question 1 for the pattern  $P = P_k$  is  $\Omega(n^{1/3-1/k})$ , which disproves Conjecture 1. We do not know if any pattern with higher ratio, say  $\Omega(n^{1/3})$  exists. For an upper bound, we trivially have  $O(n)$ , as both the enumerator and the denominator are functions between  $n$  and  $n^2$ . In fact, they are  $O(n^{2-\epsilon})$  for some  $\epsilon > 0$  depending on the size of  $P$  by the Kővári–Sós–Turán theorem Kővari, Sós, and Turán [1954], so the ratio is always  $O(n^{1-\epsilon})$ , but no better upper bound is known.

## 5 Forests

The Füredi and Hajnal [1992] formulated the special case of Conjecture 1 for cycle-free patterns  $P$  separately. Here we call a 0-1 matrix  $P$  cycle-free if the corresponding simple graph  $\overline{G}_P$  is cycle-free, that is a forest. In this case,  $\text{ex}(n, \overline{G}_P)$  (the extremal function of an unordered forest) is trivially linear. Concerning the corresponding question for ordered graphs, we formulate the following conjecture:

**Conjecture 2** *For an ordered bipartite forest  $P$  and any  $c > 1$ , we have*

$$\text{ex}_<(P, n) = o(n^c).$$

Note first that if this conjecture is true, then it characterizes the ordered graphs with almost linear extremal functions. Indeed, if  $P$  is not ordered bipartite, then  $\text{ex}_<(P, n) = \Theta(n^2)$  by Theorem 3, while if the underlying graph  $\overline{P}$  contains a cycle, then  $\text{ex}_<(P, n) \geq \text{ex}(\overline{P}, n) = \Omega(n^c)$  for some  $c > 1$ .

Note that  $o(n^c)$  for all  $c > 1$  is not the only possible way to quantify the notion that a function is “close to linear”. One could formulate a stronger conjecture with a bound  $O(n \log^c n)$  for a constant  $c = c_P$  depending on  $P$ , or even with an  $O(n \log n)$  bound. Conjecture 2 and the conjecture with the  $O(n \log^c n)$  bound are still open and by Theorem 4 are equivalent to the similar conjectures about  $\text{Ex}(P, n)$  for cycle-free 0-1 matrices  $P$ . The strongest form of the conjecture (an  $O(n \log n)$  bound) was also considered for a while and was supported by the fact that it was easy to find an extremal function of the order  $\Theta(n \log n)$ , but there was no known example of an ordered bipartite forest whose extremal function grows faster. Note that here the distinction between cycle-free 0-1 matrices and ordered bipartite forests is meaningful. As we have seen above, there exists a three-edge ordered bipartite path whose extremal function is  $\Theta(n \log n)$ . Although the extremal function of the corresponding 2-by-2 matrix is linear, there is a 3-by-2 0-1 matrix whose extremal function is  $\Theta(n \log n)$ . This was the first 0-1 matrix considered in the context of extremal functions in the papers Füredi [1990] and Bienstock and Györi [1991].

Pettie [2011] found a cycle-free 0-1 matrix  $P$  with extremal function slightly higher than  $n \log n$ : for this matrix  $P$  one has  $\text{Ex}(P, n) = \Omega(n \log n \log \log n)$ . By this, he

disproved the strengthening of Conjecture 2 with the  $O(n \log n)$  upper bound, but the conjecture may still hold with the bound  $O(n \log^2 n)$ . Pettie's result was slightly improved and the current best lower bound is due [Park and Shi \[n.d.\]](#). They found a cycle-free 0-1 matrix  $P_m$  with  $\text{Ex}(P_m, n) = \Omega(n \log n \log \log n \cdots \log^{(m)} n)$ , where  $\log^{(m)}$  denotes the  $m$ -times-iterated logarithm function.

On the positive side,  $\text{ex}_{<}(P, n) = O(n \log^c n)$  was established in [Pach and Tardos \[2006\]](#) for all ordered bipartite forests with at most 6 vertices. The most general result in this direction is due to [Korándi, Tardos, Tomon, and Weidert \[2017\]](#). They call a 0-1 matrix  $M$  *vertically degenerate* if for any submatrix  $M' = (a_{ij})$  of  $M$  consisting of  $l > 1$  rows one can find  $1 \leq k < l$  such that  $M'$  has at most one column  $j$  with two 1-entries  $a_{ij} = a_{i'j} = 1$  satisfying  $1 \leq i \leq k < i' \leq l$ . Note that all vertically degenerate 0-1 matrices are cycle-free. All cycle-free 0-1 matrices with at most three rows are vertically degenerate, but there are 4-row cycle-free 0-1 matrices that are not vertically degenerate. Using a density increment argument they prove the following theorem.

**Theorem 5** ([Korándi, Tardos, Tomon, and Weidert \[ibid.\]](#)). *Let  $M$  be a vertically degenerate 0-1 matrix with  $l$  rows. We have*

$$\text{Ex}_{<}(M, n) = n2^{O(\log^{1-1/l} n)}.$$

This result implies that Conjecture 2 holds for all ordered graphs  $G_M$ , where  $M$  is a vertically degenerate 0-1 matrix. By symmetry, Conjecture 2 is also true for all  $G_M$ , where  $M$  is *horizontally degenerate*, that is, the transpose of  $M$  is vertically degenerate. Conjecture 2 has not been verified for any other ordered bipartite forest. The smallest of these open cases is an ordered path on 8 vertices.

## 6 Linear extremal functions

[Füredi and Hajnal \[1992\]](#) conjectured, and later [Marcus and Tardos \[2004\]](#) proved, that  $\text{Ex}(P, n) = O(n)$  for permutation matrices  $P$ . It is not hard to see that this result can be restated in the following equivalent form (although [Theorem 4](#) does not directly imply this equivalence).

**Theorem 6.** *The extremal function of any ordered bipartite matching  $P$  is linear. That is,*

$$\text{Ex}(P, n) = O(n).$$

Conjecture 2, if true, characterizes all ordered graphs with almost linear extremal functions. It would be nice to find a characterization of ordered graphs or 0-1 matrices with linear extremal functions. One possibility is finding all *minimally nonlinear matrices*. We call a 0-1 matrix  $P$  *minimally nonlinear*, if its extremal function  $\text{Ex}(P, n)$  is nonlinear,

but  $\text{Ex}(P', n) = O(n)$  for all 0-1 matrices  $P' \neq P$  contained in  $P$ . It might be possible to find such a characterization, but the following theorem indicates that this might be a difficult task:

**Theorem 7** (Geneson [2009] and Keszegh [2009]). *There are infinitely many minimally nonlinear matrices.*

## 7 Interaction between ordered graphs

We finish this survey with a few remarks on interactions between extremal functions of different forbidden patterns. Let us start with the classical extremal theory of graphs. Clearly, we have

$$\text{ex}(\{G, H\}, n) \leq \min(\text{ex}(G, n), \text{ex}(H, n)). \quad (*)$$

By Theorem 2, the two sides are asymptotically the same for non-bipartite graphs  $G$  and  $H$ . It is easy to see that they differ by a factor of less than 2 if only one of the graphs is bipartite. For bipartite graphs, the situation is more complicated. We say that  $G$  and  $H$  *interact* if the two sides differ more than by a constant factor. It is not known if there exists any interacting pair of graphs, but Faudree and Simonovits [n.d.] conjecture that the cycle  $C_4$  and the subdivision of the complete graph  $K_4$ , in which each edge is subdivided with a single new vertex, do interact.

In contrast, for 0-1 matrices it is not hard to find a lot of interactions. Consider the 3-by-2 matrix  $M_1 = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \end{pmatrix}$ . Füredi [1990] and Bienstock and Györi [1991] proved that  $\text{Ex}(M_1) = \Theta(n \log n)$ . By symmetry, the extremal functions of the matrices  $M_2 = \begin{pmatrix} 1 & 0 & 1 \\ 1 & 1 & 0 \end{pmatrix}$ ,  $M_3 = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$  and  $M_4 = \begin{pmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix}$  are same. The following theorem implies that each of  $M_2$ ,  $M_3$  and  $M_4$  interacts with  $M_1$ :

**Theorem 8** (Tardos [2005]).

$$\text{Ex}(\{M_1, M_2\}, n) = \Theta(n)$$

$$\text{Ex}(\{M_1, M_3\}, n) = \Theta(n \log n / \log \log n)$$

$$\text{Ex}(\{M_1, M_4\}, n) = \Theta(n \log \log n)$$

These results represent the first step toward exploring interactions between different patterns. It would be interesting to find “stronger” interactions, where the ratio between the right and left sides of (\*) is larger than logarithmic.

**Question 2** *Are there ordered graphs  $G$  and  $H$  such that*

$$\text{ex}_{<}(\{G, H\}, n) = O(\min(\text{ex}_{<}(G, n), \text{ex}_{<}(H, n))/n^\epsilon)$$

*for some  $\epsilon > 0$ ?*

## References

- Martin Balko, Josef Cibulka, Karel Král, and Jan Kynčl (2015). “[Ramsey numbers of ordered graphs](#)”. *Electronic Notes in Discrete Mathematics* 49, pp. 419–424 (cit. on p. [3254](#)).
- Dan Bienstock and Ervin Györi (1991). “[An extremal problem on sparse 0-1 matrices](#)”. *SIAM J. Discrete Math.* 4.1, pp. 17–27. MR: [1090285](#) (cit. on pp. [3257](#), [3258](#), [3260](#)).
- J. A. Bondy and M. Simonovits (1974). “Cycles of even length in graphs”. *J. Combinatorial Theory Ser. B* 16, pp. 97–105. MR: [0340095](#) (cit. on p. [3257](#)).
- David Conlon, Jacob Fox, Choongbum Lee, and Benny Sudakov (2017). “[Ordered Ramsey numbers](#)”. *J. Combin. Theory Ser. B* 122, pp. 353–383. MR: [3575208](#) (cit. on p. [3254](#)).
- P. Erdős and M. Simonovits (1966). “A limit theorem in graph theory”. *Studia Sci. Math. Hungar* 1, pp. 51–57. MR: [0205876](#) (cit. on p. [3255](#)).
- P. Erdős and A. H. Stone (1946). “[On the structure of linear graphs](#)”. *Bull. Amer. Math. Soc.* 52, pp. 1087–1091. MR: [0018807](#) (cit. on p. [3255](#)).
- R. J. Faudree and M. Simonovits (n.d.). “On a class of degenerate extremal graph problems II” (cit. on p. [3260](#)).
- Zoltán Füredi (1990). “[The maximum number of unit distances in a convex  \$n\$ -gon](#)”. *J. Combin. Theory Ser. A* 55.2, pp. 316–320. MR: [1075714](#) (cit. on pp. [3257](#), [3258](#), [3260](#)).
- Zoltán Füredi and Péter Hajnal (1992). “[Davenport-Schinzel theory of matrices](#)”. *Discrete Math.* 103.3, pp. 233–251. MR: [1171777](#) (cit. on pp. [3257](#)–[3259](#)).
- Jesse T. Geneson (2009). “[Extremal functions of forbidden double permutation matrices](#)”. *J. Combin. Theory Ser. A* 116.7, pp. 1235–1244. MR: [2527608](#) (cit. on p. [3260](#)).
- Balázs Keszegh (2009). “[On linear forbidden submatrices](#)”. *J. Combin. Theory Ser. A* 116.1, pp. 232–241. MR: [2469261](#) (cit. on p. [3260](#)).
- Dániel Korándi, Gábor Tardos, István Tomon, and Craig Weidert (Nov. 2017). “[On the Turán number of ordered forests](#)”. arXiv: [1711.07723](#) (cit. on p. [3259](#)).
- T. Kövari, V. T. Sós, and P. Turán (1954). “On a problem of K. Zarankiewicz”. *Colloquium Math.* 3, pp. 50–57. MR: [0065617](#) (cit. on p. [3258](#)).
- Adam Marcus and Gábor Tardos (2004). “[Excluded permutation matrices and the Stanley-Wilf conjecture](#)”. *J. Combin. Theory Ser. A* 107.1, pp. 153–160. MR: [2063960](#) (cit. on p. [3259](#)).
- János Pach and Gábor Tardos (2006). “[Forbidden paths and cycles in ordered graphs and matrices](#)”. *Israel J. Math.* 155, pp. 359–380. MR: [2269435](#) (cit. on pp. [3255](#)–[3257](#), [3259](#)).
- S. G. Park and Q. Shi (n.d.). “[New bounds on extremal numbers in acyclic ordered graphs](#)” (cit. on p. [3259](#)).



- Seth Pettie (2011). “Degrees of nonlinearity in forbidden 0-1 matrix problems”. *Discrete Math.* 311.21, pp. 2396–2410. MR: [2832138](#) (cit. on p. [3258](#)).
- Gábor Tardos (2005). “On 0-1 matrices and small excluded submatrices”. *J. Combin. Theory Ser. A* 111.2, pp. 266–288. MR: [2156213](#) (cit. on p. [3260](#)).
- P. Turán (1941). “On an extremal problem in graph theory”. *Matematikai és Fizikai Lapok* 48, pp. 436–452 (cit. on p. [3254](#)).

Received 2018-03-01.

GÁBOR TARDOS  
RÉNYI INSTITUTE OF MATHEMATICS  
BUDAPEST  
HUNGARY  
[tardos@renyi.hu](mailto:tardos@renyi.hu)  
[tardosgabor@gmail.com](mailto:tardosgabor@gmail.com)

# ASYMPTOTIC ENUMERATION OF GRAPHS WITH GIVEN DEGREE SEQUENCE

NICHOLAS WORMALD

## Abstract

We survey results on counting graphs with given degree sequence, focusing on asymptotic results, and mentioning some of the applications of these results. The main recent development is the proof of a conjecture that facilitates access to the degree sequence of a random graph via a model incorporating independent binomial random variables. The basic method used in the proof was to examine the changes in the counting function when the degrees are perturbed. We compare with several previous uses of this type of method.

## 1 Introduction

We sometimes count objects in a class simply because they are there. This is especially true if they are abundantly occurring as mathematical objects (e.g. partitions, sets, or graphs with certain properties), and then we are often pleased if we obtain a simple formula. For example, the number of trees on  $n$  vertices is  $n^{n-2}$ . But we cannot hope for simple formulae in all cases, and even if we are extremely lucky and the formula is not very complicated, it may be hard to find or difficult to prove. Yet a formula is often useful in order to prove other things, and for such purposes we are frequently satisfied with an approximate or asymptotic formula. For instance, many results in probabilistic combinatorics (see e.g. [Alon and Spencer \[2000\]](#)) use such estimates.

The problems considered here involve graphs or, in an alternate guise, matrices. A non-negative integer  $m \times n$  matrix  $A$  with row sums  $\mathbf{r} = (r_1, \dots, r_m)$  and column sums  $\mathbf{s} = (s_1, \dots, s_n)$  is equivalent to a bipartite multigraph  $G$  with vertex set  $V_1 \cup V_2$  where  $V_1 = \{u_1, \dots, u_m\}$  and  $V_2 = \{v_1, \dots, v_n\}$ . The  $(i, j)$  entry of the matrix is the multiplicity of the edge  $u_i v_j$ . Here  $A$  is the adjacency matrix of  $G$ . If  $A$  is 0-1 (binary) then  $G$  has no multiple edges and is thus a (simple) bipartite graph. If every row sum is 2, then  $A$  is the

incidence matrix of a multigraph with degree sequence  $\mathbf{s}$ . If  $A$  is symmetric, then it is also the adjacency matrix of a pseudograph  $H$  (where loops and multiple edges are permitted). We can obtain  $H$  from  $G$  by identifying  $u_i$  with  $v_i$ ,  $i = 1, \dots, n$ . If  $A$  furthermore has zero diagonal, then  $H$  is a multigraph, and if it is in addition 0-1, then  $H$  is a graph. For the bulk of this article, we discuss only the enumeration of graphs, bipartite graphs and multigraphs on a given set of vertices, without mentioning the immediate corollaries for matrices.

We focus on graphs with given degree sequence  $\mathbf{d} = (d_1, \dots, d_n)$ , where  $d_i$  is the degree of vertex  $i$ . For this purpose we normally implicitly assume that  $n$  is the number of vertices in the graphs concerned, and that they have vertex set  $\{1, \dots, n\}$ . An early example of a formula concerning such graphs is from Moon [1970]: the number of trees with degree sequence  $\mathbf{d}$  is precisely

$$\binom{n-2}{d_1-1, \dots, d_n-1} = \frac{(n-2)!}{(d_1-1)! \cdots (d_n-1)!}.$$

A graph is  $d$ -regular if its degree sequence is  $(d, d, \dots, d)$ . No such neat formula is known for the number  $g_{d,n}$  of  $d$ -regular graphs on  $n$  vertices, but we do have the asymptotic formula

$$(1-1) \quad g_{d,n} \sim \frac{(dn)! e^{-(d^2-1)/4}}{(dn/2)! 2^{dn/2} \prod d_i!}$$

as  $n \rightarrow \infty$  with  $d$  fixed. Here  $a \sim b$  means  $a = b(1 + o(1))$ , with  $o(\cdot)$  the Landau notation. This and many more developments are described in Section 2.

Counting graphs by degree sequence is strongly related to finding the distribution of the degree sequence of a random graph on  $n$  vertices in either of the two most common random graph models. In the model  $\mathcal{G}(n, p)$ , where edges occur independently and each with probability  $p$ , the degree of a vertex is distributed binomially as  $\text{Bin}(n-1, p)$ , but the degrees of the vertices are not independent of each other. Bollobás [2001] devotes an early chapter to this topic. The model  $\mathcal{G}(n, p)$  can be viewed as a mixture of the models  $\mathcal{G}(n, m)$ , where  $m$  is distributed binomially as  $\text{Bin}(n(n-1)/2, p)$ . Here,  $\mathcal{G}(n, m)$  has  $m$  edges selected from all  $\binom{n}{2}$  possible positions uniformly at random. We use  $g(\mathbf{d})$  to denote the number of graphs with degree sequence  $\mathbf{d}$ . Then the probability that a random graph  $G \in \mathcal{G}(n, m)$  has degree sequence  $\mathbf{d}$  can be evaluated precisely as  $g(\mathbf{d}) / \binom{\binom{n}{2}}{m}$ .

McKay and Wormald [1990a] made a conjecture (which has now been verified) on the number of graphs with degree sequence  $\mathbf{d}$ , for a wide-ranging choice of possible vectors  $\mathbf{d}$  stated by Liebenau and Wormald [2017]. Let  $A_n$  and  $B_n$  be two sequences of probability spaces with the same underlying set for each  $n$ . Suppose that whenever the event  $H_n$  satisfies  $\mathbb{P}(H_n) = n^{-O(1)}$  in either model, it is true that  $\mathbb{P}_{A_n}(H_n) \sim \mathbb{P}_{B_n}(H_n)$ . Then we

say that  $A_n$  and  $B_n$  are *asymptotically quite equivalent* (a.q.e.). Also, throughout this paper we use  $\omega(f(n))$  to denote a function of  $n$  such that  $\omega(f(n))/f(n) \rightarrow \infty$  as  $n \rightarrow \infty$ . The conjecture from [McKay and Wormald \[1997\]](#) implies the following two propositions.

Let  $\mathfrak{D}(\mathfrak{G})$  denote the (random) degree sequence of a random graph  $\mathfrak{G}$ . Define  $\mathfrak{B}_p(n)$  to be the random sequence consisting of  $n$  independent binomial variables  $\text{Bin}(n-1, p)$ . We use  $\mathbb{Q}|_F$  to denote the restriction of the probability space  $\mathbb{Q}$  to an event  $F$ . The lack of specification of  $p$  in the following is due to the obvious fact that the restricted space  $B_p(n) \mid_{\Sigma=2m}$  is independent of  $p$ . The restriction imposed on  $m$  in this proposition only excludes graphs with so few or so many edges that their degree sequence is excruciatingly boring.

**Proposition 1.1.** *Let  $\Sigma$  denote the sum of the components of the random vector  $\mathfrak{B}_p(n)$ . Let  $0 < p < 1$ . Then  $\mathfrak{D}(\mathfrak{G}(n, m))$  and  $B_p(n) \mid_{\Sigma=2m}$  are a.q.e. provided that  $\max\{m, \binom{n}{2} - m\} = \omega(\log n)$ .*

This gives a very appealing way to derive properties of the degree sequence of a random graph with  $n$  vertices and  $m$  edges: consider independent binomials as above with  $p = 2m/n(n-1)$  and condition on the event  $\Sigma = 2m$ , which has (not very small) probability  $\Theta(1/\sqrt{p(1-p)n^2})$ . Even more appealing, it was also shown in [McKay and Wormald \[ibid.\]](#) that a statement like the above proposition would imply the following one.

**Proposition 1.2.** *Let  $\Sigma$  denote the sum of the components of the random vector  $B_{\hat{p}}(n)$ , where  $\hat{p}$  is randomly chosen according to the normal distribution with mean  $p$  and variance  $p(1-p)/n(n-1)$ , truncated at 0 and 1. Then  $\mathfrak{D}(\mathfrak{G}(n, p))$  and  $B_{\hat{p}}(n) \mid_{\Sigma \text{ is even}}$  are a.q.e. provided that  $p(1-p) = \omega(\log^3 n/n^2)$ .*

This second proposition gives even easier access to properties of the degree sequence of the random graph  $\mathfrak{G}(n, p)$ , as conditioning on the parity of  $\Sigma$  is insignificant for many properties, and the effect of the random choice of  $\hat{p}$  can be evaluated by integration. This was all made explicit in [McKay and Wormald \[ibid.\]](#), where there are a number of helper theorems to make it easy to transfer results from the independent binomial model  $\mathfrak{B}_p(n)$ . It was also observed in [McKay and Wormald \[ibid.\]](#), from known asymptotic formulae, that the main conjecture (and hence also Propositions 1.1 and 1.2) holds when  $p = o(1/\sqrt{n})$  or  $p(1-p) > n/c \log n$ . The gap between these ranges, where  $p(1-p)$  is between roughly  $1/\sqrt{n}$  and  $1/\log n$ , was only recently plugged (see [Section 4.1](#)), which proved the main conjecture of [McKay and Wormald \[1990a\]](#) in full.

This gives the following asymptotic formula for the number of graphs with given degree sequence, that holds provided the degrees are reasonably close to each other. The degree sequence of a random graph in  $\mathfrak{G}(n, p)$  with high probability falls into the range covered,

for the  $p$  considered here. Given  $\mathbf{d}$ , let

$$d = \frac{1}{n} \sum_{i=1}^n d_i,$$

$$\mu = \mu(n) = d/(n-1)$$

and

$$\gamma_2 = (n-1)^{-2} \sum_{i=1}^n (d_i - d)^2.$$

Also, given a sequence  $\mathbf{d}$  with even sum  $2m$ , let  $P(\mathbf{d})$  denote the probability that  $\mathbf{d}$  occurs in the model  $B_p(n) \mid_{\Sigma=2m}$ . Using Stirling's formula, we easily find

$$P(\mathbf{d}) \sim \sqrt{2}(\mu^\mu(1-\mu)^{1-\mu})^{n(n-1)/2} \prod \binom{n-1}{d_i}.$$

**Proposition 1.3.** *For some absolute constant  $\epsilon > 0$ ,*

$$(1-2) \quad g(\mathbf{d}) \sim \exp\left(\frac{1}{4} - \frac{\gamma_2^2}{4\mu^2(1-\mu)^2}\right) P(\mathbf{d})$$

*provided that  $dn$  is even,  $\max_j |d_j - d| = o(n^\epsilon \min\{d, n-d-1\}^{1/2})$  and  $n^2 \min\{\mu, 1-\mu\} \rightarrow \infty$ .*

**Note.** Where we state asymptotics such as in (1-2) where we do not explicitly give  $\mathbf{d}$  as a function of  $n$ , there are two possible interpretations. One is that we do, indeed, consider  $\mathbf{d}$  a function of  $n$ . Then the limit easily makes sense, and the interpretation should be that this holds for any  $\mathbf{d}(n)$  satisfying the given constraints. The other interpretation is that the asymptotic convergence should contain a bound that is uniform over all  $\mathbf{d}$  under consideration. The first interpretation is the default, and two interpretations are easily seen to be equivalent when the permitted domain of  $\mathbf{d}$  is suitably closed, such that one can consider the ‘worst’ sequence  $\mathbf{d}(n)$  for each  $n$ .

The validity of this formula cannot possibly extend to very “eccentric” degree sequences, in particular certain non-graphical degree sequences (i.e. sequences for which no graph exists). Examples of degree sequences at the fringe of the formula's validity can be obtained as follows. Consider the degree sequence  $\mathbf{d}$  with  $n/2$  entries  $d+x$  and  $n/2$  of  $d-x$ , where  $d = o(\sqrt{n})$  and  $x \leq d$ . By expanding the formula in McKay and Wormald [1991a, Theorem 5.2] appropriately, we can check that  $g(\mathbf{d})$  differs from the formula in (1-2) by a factor  $\exp(x^6(2d^2 - x^2)/nd^5 + o(1))$ . Hence, for such  $d$  and  $x$ , (1-2) is correct iff  $x = o(n^{1/6}d^{1/2})$ , which is for all  $x \leq d$  when  $d = o(n^{1/3})$ , but not for larger  $d$ .

Formulae similar to (1-2) are also now known for bipartite graphs and loopless directed graphs; see [Section 2.3](#).

The next section traces the development of results on this problem. [Section 3](#) gives a description and observations on the basic approach used in [Liebenau and Wormald \[2017\]](#), and [Section 4](#) discusses applications of the results (and methods). Some open problems are mentioned in the last section.

## 2 Results on enumeration of graphs by degrees

We focus on asymptotic results mainly because these formulae, for the problems of our concern, are much simpler than the corresponding known exact formulae. Due to the complexity of the exact formulae, they tend to be useless in proving results such as [Propositions 1.1 and 1.2](#) and the applications in [Section 4](#). For these, simple formulae are generally required, even if only approximate. The interest in limiting behaviour is thus prominent, as in many areas of mathematics.

**2.1 Asymptotic results for graphs.** Most of the following results come with explicit error bounds in the asymptotic approximations. To keep the description simple, we omit these error bounds, and similarly make little mention of a number of variations and extensions given in the papers quoted. The description in this section is basically in order of increasing maximum degree of the graphs being treated. This largely corresponds to chronological order, the main exception being for very dense graphs with degrees approximately  $cn$ .

Our story begins with [Read \[1958\]](#) thesis. Using Polya's cycle index theory and manipulation of generating functions, Read found a formula for the number of graphs with given degree sequence, from which he was able to obtain a simple asymptotic formula in the case of the number  $g_3(n)$  of 3-regular graphs:

$$(2-1) \quad g_3(n) \sim \frac{(3n)!e^{-2}}{(3n/2)!288^{n/2}}.$$

(Here  $n$  is restricted to being even, as it is in all our formulae when the total degree parity condition forces it.)

Further progress on enumeration of regular graphs was stymied by the lack of an amenable approach. However, before long significant developments occurred in enumeration of  $m \times n$  non-negative integer matrices, which in the case of 0-1 matrices correspond to bipartite graphs. Let  $b(\mathbf{r}, \mathbf{s})$  denote the number of 0-1 matrices, of dimensions  $m \times n$ , with row sum vector  $\mathbf{r}$  and column sum vector  $\mathbf{s}$ . We refer to the entries of a vector  $\mathbf{r}$  as  $r_i$ , entries of  $\mathbf{s}$  as  $s_i$ , and so on. We can assume these vectors have equal sums, and define for

a vector  $\mathbf{d}$

$$M_1(\mathbf{d}) = \sum_i d_i, \quad \text{and in general} \quad M_j(\mathbf{d}) = \sum_i [d_i]_j$$

where  $[x]_j = x(x-1)\cdots(x-j+1)$ . O'Neil [1969] showed that, as long as  $m = n$  and some  $\epsilon > 0$  satisfies  $r_i, s_i \leq (\log n)^{1/4-\epsilon}$  for all  $i$ ,

$$(2-2) \quad b(\mathbf{r}, \mathbf{s}) \sim \frac{M_1! e^{-\alpha}}{\prod_{i=1}^m r_i! \prod_{i=1}^n s_i!}$$

as  $M_1 \rightarrow \infty$ , where  $\alpha = M_2(\mathbf{r})M_2(\mathbf{s})/2M_1^2$  and  $M_1 = M_1(\mathbf{r}) = M_1(\mathbf{s})$ . Of course, in the corresponding bipartite graphs,  $M_1$  is the number of edges and  $\mathbf{r}$  and  $\mathbf{s}$  are the degree sequences of the vertices in the two parts.

For fixed  $r$ , Everett and P. R. Stein [1971/72] gave a different proof for the  $r$ -regular case of (2-2), i.e.  $r_i = s_i = r$  for all  $i$  (in particular this again requires  $m = n$ ), using symmetric function theory. This is the bipartite version of (1-1).

A. Békéssy, P. Békéssy, and Komlós [1972] showed that (2-2) holds even for  $m \neq n$ , as long as there is a constant upper bound on maximum degree of the corresponding graphs, i.e.  $\max_i r_i$  and  $\max_i s_i$ . They used the following model. Consider  $n$  buckets, and  $M_1$  balls with  $r_i$  labelled  $i$ ,  $1 \leq i \leq m$ . Distribute the balls at random in the buckets by starting with a random permutation of the balls, placing the first  $s_1$  into bucket 1, the next  $s_2$  into bucket 2, and so on. (The balls and buckets are all mutually distinguishable.) Each distribution corresponds to a matrix whose  $(i, j)$  entry is the number of balls labelled  $i$  falling into bucket  $j$ . In this model, it is easy to see that the 0-1 matrices with row sum vector  $\mathbf{r}$  and column sum vector  $\mathbf{s}$  are equiprobable. The number of permutations is  $M_1!$ , and the number of these corresponding to any one 0-1 matrix is the denominator of (2-2). Hence (2-2) follows once we show that the event that no entry is at least 2 has probability  $e^{-\alpha+o(1)}$ . This is done using an inclusion-exclusion technique, equivalent to applying Bonferroni's inequalities or Brun's sieve.

Mineev and Pavlov [1976] used more accurate analysis of the same model to show that (2-2) still holds with maximum degree  $(\gamma \log n)^{1/4}$  for  $\gamma < 2/3$ , and a slightly more extended range in the regular case.

Bender [1974] used a model equivalent to that in A. Békéssy, P. Békéssy, and Komlós [1972] to obtain results for matrices with integer entries in the range  $[0, \dots, t]$ , but with bounded row and column sums. He allowed some entries to be forced to be 0. This permits the diagonal to be forced to be 0 in the case  $m = n$ , hence giving a formula for the number of loopless digraphs with given in- and out-degree sequence.

About 20 years passed from Read's result (2-1) for the 3-regular case, before any advance was made in the case of non-bipartite graphs. In 1978 Bender and Canfield [1978] showed that the number  $g(\mathbf{d})$ , of graphs with degree sequence  $\mathbf{d}$  with the maximum degree

bounded, is given by

$$(2-3) \quad g(\mathbf{d}) \sim \frac{M_1! e^{-\alpha}}{(M_1/2)! 2^{M_1/2} \prod_{i=1}^n d_i!}$$

where  $\alpha = M_2/2M_1 + M_2^2/4M_1^2$ , and  $M_j = M_j(\mathbf{d})$  for all statements about graphs. For this result, they used a model of involutions of a set of cardinality  $M_1$  partitioned into blocks of sizes  $d_1, d_2, \dots, d_n$ . When there are no fixed points, we can regard the two elements in a 2-cycle of an involution as two balls of the same label, to recover the model of Békéssy et al. applied to the incidence matrix of graphs. (Bender and Canfield obtained other results for symmetric matrices with nonzero diagonal, in which the involutions are permitted to have fixed points.) We easily see that the number of such involutions corresponding to a given simple graph is precisely the denominator in (2-3), since the labels of the edges are immaterial, giving a factor  $(M_1/2)!$  in addition to the considerations applied for the matrix counting in (2-2). The factor  $e^{-\alpha}$  is shown to be asymptotically the probability that the graph obtained in the model is simple. We call this event  $\mathcal{S}$ .

Independently, in my PhD thesis [Wormald \[1978\]](#) I used the asymptotic results of [A. Békéssy, P. Békéssy, and Komlós \[1972\]](#) for bipartite graphs, to derive (2-3) for bounded degrees.

[Bollobás \[1980\]](#) gave the *configuration model*, in which  $d_i$  objects, commonly called half-edges, are assigned to each vertex  $i$ , and then paired up at random. Two paired half-edges form an edge joining the corresponding vertices. It is readily seen that conditioning on no loops or multiple edges gives a uniformly random graph. This is clearly equivalent to the earlier models, such as the involution model of Bender and Canfield, where each 2-cycle corresponds to two paired half-edges. The model was used in [Bollobás \[ibid.\]](#) to extend the validity of (2-3) to maximum degree  $\sqrt{2 \log n} - 1$ , provided that a certain lower bound on the number of edges  $(M_1/2)$  is satisfied. In place of the inclusion-exclusion based arguments in the earlier papers, Bollobás used the method of moments for Poisson random variables, which is essentially equivalent. Using this model, we can write

$$(2-4) \quad g(\mathbf{d}) = \frac{|\Phi| \mathbb{P}(\mathcal{S})}{\prod d_i!}$$

where  $\Phi$  is the set of pairings in the model, with  $|\Phi| = M_1! / ((M_1/2)! 2^{M_1/2})$ .

Aside from enumeration results, [Bollobás \[1981\]](#), and then many others, found the configuration model a convenient starting point to prove properties of random graphs with given degrees.

Several further developments involved estimating  $\mathbb{P}(\mathcal{S})$  for a wider range of degree sequences.



For the bipartite case, [Bollobás and McKay \[1986\]](#) extended the validity of the formula to cover  $m \leq n$  when maximum degree is at most  $\log^{1/3} m$  and a certain lower bound on the number of edges is satisfied.

Again for the bipartite case, [McKay \[1984\]](#) then took a much bigger step, by introducing switchings as a technique for this problem. (See [Section 3.1](#) for a detailed description of this method, including some subsequent developments mentioned below.) He obtained (2-2) under conditions that are a little complicated to state in general, but apply for all  $r = o(n^{1/3})$  in the  $r$ -regular case (i.e.  $r_i = s_i = r$  for all  $i$ , and  $m = n$ ).

[McKay \[1985\]](#) applied the same technique to the graph case, with a similar restriction on degrees, again obtaining (2-3) with the same formula for  $\alpha$ .

From this point onwards, the results for the two cases, graphs versus bipartite graphs, have generally been obtained more or less in tandem using the same methods, so we continue tracing only the graph case in detail.

It was evident that with considerably more effort, the switching approach should extend to higher degrees, but at the expense of much case analysis, which was enough of a deterrent to stifle such further development. Instead, after several years, [McKay and Wormald \[1991a\]](#) found a different version of switchings that enabled a much easier advance. The result was that for degree sequences with  $\Delta = o(M_1^{1/3})$ , (2-4) holds with

$$\mathbb{P}(\mathcal{S}) = \exp\left(-\frac{M_2}{2M} - \frac{M_2^2}{4M^2} - \frac{M_2^2 M_3}{2M^4} + \frac{M_2^4}{4M^5} + \frac{M_3^2}{6M^3} + O\left(\frac{\Delta^3}{M}\right)\right).$$

This covers the  $d$ -regular case for  $d = o(\sqrt{n})$ .

[Janson \[2009, 2014\]](#) was interested in characterising the degree sequences for which  $\mathbb{P}(\mathcal{S}) \rightarrow 0$ . He showed by analysing the configuration model for degree sequence  $\mathbf{d}$ , using the method of moments, that for  $M_1 = \Theta(n)$ ,

$$\mathbb{P}(\mathcal{S}) \rightarrow 0 \quad \text{iff} \quad \sum d_i^2/n \rightarrow \infty,$$

and that for  $M_2 = O(M_1)$  and  $M_1 \rightarrow \infty$ , we have the asymptotic formula

$$\mathbb{P}(\mathcal{S}) = \exp\left(-\frac{1}{2} \sum \lambda_{ii} - \sum_{i < j} (\lambda_{ij} - \log(1 + \lambda_{ij}))\right) + o(1)$$

where  $\lambda_{ij} = \sqrt{d_i(d_i - 1)d_j(d_j - 1)}/(2M_1)$ . This was the first general result to apply to some sequences with maximum degree as large as  $\sqrt{n}$ .

[Gao and Wormald \[2016\]](#) analysed cases in which the configuration model produces edges of much higher multiplicity than previous studies, using a major extension of the switching method mentioned above. This resulted in an asymptotic formula for  $g(\mathbf{d})$  when  $\mathbf{d}$  satisfies some very complicated conditions. We describe one simple consequence. We

say  $(d_1, \dots, d_n)$  is *power-law distribution-bounded with parameter  $\gamma$*  if there exists  $C > 0$  such that the number of  $d_i$  taking value at least  $i$  is at most  $Cn \sum_{j \geq i} j^{-\gamma}$  for all  $i$  and  $n$ . This condition restricts the maximum  $d_i$  to  $O(n^{1/(\gamma-1)})$ . Before [Gao and Wormald \[ibid.\]](#), there were no asymptotic enumeration formulae for such sequences when  $\gamma \leq 3$ . However, many naturally occurring networks seem to have power law degree sequences with  $\gamma < 3$ . From [Gao and Wormald \[ibid., Theorem 3\]](#) (corrected), we see that if  $\mathbf{d}$  is a power-law distribution-bounded sequence with parameter  $3 > \gamma > 1 + \sqrt{3} \approx 2.732$ , then with  $\Phi$  given just after (2-4),

$$g(\mathbf{d}) = \frac{|\Phi|}{\prod d_i!} \exp \left( -\frac{M_1}{2} + \frac{M_2}{2M_1} + \frac{3}{4} + \sum_{i < j} \log(1 + d_i d_j / M_1) + O(\xi) \right),$$

where  $\xi = n^{(2+2\gamma-\gamma^2)/(\gamma-1)}$ .

Recently, [Burstein and Rubin \[2015\]](#) presented an approach that would give a formula that is valid up to maximum degree  $n^{1-\delta}$  for any fixed  $\delta > 0$ , using a finite amount of computation. We say a little more about this in [Section 3.3](#).

[Liebenau and Wormald \[2017\]](#) introduced a new approach and “plugged the gap” in the formulae with the following result.

**Theorem 2.1.** *Let  $\mu_0 > 0$  be a sufficiently small constant, and let  $1/2 \leq \alpha < 3/5$ . If  $\mu = d/(n-1)$  satisfies  $\mu \leq \mu_0$  and, for all fixed  $K > 0$ ,  $(\log n)^K/n = O(\mu)$ , and  $|d_i - d| \leq d^\alpha$  for all  $i \in [n]$  then (1-2) holds (provided  $dn$  is even).*

Together with the previous results, this establishes [Proposition 1.3](#) and hence [Propositions 1.1 and 1.2](#). The method seems strong enough to cover all the results mentioned above, though some significant tinkering would need to be done to obtain the results for eccentric degree sequences in [Janson \[2009, 2014\]](#) and [Gao and Wormald \[2016\]](#).

At this point, we travel slightly back in time to consider results for dense graphs. Of course, extremely dense cases are covered by simply complementing the sparse cases above. All results for graphs of average degree comparable with  $n$  are based on extracting coefficients from the ‘obvious’ generating function:

$$(2-5) \quad g(\mathbf{d}) = [x_1^{d_1} \cdots x_n^{d_n}] \prod_{i < j} (1 + x_i x_j).$$

The generating function is derived by letting  $x_i$  mark the degree of vertex  $i$ , so the term  $x_i x_j$  denotes the presence of the edge  $ij$  and the term 1 denotes its absence. Coefficients are extracted using Cauchy’s integral formula, for multiple dimensions.

[McKay and Wormald \[1990a\]](#) evaluated the integrals to obtain the result of [Proposition 1.3](#) in the case that  $\mu(1-\mu) > c/\log n$  for fixed  $c > 2/3$ .

Barvinok and Hartigan [2013] used a similar approach, with different analysis, to obtain a result for a wider range of degrees when  $\min\{\mu, 1 - \mu\} = \Theta(1)$ .

Independently and almost simultaneously with the completion of the proof of Proposition 1.3 in Liebenau and Wormald [2017], Isaev and McKay [2016] made an exciting new advance, by developing the theory of martingale concentration for complex martingales, which (amongst other things) enabled them to reproduce and extend the results in McKay and Wormald [1990a] and Barvinok and Hartigan [2013], by better analysis of the integrals involved. Moreover, they recently report (by private communication) being able to obtain a result, in a certain implicit form, that applies to a wide range of degree sequences with average degree at least  $n^a$  for any fixed  $a > 0$ . They have used this to show in particular that (1-2) is valid for the  $d$ -regular case with  $d = \omega(n^{1/7})$ .

**2.2 Exact enumeration for graphs.** We include here only a selection of results that bear some relation to our main topic of simple asymptotic formulae. These concern the number  $g_d(n) = g(d, d, \dots, d)$  of  $d$ -regular graphs on  $n$  vertices.

As mentioned above, Read [1960] found a formula for the number of graphs with given degree sequence. His approach was to count the incidence matrices of the graphs, i.e. 0-1 matrices with column sums vector  $\mathbf{d}$  and all row sums 2. Pólya's Hauptsatz for enumeration under the action of the symmetric group was used to eliminate the distinction between matrices that are equivalent up to permuting the rows (edges). To eliminate multiple edges, Read used a version of the Hauptsatz in which the 'figures' are distinct. Unfortunately, this gives a very complicated formula, involving generating functions for which the extraction of coefficients is difficult. Nevertheless, in the case of  $g_3(n)$ , Read obtained a formula containing a single summation that he analysed to obtain (2-1).

One can also ask for simple recurrence relations. There are two direct uses of these: for efficient computation of the numbers, and also, as Read [1958] shows, the recurrence relation can be combined with an asymptotic formula to deduce an asymptotic series expansion. Read did this in the case of 3-regular graphs, finding the first few terms of a series in powers of  $n^{-1}$ . Read and Wormald [1980] found a similar recurrence for  $g_4(n)$ .

Goulden, Jackson, and Reilly [1983] considered the generating function (2-5), in the case of regular graphs, and obtained a different recurrence relation for  $g_4(n)$ .

Gessel (90) showed recurrence relations exist for  $g_d(n)$  when  $d$  is fixed. (To be precise, he showed that the generating function for  $d$ -regular graphs is D-finite.) However, to our knowledge, these recurrences have not been found explicitly for any  $d \geq 5$ .

Chen and Louck [1999] obtained a formula for  $g_3(n)$  by first getting a formula for matrices with row sums 3 and column sums 2 (and related problems) using a little symmetric function theory, and then using inclusion-exclusion to delete the multiple edges. The same method should work for  $d > 3$  but would appear to get rapidly much more complicated.

**2.3 Asymptotics for special types of graphs.** Aside from Moon’s formula for trees with given degree sequence mentioned in the introduction, many results are known on the asymptotic number of trees of some given variety with given degree sequence. We refer the interested reader to [Drmotá \[2009\]](#).

As implied in [Section 2](#), there have been further results on bipartite graphs, and a recent summary may be found in [Liebenau and Wormald \[2017\]](#). In particular, that paper completes the proof of the analogue of [Proposition 1.1](#) in the case that the two sides of the bipartite graph have reasonably similar cardinalities. See also [Section 5](#).

**2.4 Similar results for other structures.** [Greenhill and McKay \[2013\]](#) obtained results for multigraphs analogous to the graph results of [McKay and Wormald \[1991a\]](#), for a similar range of degrees. Already in [A. Békéssy, P. Békéssy, and Komlós \[1972\]](#), the number of bipartite multigraphs with given degree sequence (with bounded maximum degree) was also obtained.

The results on graphs obtained in [Liebenau and Wormald \[2017\]](#) were accompanied by similar results on loopless directed graphs.

Some results on hypergraphs have been obtained by means similar to those discussed for graphs. For simplicity we omit these from the scope of this article.

A  $k \times n$  Latin rectangle can be defined as an ordered set of  $k$  disjoint perfect matchings which partition the edges of a bipartite graph (i.e. a properly  $k$ -edge-coloured bipartite graph) on vertex sets  $V_1 = \{1, \dots, n\}$  and  $V_2 = \{n+1, \dots, 2n\}$ . Asymptotic estimates of the numbers of these were obtained for ever-increasing  $k$  by [Erdős and Kaplansky \[1946\]](#), [Yamamoto \[1951\]](#), [C. M. Stein \[1978\]](#), culminating in the result of [Godsil and McKay \[1990\]](#) for  $k = o(n^{6/7})$ . In a recent preprint, [Leckey, Liebenau, and Wormald \[n.d.\]](#) reached  $k = o(n/\log n)$  using the method described above for graph enumeration. See [Section 3.3](#).

[Kuperberg, Lovett, and Peled \[2017\]](#) have a different probabilistic approach to enumeration of several other kinds of regular combinatorial structures such as orthogonal arrays,  $t$ -designs and certain regular hypergraphs.

### 3 The perturbation method

The author coined this term in [Wormald \[1996\]](#), to refer to enumeration methods based on comparing the number of structures with a given parameter set to the numbers of structures with slightly perturbed parameter sets. Overall, it can be expressed as estimating the ratio of probabilities of “adjacent” points in a discrete probability space. How to estimate the ratio depends on the application, and some examples are discussed below. It is relatively

straightforward to use the information on ratios for adjacent points. We may use the following result, in which the probabilities actually occurring relate to  $\mathcal{P}'$  and are compared with those in some “ideal” probability space  $\mathcal{P}$ . Here  $\text{diam}(F)$  is the diameter of  $F$ .

**Lemma 3.1** (Liebenau and Wormald [2017]). *Let  $\mathcal{P}$  and  $\mathcal{P}'$  be probability spaces with the same underlying set  $\Omega$ . Let  $F$  be a graph with vertex set  $\mathcal{W} \subseteq \Omega$  such that  $\mathbb{P}_{\mathcal{P}}(v), \mathbb{P}_{\mathcal{P}'}(v) > 0$  for all  $v \in \mathcal{W}$ . Suppose that  $1 > \epsilon > 0$  with  $\min\{\mathbb{P}_{\mathcal{P}}(\mathcal{W}), \mathbb{P}_{\mathcal{P}'}(\mathcal{W})\} \geq e^{-\epsilon}$ , and that  $C, \delta > 0$  satisfy, for every edge  $uv$  of  $F$ ,*

$$(3-1) \quad \left| \log \frac{\mathbb{P}_{\mathcal{P}'}(u)}{\mathbb{P}_{\mathcal{P}'}(v)} - \log \frac{\mathbb{P}_{\mathcal{P}}(u)}{\mathbb{P}_{\mathcal{P}}(v)} \right| \leq C\delta.$$

*Suppose further that  $\text{diam}(F) \leq r < \infty$ . Then for each  $v \in \mathcal{W}$  we have*

$$|\log \mathbb{P}_{\mathcal{P}'}(v) - \log \mathbb{P}_{\mathcal{P}}(v)| \leq rC\delta + O(\epsilon).$$

The proof is entirely straightforward, using a telescoping product of ratios along a path joining  $u$  to  $v$  of length at most  $r$ .

**Note:** if the error function  $\delta$  depends on  $u$  and  $v$ , it might be possible to take advantage of this and finish with a smaller error term than what is suggested by the lemma. In the applications so far, this would not give any significant gain.

Estimating the ratios of adjacent probabilities, and hence  $\delta$ , is the main requirement for applying the method. We describe several different but related examples. The overall structure of the arguments in most cases was phrased differently from Lemma 3.1, but is essentially equivalent. In most cases, ratios of adjacent probabilities were found to approximate the ratio of corresponding probabilities of a Poisson distributed random variable. In such cases we may take  $\Omega = \mathbb{N}$  and  $\mathbb{P}_{\mathcal{P}}(i) = \mathbb{P}(X = i)$  for a given Poisson random variable  $X$ . The graph  $F$  then has edges  $\{i, i + 1\}$  for all  $i$  in some suitably defined set  $\mathcal{W}$ .

**3.1 Switchings for pairings.** Arguments involving switchings or similar concepts have been used in many places in combinatorics. An early example close to our topic is provided by the bounds on the probabilities of subgraphs of random graphs obtained by McKay [1981].

McKay [1985] applied switchings to estimate the probability of the event  $\mathcal{S}$  (that a simple graph results) in the configuration model for degree sequence  $\mathbf{d}$  described in Section 2.1. We can describe this as two rounds of the perturbation method, first estimating the number of pairings with no loops, and then, among those, the number with no double edge. (To be precise, even higher multiplicities were eliminated first.) Roughly, the double edge round was as follows. Let  $\mathcal{C}_i$  denote set of pairings with  $i$  double edges. Take

any pairing  $P$  in  $\mathcal{C}_i$  for  $i \geq 1$ . To apply a *switching* to  $P$ , choose a random pair from a random double pair, call it  $ab$ , and any other random pair, call it  $cd$ , and then delete those two pairs and replace them by  $ac$  and  $bd$ . This is depicted in Figure 1. We can count the

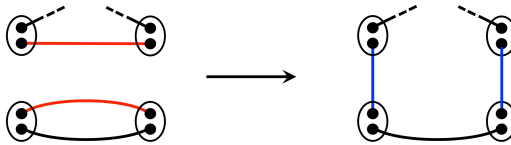


Figure 1: *Switching two pairs*

switchings that produce members of  $\mathcal{C}_{i-1}$  in two ways: the number that can be applied to such a  $P \in \mathcal{C}_i$ , and the number that can *produce* a pairing  $P' \in \mathcal{C}_{i-1}$ . By estimating the ratio of these two numbers, McKay obtained an estimate of the ratio  $\mathbb{P}(\mathcal{C}_i)/\mathbb{P}(\mathcal{C}_{i-1})$ , in the model. In his argument, the ratios were used to estimate the ratios  $\mathbb{P}(\mathcal{C}_i)/\mathbb{P}(\mathcal{C}_0)$ , and the absolute sizes were obtained by arguing about the sum of these ratios. Alternatively and equivalently, we could apply Lemma 3.1 by letting  $\Omega = \mathbb{N}$  and assigning probabilities in  $\mathcal{P}'$ , according to the distribution of the random number,  $X$ , of double edges in a pairing. Then, with  $\mathcal{P}$  defined as an appropriate Poisson random variable, we can deduce (3-1) for quite small values of  $\delta$  and  $C$ . The set  $\mathcal{W}$  is defined to be  $[0, M_0]$  for a suitable constant  $M_0$ , and  $\epsilon$  can be bounded by estimating  $\mathbb{E}[X]_k$  for suitable large  $k$ .

To obtain sufficient accuracy, McKay introduced secondary switchings that enabled him to argue about the local structure of random elements of  $\mathcal{C}_i$ . This was needed because the number of reverse switchings depends heavily on the structure of  $P'$  mentioned above, though it is quite stable for typical  $P'$ . This enabled him to obtain the asymptotic formula in the case of  $d$ -regular graphs for  $d = o(n^{1/3})$ .

McKay and Wormald [1991a] introduced fancier switchings, involving more pairs, for which local structure had little effect on the number of reverse switchings, and hence directly gave accuracy comparable to the secondary switchings of McKay. Further secondary switchings produced results that, in the  $d$ -regular case, reached  $d = o(\sqrt{n})$ . An additional benefit of the fancier switchings was that they led to a useful uniform generator of  $d$ -regular graphs for  $d = O(n^{1/3})$ . (See Section 4.3.)

**3.2 Relation to Stein's method.** Stein's asymptotic formula for  $k \times n$  Latin rectangles in C. M. Stein [1978] was based on Chen's method (which is elsewhere called the Stein-Chen or Chen-Stein method) for Poisson approximation, which was based in turn on a more general method of Stein for approximating a random variable. His basic strategy was

the same as all results on this problem before [Leckey, Liebenau, and Wormald \[n.d.\]](#): given  $k - 1$  rows of a rectangle, estimate the probability that a random new row is compatible with the given rows. Stein approached this as follows. Given the first  $k - 1$  rows, define  $X$  to be the number of column constraints that a random permutation, upon being inserted as the  $k$ th row, would violate. A permutation  $\Pi$  is associated with a random permutation  $\Pi'$  using a random pair  $(j_1, j_2)$  of distinct elements of  $[n]$ :  $\Pi'(i)$  is defined as  $\Pi(j_1)$  if  $i = j_2$ ,  $\Pi(j_2)$  if  $i = j_1$ , and  $\Pi(i)$  otherwise. Then, with  $X'$  the number of column constraints violated by  $\Pi'$ , [C. M. Stein \[1978, \(3\) in Section 2\]](#) is

$$\frac{\mathbb{P}(X = x + 1)}{\mathbb{P}(X = x)} = \frac{\mathbb{P}(X' = x + 1 \mid X = x)}{\mathbb{P}(X' = x \mid X = x + 1)}.$$

This is clearly equivalent to restricting to pairs  $(j_1, j_2)$  where  $\Pi'$  violates one less column constraint than  $\Pi$ , which is a direct analogue of the switchings for pairings described above. Stein also uses secondary randomisation in [C. M. Stein \[ibid., \(20\), Section 2\]](#), which corresponds to the secondary switchings described above. However, he does not seem to use the analogue of the fancier switchings. On the other hand, analogues of the fancier switchings were applied to the problem of uniformly generating Latin rectangles by [McKay and Wormald \[1991b\]](#).

**3.3 Iterated applications.** In the ideal situation, we can start with a set of initial estimates for the ratios of adjacent probabilities, and then iteratively feed these estimates into equations that are derived from some operation, such as switchings, to improve the accuracy of the estimates.

We can view the equations as specifying an operator on functions which fixes the true ratios, except perhaps for a quantified error term. Normally we could then hope to apply standard concepts of fixed point analysis to show that the true ratios are close to a fixed point of this operator. Initial bounds are required on the true ratios, as they provide an initial “guess”, and also simultaneously they can be used to guarantee that, essentially, this guess is in the domain of attraction of the correct fixed point of the operator.

Here are some examples.

(i) Nonexistence of subgraphs of  $\mathcal{G}(n, p)$  and  $\mathcal{G}(n, m)$ .

Consider estimating the number  $t_{n,m}$  of graphs with  $n$  vertices,  $m$  edges and no triangles (cycles of length 3). [A. Frieze \[1992\]](#) switched edges of subgraphs to different positions and then argued in a similar fashion to the argument for double edges in pairings described above. This gave an asymptotic formula for  $t_{n,m}$ , and similarly for other strictly balanced subgraphs, when  $m = n^{1+\theta}$  for  $\theta$  quite small. A result for  $t_{n,m}$ , with much larger  $\theta$  was achieved by the author in [Wormald \[1996\]](#), using a different version of the perturbation method. Instead of moving edges to new positions, one of the basic operations considered

was to add a triangle at a random position. This generally transforms the graph into one with an extra triangle, but sometimes two or more new ones can be created. The number of copies of two triangles sharing an edges is also recorded, as well as numbers of some other small clusters of triangles sharing edges. In this case vertices of  $F$  in [Lemma 3.1](#) are vectors specifying the numbers of each type of cluster, and each vertex is adjacent to several others, being those adjacent in the integer lattice. The estimate for a ratio between adjacent vertices contains lower order terms involving similar ratios for other nearby vertices. The ratio for vectors adjacent in the vector component corresponding to a given cluster is estimated using a “switching” operation in which a copy of the corresponding cluster is added at random. Fixed points are not used explicitly, but estimates of lower and upper bounds on ratios are iterated. These arguments are complicated by a careful induction that yields the required initial bounds on the ratios. [Stark and Wormald \[2016\]](#) strengthened the method and also extended it to all strictly balanced subgraphs.

(ii) Degree switchings for counting graphs.

This is a rough outline of the argument in [Liebenau and Wormald \[2017\]](#). Suppose we have a random graph  $G$  with degree sequence  $\mathbf{d} - \mathbf{e}_b$  where  $\mathbf{e}_v$  denotes the elementary unit vector with 1 in its  $v^{\text{th}}$  coordinate. Pick a random edge  $e$  incident with vertex  $a$ , and with  $v$  denoting the other end of  $e$ , remove  $e$  and add the edge  $bv$ . Let  $B(a, b, \mathbf{d} - \mathbf{e}_b)$  denote the probability of the “bad” event that a loop or multiple edge is produced. If this event fails, the graph now has degree sequence  $\mathbf{d} - \mathbf{e}_a$ . We call this a *degree switching*. Simple counting shows that

$$(3-2) \quad R(a, b; \mathbf{d}) := \frac{g(\mathbf{d} - \mathbf{e}_a)}{g(\mathbf{d} - \mathbf{e}_b)} = \frac{d_a}{d_b} \cdot \frac{1 - B(a, b, \mathbf{d} - \mathbf{e}_b)}{1 - B(b, a, \mathbf{d} - \mathbf{e}_a)}.$$

Let  $P_{av}(\mathbf{d})$  denote the probability that edge  $av$  occurs in a random graph with degree dequence  $\mathbf{d}$ . With a little work, we can express  $B(a, b, \mathbf{d} - \mathbf{e}_b)$  using a combination of such probabilities, resulting in a formula for  $R(a, b; \mathbf{d})$  in terms of the  $P_{av}(\mathbf{d})$  for various  $a, v$  and  $\mathbf{d}$ .

By noting  $d_a = \sum_v P_{av}$ , we can also obtain

$$P_{av}(\mathbf{d}) = d_v \left( \sum_{b \in V \setminus \{a\}} R(b, a; \mathbf{d} - \mathbf{e}_v) \frac{1 - P_{bv}(\mathbf{d} - \mathbf{e}_b - \mathbf{e}_v)}{1 - P_{av}(\mathbf{d} - \mathbf{e}_a - \mathbf{e}_v)} \right)^{-1}$$

where  $V = [n]$  is the set of all vertices. Iterating these two formulae produces a sequence of approximations to the functions  $P$  and  $R$  that converges, and it is convenient to consider, as described above, fixed points of the operators defined by these equations. Then the argument becomes one of proving that the limits of the convergent solutions are close to the fixed points of the operators, and that these are close to the true values of the ratios.



The required initial bounds on the true ratios were obtained in this case by a primitive analysis of a switching-type operation similar to that in [Figure 1](#).

[Lemma 3.1](#) was applied here, with  $\Omega$  being the set of sequences on non-negative integers of sum  $2m$ . The probability of  $\mathbf{d}$  in  $\mathcal{P}'$  is proportional to  $g(\mathbf{d})$ , and in  $\mathcal{P}$  is proportional to the target formula, i.e. the right hand side of (1-2).

(iii) A related approach for graphs.

While work on [Liebenau and Wormald \[2017\]](#) was under way, [Burstein and Rubin \[2015\]](#) (mentioned above) considered ratios of ‘adjacent’ degree sequences, and iterated their formulae to obtain higher accuracy ratios. Having estimated the ratios, they used them to compare  $g(\mathbf{d})$  with some known value  $g(\mathbf{d}')$ . In terms of the ratios, they give a formula for  $g(\mathbf{d})$ , however it was not explicit enough in the general case to enable the derivation of results as simple as (1-2). In particular, they did not extend the range of validity of (1-2) past what was known at the time.

(iv) Counting Latin rectangles.

All results before [Leckey, Liebenau, and Wormald \[n.d.\]](#) considered adding a random row to a valid  $(k-1) \times n$  Latin rectangle, and estimating the probability that the new row causes no conflicts with the previous rows. (The paper [McKay and Wormald \[1991b\]](#) is almost an exception, since all rows are considered at random, and switching operations were used to generate a random Latin rectangle. A similar analysis on numbers would have yielded the asymptotic formula for  $k = o(n^{1/3})$  which was already known.) The approach in [Leckey, Liebenau, and Wormald \[n.d.\]](#) is different. As mentioned above, enumerating  $k \times n$  Latin rectangles is equivalent to counting  $k$ -regular bipartite graphs with vertex parts  $V_1$  and  $V_2$ , both of cardinality  $n$ , which have been properly  $k$ -edge-coloured. Now consider all bipartite graphs with vertex parts  $V_1$  and  $V_2$ , with  $n$  edges of each of  $k$  colours. The *colour-degree sequence* is the array of numbers  $d_{ij}$ ,  $i = 1, \dots, n$ ,  $j = 1, \dots, k$  such that  $d_{ij}$  is the number of edges of colour  $j$  incident with vertex  $i$ . Then the problem becomes one of enumerating those graphs with the all-1’s degree sequence. Equivalently, consider the probability of such a colour-degree sequence arising when an edge-coloured bipartite graph with  $n$  edges of each colour is chosen uniformly at random. The problem is simplified somewhat by restricting to those graphs in which the degrees on  $V_2$  are all 1; there is a simple model for selecting a random graph subject to this condition. To solve it, degree-switchings are applied to the vertices in  $V_1$ , and the perturbation method is applied as in [Lemma 3.1](#) with  $\mathcal{P}$  being a certain multinomial probability distribution. The number of variables (colour-degrees) is much larger than the graph case discussed above, with the consequence that the argument “only” succeeds for  $k = o(n/\log n)$ .

## 4 Applications of the enumeration results

**4.1 Models for joint degree distribution.** It was shown in [McKay and Wormald \[1997\]](#) how to use [Proposition 1.2](#) to transfer properties of a sequence of independent binomial variables to the degree sequence of the random graph  $\mathcal{G}(n, p)$  (in the range of  $p$  for which the proposition holds). In a follow-up paper, the authors will use this to obtain results on the order statistics of the degree sequence, and also the distribution of the number of vertices of given degree, a problem discussed at some length by [Barbour, Holst, and Janson \[1992\]](#).

Similar models for bipartite graphs and loopless directed graphs have been investigated in the dense range by [McKay and Skerman \[2016\]](#), who observed that once we have the results on enumeration that are now supplied by [Liebenau and Wormald \[2017\]](#), these models can presumably be extended to all interesting ranges of density.

**4.2 Subgraphs and properties of random graphs.** The methods of counting graphs can frequently be modified to include certain edges as specified (or forbidden). This lets us estimate moments of random variables that count copies of given subgraphs. Armed with such formulae, if they are simple enough, we can derive properties of the subgraph counts of random graphs with given degrees.

Examples include perfect matchings in regular graphs (see [Bollobás and McKay \[1986\]](#)). [Robinson and Wormald \[1994\]](#) used this approach (and a new technique for analysing variance) to show that a random  $d$ -regular graph is highly likely to have a Hamilton cycle. Enumeration results were used similarly to prove various properties of random regular graphs of high degree by [Krivelevich, Sudakov, Vu, and Wormald \[2001\]](#). There many other examples.

**4.3 Random generation.** The uniform generation of random graphs with given degree sequence, and related objects, has statistical uses (see [Blitzstein and Diaconis \[2010\]](#) for example). Methods and results of enumeration can be useful, or sometimes adapted, to this problem. See [Wormald \[1984\]](#) for examples with exact enumeration; the applicability of the configuration model and its earlier bipartite versions are obvious. [McKay and Wormald \[1990b\]](#) used switchings to generate random regular graphs uniformly, and this was extended by [Gao and Wormald \[2017\]](#).

**4.4 Relations to graphs *without* specified degrees.** [Kim and Vu \[2004\]](#) used asymptotic enumeration results, amongst other things, to show that a random  $d$ -regular graph is “sandwiched” in between two random graphs  $\mathcal{G}(n, m)$  for two different values of  $m$  close to  $d$ , as long as  $d$  does not grow too slowly or too quickly. Their result was extended by

Dudek, A. Frieze, Ruciński, and Šileikis [2017] to larger  $d$ , which might have been easier had the results in Liebenau and Wormald [2017] been available.

One approach to obtaining results for graphs with a given number of edges is to prove results for given degree sequences and then essentially sum over all degree sequences. This was used for instance by Pittel and Wormald [2005] to obtain the distribution of the size of the 2-core in the random graphs  $\mathcal{G}(n, p)$  and  $\mathcal{G}(n, m)$ . Asymptotic enumeration results are used heavily in such arguments.

## 5 Final questions

An obvious question that might soon be within reach, given the new methods arising in the past year or so, is to find necessary and sufficient conditions on  $\mathbf{d}$  for (1-2) to hold.

In McKay [2010], McKay points out that the asymptotic formula for the number of bipartite graphs with given degree sequence is unknown when the two vertex parts have very different cardinalities. The method in Liebenau and Wormald [2017] goes some way towards alleviating this defect in our knowledge, but further work can still be done. In particular, Canfield and McKay [2005] suggest a formula of Good and Crook that might be valid to within a constant factor in all cases for the biregular case.

One problem that arose in some discussions with Boris Pittel and is still unsolved, is to construct a nice model for the degree sequence of a random connected graph with  $n$  vertices and  $m$  edges, in the sparse range, particularly when  $m = \Theta(n)$ .

The asymptotic number of acyclic digraphs with  $n$  vertices and  $cn$  edges is essentially unknown. Can we nevertheless find an asymptotic formula for the number of acyclic digraphs with given in- and out-degree sequence, with  $cn$  edges?

**Acknowledgments.** I would like to thank all my coauthors on this topic, and Brendan McKay in particular for highly productive collaborations at various stages of the work and also for checking an early draft of this article.

## References

- N. Alon and J. H. Spencer (2000). *The probabilistic method*. Second. Wiley-Interscience Series in Discrete Mathematics and Optimization. With an appendix on the life and work of Paul Erdős. Wiley-Interscience [John Wiley & Sons], New York, pp. xviii+301. MR: [1885388](#) (cit. on p. [3263](#)).
- A. D. Barbour, L. Holst, and S. Janson (1992). *Poisson approximation*. Vol. 2. Oxford Studies in Probability. Oxford Science Publications. The Clarendon Press, Oxford University Press, New York, pp. x+277. MR: [1163825](#) (cit. on p. [3279](#)).

- A. Barvinok and J. A. Hartigan (2013). “The number of graphs and a random graph with a given degree sequence”. *Random Structures Algorithms* 42.3, pp. 301–348. MR: [3039682](#) (cit. on p. [3272](#)).
- A. Békéssy, P. Békéssy, and J. Komlós (1972). “Asymptotic enumeration of regular matrices”. *Studia Sci. Math. Hungar.* 7, pp. 343–353. MR: [0335342](#) (cit. on pp. [3268](#), [3269](#), [3273](#)).
- E. A. Bender (1974). “The asymptotic number of non-negative integer matrices with given row and column sums”. *Discrete Math.* 10, pp. 217–223. MR: [0389621](#) (cit. on p. [3268](#)).
- E. A. Bender and E. R. Canfield (1978). “The asymptotic number of labeled graphs with given degree sequences”. *J. Combinatorial Theory Ser. A* 24.3, pp. 296–307. MR: [0505796](#) (cit. on p. [3268](#)).
- J. Blitzstein and P. Diaconis (2010). “A sequential importance sampling algorithm for generating random graphs with prescribed degrees”. *Internet Math.* 6.4, pp. 489–522. MR: [2809836](#) (cit. on p. [3279](#)).
- B. Bollobás (1980). “A probabilistic proof of an asymptotic formula for the number of labelled regular graphs”. *European J. Combin.* 1.4, pp. 311–316. MR: [595929](#) (cit. on p. [3269](#)).
- (1981). “The independence ratio of regular graphs”. *Proc. Amer. Math. Soc.* 83.2, pp. 433–436. MR: [624948](#) (cit. on p. [3269](#)).
- (2001). *Random graphs*. Second. Vol. 73. Cambridge Studies in Advanced Mathematics. Cambridge University Press, Cambridge, pp. xviii+498. MR: [1864966](#) (cit. on p. [3264](#)).
- B. Bollobás and B. D. McKay (1986). “The number of matchings in random regular graphs and bipartite graphs”. *J. Combin. Theory Ser. B* 41.1, pp. 80–91. MR: [854605](#) (cit. on pp. [3270](#), [3279](#)).
- D. Burstein and J. Rubin (Nov. 2015). “Degree switching and partitioning for enumerating graphs to arbitrary orders of accuracy”. arXiv: [1511.03738](#) (cit. on pp. [3271](#), [3278](#)).
- E. R. Canfield and B. D. McKay (2005). “Asymptotic enumeration of dense 0-1 matrices with equal row sums and equal column sums”. *Electron. J. Combin.* 12, Research Paper 29, 31. MR: [2156683](#) (cit. on p. [3280](#)).
- W. Y. C. Chen and J. D. Louck (1999). “Enumeration of cubic graphs by inclusion-exclusion”. *J. Combin. Theory Ser. A* 86.1, pp. 151–157. MR: [1682968](#) (cit. on p. [3272](#)).
- C. Cooper, A. Frieze, B. Reed, and O. Riordan (2002). “Random regular graphs of non-constant degree: independence and chromatic number”. *Combin. Probab. Comput.* 11.4, pp. 323–341. MR: [1918719](#).
- M. Drmota (2009). *Random trees*. An interplay between combinatorics and probability. SpringerWienNewYork, Vienna, pp. xviii+458. MR: [2484382](#) (cit. on p. [3273](#)).

- A. Dudek, A. Frieze, A. Ruciński, and M. Šileikis (2017). “[Embedding the Erdős-Rényi hypergraph into the random regular hypergraph and Hamiltonicity](#)”. *J. Combin. Theory Ser. B* 122, pp. 719–740. MR: [3575225](#) (cit. on p. [3280](#)).
- P. Erdős and I. Kaplansky (1946). “[The asymptotic number of Latin rectangles](#)”. *Amer. J. Math.* 68, pp. 230–236. MR: [0015356](#) (cit. on p. [3273](#)).
- C. J. Everett and P. R. Stein (1971/72). “[The asymptotic number of integer stochastic matrices](#)”. *Discrete Math.* 1.1, pp. 55–72. MR: [0389620](#) (cit. on p. [3268](#)).
- A. Frieze (1992). “On small subgraphs of random graphs”. In: *Random graphs, Vol. 2 (Poznań, 1989)*. Wiley-Intersci. Publ. Wiley, New York, pp. 67–90. MR: [1166608](#) (cit. on p. [3276](#)).
- A. M. Frieze and T. Łuczak (1992). “[On the independence and chromatic numbers of random regular graphs](#)”. *J. Combin. Theory Ser. B* 54.1, pp. 123–132. MR: [1142268](#).
- P. Gao, Y. Su, and N. Wormald (2012). “[Induced subgraphs in sparse random graphs with given degree sequences](#)”. *European J. Combin.* 33.6, pp. 1142–1166. MR: [2904981](#).
- P. Gao and N. Wormald (2016). “[Enumeration of graphs with a heavy-tailed degree sequence](#)”. *Adv. Math.* 287, pp. 412–450. MR: [3422681](#) (cit. on pp. [3270](#), [3271](#)).
- (2017). “[Uniform generation of random regular graphs](#)”. *SIAM J. Comput.* 46.4, pp. 1395–1427. MR: [3686817](#) (cit. on p. [3279](#)).
- C. D. Godsil and B. D. McKay (1990). “[Asymptotic enumeration of Latin rectangles](#)”. *J. Combin. Theory Ser. B* 48.1, pp. 19–44. MR: [1047551](#) (cit. on p. [3273](#)).
- I. P. Goulden, D. M. Jackson, and J. W. Reilly (1983). “[The Hammond series of a symmetric function and its application to  \$P\$ -recursiveness](#)”. *SIAM J. Algebraic Discrete Methods* 4.2, pp. 179–193. MR: [699770](#) (cit. on p. [3272](#)).
- C. Greenhill and B. D. McKay (2013). “[Asymptotic enumeration of sparse multigraphs with given degrees](#)”. *SIAM J. Discrete Math.* 27.4, pp. 2064–2089. MR: [3141749](#) (cit. on p. [3273](#)).
- M. Isaev and B. D. McKay (Apr. 2016). “[Complex martingales and asymptotic enumeration](#)”. arXiv: [1604.08305](#) (cit. on p. [3272](#)).
- S. Janson (2009). “[The probability that a random multigraph is simple](#)”. *Combin. Probab. Comput.* 18.1-2, pp. 205–225. MR: [2497380](#) (cit. on pp. [3270](#), [3271](#)).
- (2014). “[The probability that a random multigraph is simple. II](#)”. *J. Appl. Probab.* 51A. Celebrating 50 Years of The Applied Probability Trust, pp. 123–137. MR: [3317354](#) (cit. on pp. [3270](#), [3271](#)).
- J. H. Kim and V. H. Vu (2004). “[Sandwiching random graphs: universality between random graph models](#)”. *Adv. Math.* 188.2, pp. 444–469. MR: [2087234](#) (cit. on p. [3279](#)).
- M. Krivelevich, B. Sudakov, V. H. Vu, and N. Wormald (2001). “[Random regular graphs of high degree](#)”. *Random Structures Algorithms* 18.4, pp. 346–363. MR: [1839497](#) (cit. on p. [3279](#)).

- G. Kuperberg, S. Lovett, and R. Peled (2017). “Probabilistic existence of regular combinatorial structures”. *Geom. Funct. Anal.* 27.4, pp. 919–972. MR: [3678505](#) (cit. on p. [3273](#)).
- K. Leckey, A. Liebenau, and N. Wormald (n.d.). *The asymptotic number of Latin rectangles*. Preprint (cit. on pp. [3273](#), [3276](#), [3278](#)).
- A. Liebenau and N. Wormald (Feb. 2017). “Asymptotic enumeration of graphs by degree sequence, and the degree sequence of a random graph”. arXiv: [1702.08373](#) (cit. on pp. [3264](#), [3267](#), [3271–3274](#), [3277–3280](#)).
- B. D. McKay (1981). “Subgraphs of random graphs with specified degrees”. *Congr. Numer.* 33, pp. 213–223. MR: [681916](#) (cit. on p. [3274](#)).
- (1984). “Asymptotics for 0-1 matrices with prescribed line sums”. In: *Enumeration and design (Waterloo, Ont., 1982)*. Academic Press, Toronto, ON, pp. 225–238. MR: [782316](#) (cit. on p. [3270](#)).
  - (1985). “Asymptotics for symmetric 0-1 matrices with prescribed row sums”. *Ars Combin.* 19.A, pp. 15–25. MR: [790916](#) (cit. on pp. [3270](#), [3274](#)).
  - (2010). “Subgraphs of random graphs with specified degrees”. In: *Proceedings of the International Congress of Mathematicians. Volume IV*. Hindustan Book Agency, New Delhi, pp. 2489–2501. MR: [2827981](#) (cit. on p. [3280](#)).
  - (2011). “Subgraphs of dense random graphs with specified degrees”. *Combin. Probab. Comput.* 20.3, pp. 413–433. MR: [2784635](#).
- B. D. McKay and F. Skerman (2016). “Degree sequences of random digraphs and bipartite graphs”. *J. Comb.* 7.1, pp. 21–49. MR: [3436194](#) (cit. on p. [3279](#)).
- B. D. McKay and N. Wormald (1990a). “Asymptotic enumeration by degree sequence of graphs of high degree”. *European J. Combin.* 11.6, pp. 565–580. MR: [1078713](#) (cit. on pp. [3264](#), [3265](#), [3271](#), [3272](#)).
- (1990b). “Uniform generation of random regular graphs of moderate degree”. *J. Algorithms* 11.1, pp. 52–67. MR: [1041166](#) (cit. on p. [3279](#)).
  - (1991a). “Asymptotic enumeration by degree sequence of graphs with degrees  $o(n^{1/2})$ ”. *Combinatorica* 11.4, pp. 369–382. MR: [1137769](#) (cit. on pp. [3266](#), [3270](#), [3273](#), [3275](#)).
  - (1991b). “Uniform generation of random Latin rectangles”. *J. Combin. Math. Combin. Comput.* 9, pp. 179–186. MR: [1111853](#) (cit. on pp. [3276](#), [3278](#)).
  - (1997). “The degree sequence of a random graph. I. The models”. *Random Structures Algorithms* 11.2, pp. 97–117. MR: [1610253](#) (cit. on pp. [3265](#), [3279](#)).
- M. P. Mineev and A. I. Pavlov (1976). “The number of  $(0,1)$ -matrices with given sums over the rows and columns”. *Dokl. Akad. Nauk SSSR* 230.2, pp. 271–274. MR: [0414381](#) (cit. on p. [3268](#)).
- J. W. Moon (1970). *Counting labelled trees*. Vol. 1969. From lectures delivered to the Twelfth Biennial Seminar of the Canadian Mathematical Congress (Vancouver).

- Canadian Mathematical Congress, Montreal, Que., pp. x+113. MR: [0274333](#) (cit. on p. [3264](#)).
- P. E. O’Neil (1969). “Asymptotics and random matrices with row-sum and column-sum restrictions”. *Bull. Amer. Math. Soc.* 75, pp. 1276–1282. MR: [0257116](#) (cit. on p. [3268](#)).
- B. Pittel and N. Wormald (2005). “Counting connected graphs inside-out”. *J. Combin. Theory Ser. B* 93.2, pp. 127–172. MR: [2117934](#) (cit. on p. [3280](#)).
- R. C. Read (1958). “Some enumeration problems in graph theory”. PhD thesis. University of London (University College of the West Indies) (cit. on pp. [3267](#), [3272](#)).
- (1960). “The enumeration of locally restricted graphs. II”. *J. London Math. Soc.* 35, pp. 344–351. MR: [0140443](#) (cit. on p. [3272](#)).
- R. C. Read and N. Wormald (1980). “Number of labeled 4-regular graphs”. *J. Graph Theory* 4.2, pp. 203–212. MR: [570354](#) (cit. on p. [3272](#)).
- R. W. Robinson and N. Wormald (1994). “Almost all regular graphs are Hamiltonian”. *Random Structures Algorithms* 5.2, pp. 363–374. MR: [1262985](#) (cit. on p. [3279](#)).
- D. Stark and N. Wormald (Aug. 2016). “The probability of nonexistence of a subgraph in a moderately sparse random graph”. arXiv: [1608.05193](#) (cit. on p. [3277](#)).
- C. M. Stein (1978). “Asymptotic evaluation of the number of Latin rectangles”. *J. Combin. Theory Ser. A* 25.1, pp. 38–49. MR: [499035](#) (cit. on pp. [3273](#), [3275](#), [3276](#)).
- N. Wormald (1978). “Some Problems in the Enumeration of Labelled Graphs”. PhD thesis (cit. on p. [3269](#)).
- (1984). “Generating random regular graphs”. *J. Algorithms* 5.2, pp. 247–280. MR: [744493](#) (cit. on p. [3279](#)).
- (1996). “The perturbation method and triangle-free random graphs”. In: *Proceedings of the Seventh International Conference on Random Structures and Algorithms (Atlanta, GA, 1995)*. Vol. 9. 1-2, pp. 253–270. MR: [1611697](#) (cit. on pp. [3273](#), [3276](#)).
- K. Yamamoto (1951). “On the asymptotic number of Latin rectangles”. *Jap. J. Math.* 21, 113–119 (1952). MR: [0051203](#) (cit. on p. [3273](#)).

Received 2017-12-04.

NICHOLAS WORMALD  
SCHOOL OF MATHEMATICAL SCIENCES  
MONASH UNIVERSITY  
AUSTRALIA  
[nick.wormald@monash.edu](mailto:nick.wormald@monash.edu)

# UNDERSTANDING QUANTUM ALGORITHMS VIA QUERY COMPLEXITY

ANDRIS AMBAINIS

## Abstract

Query complexity is a model of computation in which we have to compute a function  $f(x_1, \dots, x_N)$  of variables  $x_i$  which can be accessed via queries. The complexity of an algorithm is measured by the number of queries that it makes. Query complexity is widely used for studying quantum algorithms, for two reasons. First, it includes many of the known quantum algorithms (including Grover's quantum search and a key subroutine of Shor's factoring algorithm). Second, one can prove lower bounds on the query complexity, bounding the possible quantum advantage. In the last few years, there have been major advances on several longstanding problems in the query complexity. In this talk, we survey these results and related work, including:

- the biggest quantum-vs-classical gap for partial functions (a problem solvable with 1 query quantumly but requiring  $\Omega(\sqrt{N})$  queries classically);
- the biggest quantum-vs-deterministic and quantum-vs-probabilistic gaps for total functions (for example, a problem solvable with  $M$  queries quantumly but requiring  $\tilde{\Omega}(M^{2.5})$  queries probabilistically);
- the biggest probabilistic-vs-deterministic gap for total functions (a problem solvable with  $M$  queries probabilistically but requiring  $\tilde{\Omega}(M^2)$  queries deterministically);
- the bounds on the gap that can be achieved for subclasses of functions (for example, symmetric functions);
- the connections between query algorithms and approximations by low-degree polynomials.



# 1 Introduction

Quantum computers open new possibilities for computing, by being able to solve problems that are considered intractable classically. The most famous example is factoring large numbers which is thought to require  $\Omega(2^{n^c})$  time classically but is efficiently solvable by a quantum computer, due to Shor's quantum algorithm [Shor \[1997\]](#). Another example is simulating quantum physical systems which is thought to require  $\Omega(2^n)$  time classically but is also solvable in polynomial time quantumly [Cirac and Zoller \[2012\]](#) and [Georgescu, Ashhab, and Nori \[2014\]](#).

This naturally leads to a question: how large is the advantage of quantum computers? Can we put limits on it?

In the Turing machine model, we have  $BQTIME(f(n)) \subseteq \cup_c TIME(2^{cf(n)})$  where  $TIME$  and  $BQTIME$  denote the classes of problems that are solvable by deterministic or quantum Turing machines within the respective time bound. However, it is very difficult to prove unconditional separations between complexity classes and we cannot even show that  $BQTIME(f(n))$  is larger than  $TIME(f(n))$ .

For this reason, the power of quantum computers is often studied in the query model (also known as the decision tree model [Buhrman and de Wolf \[2002\]](#)). In this model, we have to compute a function  $f(x_1, \dots, x_N)$  of an input  $(x_1, \dots, x_N)$ , with  $x_i$  accessible via queries to a black box that, given  $i$ , outputs  $x_i$ . The complexity is measured by the number of queries that an algorithm makes.

The query model is very interesting in the quantum case because it captures most of the known quantum algorithms. Some of the problems that can be described in it are:

**Search.** Given black box access to  $x_1, \dots, x_N \in \{0, 1\}$ , determine whether there exists  $i : x_i = 1$  (or find such  $i$ ).

Search requires  $N$  queries classically but can be solved with  $O(\sqrt{N})$  queries quantumly [Grover \[1996\]](#). It can be viewed as a black box model for a generic exhaustive search problem where one has to check  $N$  possibilities (without any information which of those  $N$  possibilities are more likely) and implies quantum speedups for a variety of problems (for example, a quadratic quantum speedup over the best probabilistic algorithm for 3-SAT [Ambainis \[2004\]](#)).

**Period-finding.** Given black box access to  $x_1, \dots, x_N \in [M]$ , determine the smallest  $r$  such that  $x_i = x_{i+r}$  for all  $i$  (and  $x_i \neq x_{i+q}$  for all  $i$  and  $q < r$ ), under a promise that such  $r$  exists and is smaller than  $c\sqrt{N}$  for some  $c > 0$ .

Period-finding is solvable with  $O(1)$  queries quantumly and requires  $\Omega(\frac{N^{1/4}}{\log N})$  queries classically [Shor \[1997\]](#) and [Chakraborty, Fischer, Matsliah, and de Wolf \[2010\]](#). It is at the heart of Shor's factoring algorithm [Shor \[1997\]](#) which consists of a classical reduction from factoring to period-finding and a quantum algorithm for period-finding.

**Element distinctness.** Given black box access to  $x_1, \dots, x_N \in [M]$ , determine if there are  $i, j : i \neq j$  such that  $x_i = x_j$ .

Element distinctness requires  $N$  queries classically and  $\Theta(N^{2/3})$  queries quantumly [Ambainis \[2007\]](#) and [Aaronson and Shi \[2004\]](#). It is related to black box models of algorithms for breaking collision-resistant hash functions (an important cryptographic primitive). The quantum algorithm for element distinctness is also useful as a subroutine for other quantum algorithms, from checking matrix products [Buhrman and Špalek \[2006\]](#) to solving typical instances of subset sum (which is also important for cryptography) [Bernstein, Jeffery, Lange, and Meurer \[2013\]](#).

Many other quantum query algorithms are known, as can be seen from Quantum Algorithms Zoo (a website collecting information about all quantum algorithms [Jordan \[n.d.\]](#)). From a complexity-theoretic perspective, the query model is very interesting because it allows to prove lower bounds on quantum algorithms and it is often possible to characterize the quantum advantage within a big-O factor.

The current survey is focused on characterizing the maximum possible quantum advantage in the query model, for different types of computational tasks. Let  $Q(f)$  and  $R(f)$  denote the number of queries for the best quantum and randomized algorithm, respectively. For partial Boolean functions, we describe a gap of  $Q(f) = 1$  vs  $R(f) = \Omega(\sqrt{N}/\log N)$  [Aaronson and Ambainis \[2015\]](#). For total functions, the biggest known gap is much smaller:  $R(f) = \tilde{\Omega}(Q^{2.5}(f))$  [Aaronson, Ben-David, and Kothari \[2016\]](#). Imposing symmetry constraints on  $f$  also decreases the maximum possible gap.

As a side result, this research has led to new results on classical query algorithms. This includes solutions to two well known problems in the classical query complexity which had been open for about 30 years (such as determining the maximum gap between randomized and deterministic query complexities [Saks and Wigderson \[1986\]](#) and [Arunachalam, Briët, and Palazuelos \[2017\]](#)). We describe those developments, as well.

## 2 Computational Models

**2.1 Deterministic, randomized and quantum query algorithms.** We now formally define the models of query complexity that we use. We consider computing a function  $f(x_1, \dots, x_N)$  of variables  $x_i$ . By default, we assume that the variables  $x_i$  are  $\{0, 1\}$ -valued. (If we consider  $x_i$  with values in a larger set, this is explicitly indicated.) The function  $f(x_1, \dots, x_N)$  can be either a total function (defined on the entire  $\{0, 1\}^N$ ) or a partial function (defined on a subset of  $\{0, 1\}^N$ ).

**Deterministic algorithms.** Deterministic query algorithms are often called *decision trees*, because they can be described by a tree (as in [Figure 1](#)). At each node of this tree, we have the name of a variable that is asked if the algorithm gets to this node. Depending

on the outcome of the query, the algorithm proceeds to the  $x_i = 0$  child or to the  $x_i = 1$  child of the node. If the algorithm gets to a leaf of the tree, it outputs the value of the function listed at this leaf.

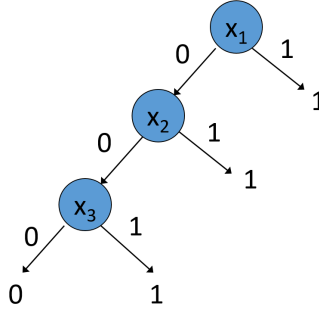


Figure 1: Example of a decision tree

The complexity of an algorithm  $\mathcal{Q}$  is the maximum number of queries that it can make. Deterministic query complexity  $D(f)$  is the smallest complexity of a deterministic  $\mathcal{Q}$  which outputs  $f(x_1, \dots, x_N)$  if the queries are answered according to  $(x_1, \dots, x_N)$ , whenever  $f(x_1, \dots, x_N)$  is defined.

**Randomized algorithms.** In a randomized query algorithm, the algorithm may choose the variable  $x_i$  for the next query randomly from some probability distribution.

Randomized algorithms are usually studied either in the zero-error setting or in the bounded error setting. In the zero-error setting, the algorithm is required to output  $f(x_1, \dots, x_N)$  with probability at least  $1/2$  and may output "don't know" otherwise but must not output a value that is different from  $f(x_1, \dots, x_N)$ . In the bounded-error setting, algorithm is required to output  $f(x_1, \dots, x_N)$  with probability at least  $2/3$  and may output anything otherwise. In both cases, the requirement has to be satisfied for every  $(x_1, \dots, x_N)$  for which  $f(x_1, \dots, x_N)$  is defined.

The complexity of an algorithm  $\mathcal{Q}$  is measured by the largest number of queries that is made by  $\mathcal{Q}$ , for the worst choice of  $(x_1, \dots, x_N)$  and the worst random choices of  $\mathcal{Q}$ .  $R_0(f)$  and  $R_2(f)$  are the smallest complexities of a zero-error randomized and a bounded error randomized algorithm for  $f$ , respectively. (Alternatively, one can define randomized query complexity via the expected number of queries for the worst case  $(x_1, \dots, x_N)$  but this changes the complexities  $R_0$  and  $R_2$  by at most a constant factor.)

**Quantum algorithms.** Unlike in the probabilistic case, different branches of a quantum algorithm can recombine at a later stage. For this reason, a quantum query algorithm cannot be described by a tree.

Instead, a quantum query algorithm is defined by an initial state  $|\psi_{start}\rangle$  and transformations  $U_0, Q, U_1, \dots, Q, U_T$ . The initial state  $|\psi_{start}\rangle$  and transformations  $U_i$  are independent of  $x_1, \dots, x_N$ .  $Q$  are the queries - transformations of a fixed form that depend on  $x_i$ 's. The algorithm consists of performing  $U_0, Q, U_1, \dots, Q, U_T$  on  $|\psi_{start}\rangle$  and measuring the result (as shown in Figure 2). The algorithm computes  $f$  if, for every  $(x_1, \dots, x_N)$  for which  $f(x_1, \dots, x_N)$  is defined, this measurement produces  $f(x_1, \dots, x_N)$ .

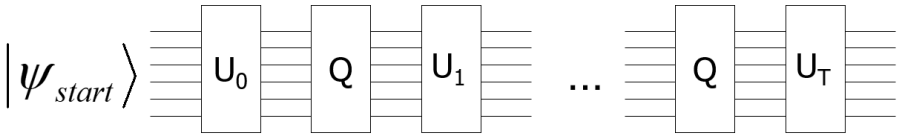


Figure 2: Structure of a quantum query algorithm

To define the model more precisely, we must define the notions of a quantum state, a transformation, and a measurement. (For more details on these notions, we refer the reader to the book [Nielsen and Chuang \[2000\]](#).) The state space of a quantum algorithm is a complex vector space of dimension  $d$  (where  $d$  can be chosen by the designer of the algorithm). Let  $|1\rangle, \dots, |d\rangle$  be an orthonormal basis for this vector space. A quantum state is a vector

$$|\psi\rangle = \alpha_1|1\rangle + \dots + \alpha_d|d\rangle = \begin{pmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_d \end{pmatrix}$$

of unit length (i.e. satisfying  $\sum_i |\alpha_i|^2 = 1$ ). A unitary transformation is a linear transformation on  $|\psi\rangle$  that preserves the length of  $|\psi\rangle$ . The principles of quantum mechanics allow to perform any unitary  $U$  on a quantum state.

A measurement is the way of obtaining information from a quantum state. Measuring a state  $|\psi\rangle$  with respect to  $|1\rangle, \dots, |d\rangle$  yields the result  $i$  with probability  $|\alpha_i|^2$ .

To define a quantum query algorithm, we allow the starting state  $|\psi_{start}\rangle$  to be an arbitrary quantum state.  $U_i$ 's can be arbitrary unitary transformations that do not depend

on  $x_1, \dots, x_N$ .  $Q$  is the query transformation, defined in a following way<sup>1</sup>. We rename the basis states from  $|1\rangle, \dots, |d\rangle$  to  $|i, j\rangle$  with  $i \in \{0, 1, \dots, N\}$  and  $j \in [d_i]$  for some  $d_i$  and define

$$Q|0, j\rangle = |0, j\rangle \text{ for all } j,$$

$$Q|i, j\rangle = \begin{cases} |i, j\rangle & \text{if } x_i = 0 \\ -|i, j\rangle & \text{if } x_i = 1 \end{cases}.$$

It can be argued that this is a natural quantum counterpart of a probabilistic query in which we choose  $i$  according to a probability distribution and get the corresponding  $x_i$ .

After the last transformation, the state of the algorithm is measured w.r.t.  $|1\rangle, \dots, |d\rangle$  and the result is transformed into the answer of the algorithm according to a predefined rule. (For example, if the answer should be  $\{0, 1\}$ -valued, we could take the first bit of the measurement result  $i$  as the answer.)

Two most frequently considered types of quantum query algorithms are *exact* and *bounded error* algorithms. A quantum query algorithm computes  $f$  exactly if its answer is always the same as  $f(x_1, \dots, x_N)$ , whenever  $f$  is defined. A quantum query algorithm  $Q$  computes  $f$  with bounded error, if for every  $(x_1, \dots, x_N)$ , for which  $f(x_1, \dots, x_N)$  is defined, the probability that  $Q$  outputs  $f(x_1, \dots, x_N)$  as the answer is at least  $2/3$ .  $Q_E(f)$  and  $Q_2(f)$  are the smallest numbers of queries in quantum algorithms that compute  $f$  exactly and with bounded error, respectively.

**2.2 Quantities that are related to query complexity.** In this section, we define several quantities that provide upper and lower bounds on different query complexities. Using them, we can prove bounds on the maximum gaps between query complexity measures (for example, that  $D(f) = O(R_2^3(f))$  [Nisan \[1991\]](#) and  $D(f) = O(Q_2^6(f))$  [Beals, Buhrman, Cleve, Mosca, and de Wolf \[2001\]](#) for any total Boolean function  $f$ ).

**Block sensitivity.** For an input  $x \in \{0, 1\}^N$  and a subset of variables  $S \subseteq [N]$ ,  $x^{(S)}$  is the input obtained from  $x$  by changing all  $x_i, i \in S$  to opposite values. The block sensitivity  $bs(f)$  is the maximum  $k$  for which there is an input  $x \in \{0, 1\}^N$  and pairwise disjoint subsets  $S_1, \dots, S_k \subseteq [N]$  with  $f(x) \neq f(x^{(S_i)})$  for all  $i \in [k]$ .

Block sensitivity is a lower bound on all the query complexity measures:  $D(f) \geq bs(f)$ ,  $R(f) = \Omega(bs(f))$  [Nisan \[1991\]](#) and  $Q(f) = \Omega(\sqrt{bs(f)})$  [Beals, Buhrman, Cleve, Mosca, and de Wolf \[2001\]](#). It also provides an upper bound on  $D(f)$  for total Boolean functions  $f$ :  $D(f) = O(bs^3(f))$  [Nisan \[1991\]](#). Combining these relations yields  $D(f) = O(R_2^3(f))$  and  $D(f) = O(Q_2^6(f))$  - the best upper bounds on the gap between  $D(f)$  and  $R_2(f)$  or  $Q_2(f)$ .

<sup>1</sup>Since most of this survey considers functions  $f(x_1, \dots, x_N)$  of variables  $x_i \in \{0, 1\}$ , we only give the definition of a query for this case.

**Certificate complexity.** For an input  $x \in \{0, 1\}^N$ , a certificate is a set  $S \subseteq [N]$  with the property that the variables  $x_i, i \in S$  determine the value of  $f(x)$ . (More precisely,  $S \subseteq [N]$  is a certificate on an input  $x$  if, for any  $y \in \{0, 1\}^N$  such that  $x_i = y_i$  for all  $i \in S$ , we have  $f(x) = f(y)$ .)  $C_x(f)$  is the minimum size  $|S|$  of a certificate  $S$  on the input  $x$ . The certificate complexity  $C(f)$  is the maximum of  $C_x(f)$  over all  $x \in \{0, 1\}^N$ .

Certificate complexity provides a better upper bound on  $D(f)$  for total  $f$ :  $D(f) \leq C^2(f)$  [Nisan \[ibid.\]](#). If one could show that  $Q_2(f) = \Omega(\sqrt{C(f)})$ , this would imply  $D(f) = O(Q_2^4(f))$ , improving the best known relation between  $D(f)$  and  $Q_2(f)$ .

**Randomized certificate complexity** [Aaronson \[2008\]](#). For an input  $x$ ,  $RC_x(f)$  is the minimum number of queries in a bounded-error randomized query algorithm that accepts  $x$  and rejects all  $y : f(x) \neq f(y)$ . The randomized certificate complexity  $RC(f)$  is the maximum of  $RC_x(f)$  over all  $x \in \{0, 1\}^N$ .

Unlike for the standard certificate complexity, it is known that  $Q_2(f) = \Omega(\sqrt{RC(f)})$  [Aaronson \[ibid.\]](#). Proving  $D(f) = O(RC^2(f))$  for total  $f$  (which is not known) would also imply  $D(f) = O(Q_2^4(f))$ .

**Polynomial degree.** The exact degree,  $\deg(f)$ , is the degree of the multilinear polynomial  $p(x_1, \dots, x_N)$  which satisfies  $f(x_1, \dots, x_N) = p(x_1, \dots, x_N)$  for all  $(x_1, \dots, x_N)$ . The approximate degree,  $\widetilde{\deg}(f)$ , is the smallest degree of a multilinear polynomial  $p(x_1, \dots, x_N)$   $\{0, 1\}^N$  which satisfies  $|f(x_1, \dots, x_N) - p(x_1, \dots, x_N)| \leq \frac{1}{3}$  for all  $(x_1, \dots, x_N) \in \{0, 1\}^N$ .

Both of these measures also provide lower bounds on quantum query complexity:  $Q_E(f) \geq \frac{\deg(f)}{2}$  and  $Q_2(f) = \Omega(\sqrt{\widetilde{\deg}(f)})$  [Beals, Buhrman, Cleve, Mosca, and de Wolf \[2001\]](#).

### 3 Maximum quantum-classical gap for partial functions

In this section, we consider the question: what is the maximum possible gap between  $Q_2(f)$  and the most general of classical complexities  $R_2(f)$ , for a partial function  $f(x_1, \dots, x_N)$  if we do not place any constraints on  $f$ ?

As we already mentioned, period finding has  $Q_2(f) = O(1)$  and  $R_2(f) = \tilde{\Omega}(\sqrt[4]{N})$ . In the form defined in [Section 1](#), period-finding is not a Boolean function (it has variables  $x_i \in [M]$  instead of Boolean variables). While it is possible to define a Boolean version of period-finding with almost the same gap, there is Boolean function with an even bigger gap:

**Theorem 1.** [Aaronson and Ambainis \[2015\]](#) *There exists  $f$  with  $Q_2(f) = 1$  and  $R_2(f) = \Omega(\sqrt{N} / \log N)$ .*

The function  $f$  is defined as follows [Aaronson and Ambainis \[ibid.\]](#). We have  $N = 2^{n+1}$  variables. For technical convenience, we denote variables  $x_0, \dots, x_{2^n-1}$ ,

$y_0, \dots, y_{2^n-1}$  and assume that the possible values for variables are  $\pm 1$  (instead of 0 and 1). Let  $F$  be the  $2^n \times 2^n$  matrix (with rows and columns indexed by  $a, b \in [0, 2^n - 1]$ ) defined by  $F_{a,b} = \frac{1}{2^{n/2}}(-1)^{a \cdot b}$  where  $a \cdot b = \sum_i a_i b_i$  is the inner product between  $a$  and  $b$  interpreted as  $n$ -bit strings  $a_{n-1} \dots a_0$  and  $b_{n-1} \dots b_0$ . (In terms of quantum computing,  $F = H^{\otimes n}$  where  $H$  is the standard  $2 \times 2$  Hadamard matrix.) We define

$$f(x_0, \dots, y_{2^n-1}) = \begin{cases} 1 & \text{if } \sum_{a,b} F_{a,b} x_a y_b \geq \frac{3}{5} 2^n \\ 0 & \text{if } \sum_{a,b} F_{a,b} x_a y_b \leq \frac{1}{100} 2^n \end{cases}.$$

The thresholds  $\frac{3}{5}$  and  $\frac{1}{100}$  are chosen so that:

- if we choose  $x_i \in \{-1, 1\}$  for  $i \in \{0, \dots, 2^n - 1\}$  uniformly at random and then choose  $y_i = \text{sgn}((Fx)_i)$ , we get  $f = 1$  with a high probability;
- if we choose both  $x_i$  and  $y_j$  uniformly at random from  $\{-1, 1\}$ , we get  $f = 0$  with a high probability.

Thus, by solving  $f$ , we are effectively distinguishing between  $\vec{y} = (y_i)_{i \in [0, 2^n-1]}$  being the vector of signs of  $F\vec{x}$  where  $\vec{x} = (x_i)_{i \in [0, 2^n-1]}$  and  $\vec{y}$  being independently random.

$Q_2(f) = 1$  is shown by a quantum algorithm that generates a quantum state

$$|\psi\rangle = \sum_{i=0}^{2^n-1} \left( \frac{x_i}{\sqrt{2^n}} |0, i\rangle + \frac{y_i}{\sqrt{2^n}} |1, i\rangle \right).$$

This quantum state can be generated by just 1 query. We then apply the transformation  $F$  to basis states  $|0, i\rangle$ , transforming the state to

$$|\psi\rangle = \sum_{i=0}^{2^n-1} \left( \frac{(Fx)_i}{\sqrt{2^n}} |0, i\rangle + \frac{y_i}{\sqrt{2^n}} |1, i\rangle \right).$$

We then use the SWAP test [Buhrman, Cleve, Watrous, and De Wolf \[2002\]](#), a well known test for testing similarity of coefficient vectors of two parts of a quantum state.

The proof of the lower bound,  $R_2(f) = \Omega(\sqrt{N} / \log N)$ , is quite intricate. We define a corresponding problem (which we call REAL FORRELATION) with real valued variables  $x_0, \dots, x_{2^n-1}, y_0, \dots, y_{2^n-1}$  in which we have to distinguish between two cases:

- all  $x_i$  and  $y_i$  are i.i.d. random with Gaussian distribution  $\mathfrak{N}(0, 1)$ ;
- $x_i$ 's are i.i.d random with Gaussian distribution  $\mathfrak{N}(0, 1)$  and  $y_i$  are obtained by applying Fourier transform to a vector consisting of  $x_i$ 's:  $y_i = ((Fx)_i)$ .

In [Aaronson and Ambainis \[2015\]](#), we show that any algorithm for FORRELATION implies an algorithm for REAL FORRELATION with a similar complexity. Thus, it suffices to show a classical lower bound on REAL FORRELATION.

REAL FORRELATION is, in turn, a special case of a more general problem, GAUSSIAN DISTINGUISHING, in which we have to determine whether a set of real-valued variables  $x_1, \dots, x_M$  has a hidden structure. Let  $\vec{v}_1, \dots, \vec{v}_M$  be a set of vectors in  $\mathbb{R}^d$  for some  $d$ . We have to distinguish between two cases:

- (a) all  $x_i$  are i.i.d. random with Gaussian distribution  $\mathfrak{N}(0, 1)$ ;
- (b)  $x_1, \dots, x_M$  are generated by choosing a random  $\vec{u} \in \mathbb{R}^d$  (whose entries are i.i.d.  $\mathfrak{N}(0, 1)$  random variables) and taking  $x_i = (\vec{u}, \vec{v}_i)$ .

The lower bound on REAL FORRELATION is a special case of

**Theorem 2.** *Let  $\vec{v}_i$  be such that  $|(\vec{v}_i, \vec{v}_j)| \leq \epsilon$  for all  $i \neq j$ . Then, GAUSSIAN DISTINGUISHING requires  $\Omega(\frac{1/\epsilon}{\log(M/\epsilon)})$  queries.*

In the case of REAL FORRELATION,  $M = 2^{n+1}$ ,  $d = 2^n$ ,  $\vec{v}_1, \dots, \vec{v}_{2^n}$  are the computational basis states  $|0\rangle, \dots, |2^n - 1\rangle$  and  $\vec{v}_{2^n+1}, \dots, \vec{v}_{2^{n+1}}$  are  $F|0\rangle, \dots, F|2^n - 1\rangle$ . Then,  $\epsilon = \frac{1}{\sqrt{2^n}} = \frac{1}{\sqrt{N/2}}$ , implying a lower bound of  $\Omega(\sqrt{N}/\log N)$  on REAL FORRELATION. This bound is nearly tight, as shown by

**Theorem 3.** *Let  $\mathcal{Q}$  be a 1-query quantum algorithm. There is a probabilistic algorithm  $\mathcal{Q}'$  that makes  $O(\sqrt{N})$  queries and, on every input  $(x_1, \dots, x_N)$ , outputs an estimate  $\tilde{p}$  such that  $|p - \tilde{p}| \leq \epsilon$  (where  $p$  is the accepting probability of  $\mathcal{Q}$  on  $(x_1, \dots, x_N)$ ) with a high probability.*

The simulation makes use of the connection between quantum algorithms and polynomials:

**Lemma 1.** [Beals, Buhrman, Cleve, Mosca, and de Wolf \[2001\]](#) *Let  $\mathcal{Q}$  be a quantum algorithm that makes  $k$  queries to an input  $(x_1, \dots, x_N)$ ,  $x_i \in \{0, 1\}$ . The accepting probability of  $\mathcal{Q}$  can be expressed as a polynomial  $p(x_1, \dots, x_N)$  in variables  $x_1, \dots, x_N$  of degree at most  $2k$ .*

Since the accepting probability of an algorithm must be between 0 and 1, we have  $0 \leq p(x_1, \dots, x_N) \leq 1$  whenever  $x_1, \dots, x_N \in \{0, 1\}$ . [Theorem 3](#) then follows from a more general result about estimating bounded polynomials:

**Lemma 2.** [Aaronson and Ambainis \[2015\]](#) *For every polynomial  $p(x_1, \dots, x_N)$  with  $\deg p \leq 2$  and  $0 \leq p(x_1, \dots, x_N) \leq 1$  for any  $x_1, \dots, x_N \in \{0, 1\}$ , there is a probabilistic algorithm  $\mathcal{Q}'$  that makes  $O(\sqrt{N})$  queries and outputs an estimate  $\tilde{p}$  such that  $|p(x_1, \dots, x_N) - \tilde{p}| \leq \epsilon$  with a high probability, for every input  $(x_1, \dots, x_N) \in \{0, 1\}^N$ .*



More generally, if we have a bounded polynomial  $p(x_1, \dots, x_N)$  with  $\deg p \leq k$ , its value can be estimated with  $O(N^{1-1/k})$  queries. Together with [Lemma 1](#), this implies a probabilistic simulation of  $t$  query quantum algorithms with  $O(N^{1-1/2t})$  queries. Unlike for  $t = 1$ , we do not know whether this is optimal.

**Open Problem 1.** *Let  $t \geq 2$ . Is there a partial function  $g(x_1, \dots, x_N)$  with  $Q_2(g) = t$  and  $R_2(g) = \tilde{\Omega}(N^{1-1/2t})$ ?*

The problem remains open, if instead of  $\tilde{\Omega}(N^{1-1/2t})$  (which matches our upper bound), we ask for a weaker lower bound of  $\Omega(N^c)$ ,  $c > 1/2$  and, even, if instead of a constant  $t$ , we allow  $t = O(\log^c N)$ .

**Open Problem 2.** *Is there a partial function  $g(x_1, \dots, x_N)$  with  $Q_2(g) = O(\log^c N)$  for some  $c$  and  $R_2(g) = \Omega(N^d)$  for  $d > 1/2$ ?*

The well known examples of problems with a large quantum-classical gap (such as Simon's problem [Simon \[1997\]](#) or period-finding) typically have  $R_2(g) = O(\sqrt{N})$ . In [Aaronson and Ambainis \[2015\]](#), we give a candidate problem,  $k$ -FOLD FORRELATION for which we conjecture that bounds of [Open Problem 1](#) hold. This is, however, the only candidate problem that we know.

## 4 Total functions: pointer function method

For total functions  $f$ , the possible gaps between  $Q(f)$ ,  $R(f)$  and  $D(f)$  are much smaller: all of these complexity measures are polynomially related.

It is well known that  $D(f) = O(Q_2^6(f))$  [Beals, Buhrman, Cleve, Mosca, and de Wolf \[2001\]](#) and  $D(f) = O(R_2^3(f))$  [Nisan \[1991\]](#). For exact/zero error algorithms we know that  $D(f) = O(Q_E^3(f))$  [Midrijanis \[2004\]](#) and  $D(f) = O(R_0^2(f))$  [Saks and Wigderson \[1986\]](#). The question is: how tight are these bounds?

For a very long time, the best separations were:

- Quantum vs. probabilistic/deterministic:  $OR(x_1, \dots, x_N)$  has  $Q_2(OR) = O(\sqrt{N})$  due to Grover's quantum search algorithm and  $R_2(f) = \Omega(N)$ .
- Probabilistic vs. deterministic: the binary AND-OR tree function of depth  $d$  has  $D(f) = 2^d$  and  $R_0(f) = O((\frac{1+\sqrt{33}}{4})^d)$ , thus implying that  $R_0(f) = O(D^{0.753\dots}(f))$  [Saks and Wigderson \[ibid.\]](#).

Both of these separations were conjectured to be optimal by a substantial part of the respective research community.

For exact quantum query complexity, the best separation was  $Q_2(XOR) = N/2$  vs.  $D(XOR) = R_2(XOR) = N$  for the  $N$ -bit XOR function [Beals, Buhrman, Cleve, Mosca,](#)

and de Wolf [2001] until 2013 when an example with  $Q_E(f) = O(R_2^{0.86\dots}(f))$  was discovered Ambainis [2016].

In 2015, major improvements to all of these bounds were achieved, via two new methods. The first of them, the *pointer function* method was first invented by Göös, Göös, Pitassi, and Watson [2015] for solving the communication vs. partition number problem in classical communication complexity. It was then quickly adapted to separating query complexity measures by Ambainis, Balodis, Belovs, Lee, Santha, and Smotrovs [2017]:

**Theorem 4.** *Ambainis, Balodis, Belovs, Lee, Santha, and Smotrovs [ibid.]*

1. *There exists a total Boolean function  $f$  with  $Q_2(f) = \tilde{O}(D^{1/4}(f))$ .*
2. *There exists a total Boolean function  $f$  with  $R_0(f) = \tilde{O}(D^{1/2}(f))$ .*
3. *There exists a total Boolean function  $f$  with  $R_2(f) = \tilde{O}(R_0^{1/2}(f))$ .*

The first two results provide major improvements over the previously known results mentioned at the beginning of this section. The third result is the first ever superlinear gap between  $R_0(f)$  and  $R_2(f)$  for a total  $f$ .

We now illustrate the method by describing the simplest function by Göös, Pitassi, and Watson [2015] and sketch a proof that it achieves  $R_2(f) = \tilde{O}(D^{1/2}(f))$ , a slightly weaker result than the second item above. Consider  $f(x_{ij}, y_{ij}, z_{ij})$ , with variables  $x_{ij} \in \{0, 1\}$ ,  $y_{ij} \in [0, N]$ ,  $z_{ij} \in [0, M]$  indexed by  $i \in [N]$ ,  $j \in [M]$ . The variables  $x_{ij}$  are interpreted as elements of an  $N \times M$  table and pairs of variables  $(y_{ij}, z_{ij})$  are interpreted as pointers to entries in this table<sup>2</sup>.

We define that  $f = 1$  if the following conditions are satisfied:

1. the  $N \times M$  table has a unique column  $i$  in which all entries  $x_{ij}$  are 1;
2. in this column, there is exactly one  $j$  for which  $(y_{ij}, z_{ij}) \neq (0, 0)$ ;
3. if we start at this  $(i, j)$  and repeatedly follow the pointers (that is, consider the sequence  $(i_k, j_k)$  defined by  $(i_0, j_0) = (i, j)$  and  $(i_k, j_k) = (y_{i_{k-1}j_{k-1}}, z_{i_{k-1}j_{k-1}})$  for  $k > 0$ ), then:
  - (a) for each  $i' \neq i$ , there is a unique  $k \in [N - 1]$  with  $i_k = i'$ ,
  - (b)  $(i_N, j_N) = (0, 0)$ ,
  - (c)  $x_{i_k j_k} = 0$  for all  $k \in [N - 1]$ .

This function  $f$  has the following properties:

---

<sup>2</sup>As described, this is a function of variables with a larger set of values but it can be converted into a function with  $\{0, 1\}$ -valued variables, with complexities changing by at most a logarithmic factor.

1.  $D(f) = NM$ : for any deterministic algorithm, an adversary may choose the values for variables so that at least one of  $x_{ij}$ ,  $y_{ij}$ ,  $z_{ij}$  needs to be queried for each  $ij$ .
2. If  $f(x_{ij}, y_{ij}, z_{ij}) = 1$ , this can be certified by showing variables  $x_{ij}$ ,  $y_{ij}$ ,  $z_{ij}$  for  $N + M - 1$  different  $(i, j)$ : the all-1 column and the cells  $(i_k, j_k)$  in the sequence of pointers. Moreover, there is one and only one way to certify this.

To show a gap between  $D(f)$  and  $R_2(f)$ , it suffices to show that a randomized algorithm can find this certificate faster than a deterministic algorithm. For that, we set  $N = M$  and consider the following randomized algorithm (due to [Mukhopadhyay and Sanyal \[2015\]](#)):

1.  $\Theta(N \log N)$  times repeat:
  - (a) Choose a random entry  $(i, j)$  of the table in a column that has not been eliminated yet.
  - (b) While  $x_{ij} = 0$ ,  $y_{ij} \neq 0$ ,  $z_{ij} \neq 0$  and  $i$  is not a column that has been already eliminated:
    - eliminate column  $i$ ;
    - set  $i = y_{ij}$  and  $j = z_{ij}$ .
  - (c) If  $x_{ij} = 0$  but  $y_{ij} = 0$  or  $z_{ij} = 0$ , eliminate column  $i$ .
2. If all columns are eliminated or more than 100 columns remain, output 0.
3. Otherwise, test each of remaining columns by checking whether it satisfies the conditions for a certificate.

If  $f = 1$ , each time when we choose a random entry in a column that is not the all-1 column, there is an  $\frac{1}{N}$  probability of choosing the entry that is a part of the pointer chain. This means that, during  $\Theta(N \log N)$  repetitions, this happens  $\Theta(\log N)$  times. Each time, the columns that are after this entry in the pointer chain get eliminated. On average, half of remaining columns are after the entry that gets chosen. This means that, with a high probability, after  $\Theta(\log N)$  times, only  $O(1)$  columns are not eliminated. Then, one can test each of them with  $O(N)$  queries.

This basic construction can be modified in several ways [Ambainis, Balodis, Belovs, Lee, Santha, and Smotrovs \[2017\]](#). To separate two models of computation, we should make the certificate for  $f = 1$  easy to find in one of them but difficult in the other model. (For example, hard to find by zero-error probabilistic algorithms but easy to find by bounded error probabilistic algorithms.) For different separations, the modifications include:

- Arranging the cells with pointers (in columns that are not the all-1 column) into a binary tree instead of a chain.
- Introducing back pointers at the end of the pointer chain or at the leaves of the tree, pointing back to the all-1 column.
- Having more than one all-1 column with pointers among the all-1 columns.

Besides the three major results in [Theorem 4](#), this approach gives better-than-before separations between exact quantum query complexity and all classical complexity measures ( $Q_E(f) = \tilde{O}(\sqrt{D(f)})$ ,  $Q_E(f) = \tilde{O}(\sqrt{R_0(f)})$ , and  $Q_E(f) = \tilde{O}(R_2^{2/3}(f))$ ), between bounded-error quantum and zero-error probabilistic complexity ( $Q_2(f) = \tilde{O}(\sqrt[3]{R_0(f)})$ ), and between polynomial degree and randomized query complexity ( $\widetilde{\deg}(f) = \tilde{O}(\sqrt[4]{R_2(f)})$ ) [Ambainis, Balodis, Belovs, Lee, Santha, and Smotrovs \[ibid.\]](#).

## 5 Total functions: cheat sheet method

**5.1 Query complexity.** After the developments described in the previous section, the biggest separation between quantum and randomized complexities still remained  $Q(f) = O(\sqrt{R_2(f)})$ . This was improved to  $Q(f) = \tilde{O}(R_2(f)^{2/5})$  in a breakthrough paper by [Aaronson, Ben-David, and Kothari \[2016\]](#), using another new method, *cheat sheets*.

The key feature of *cheat sheet* method is that it takes separations for partial functions and transforms them into separations for total functions, by adding extra variables that allow to check that the input satisfies the promise for one of two cases when the partial function  $f$  is defined. The main result is

**Theorem 5.** [Aaronson, Ben-David, and Kothari \[ibid.\]](#) *Let  $f(x_1, \dots, x_N)$  be a partial function with  $Q_2(f) = Q$ ,  $R_2(f) = R$  and  $C(f) = C$ . Then, there exists a total function  $f_{CS}$  with  $Q_2(f_{CS}) = \tilde{O}(Q + \sqrt{C})$  and  $R_2(f_{CS}) = \Omega(R)$ .*

Let  $f(x_1, \dots, x_{N^3})$  be the partial function  $f = \text{AND} \circ \text{OR} \circ \text{FORRELATION}$  obtained by composing *AND*, *OR* and *FORRELATION* on  $N$  variables each. From the complexities of *AND*, *OR* and *FORRELATION* and composition properties of the complexity measures it follows that  $Q_2(f) = O(N)$ ,  $C(f) = O(N^2)$  and  $R_2(f) = \tilde{\Omega}(N^{2.5})$ , implying

**Theorem 6.** [Aaronson, Ben-David, and Kothari \[ibid.\]](#) *There exists a total Boolean function  $f_{CS}$  with  $Q_2(f_{CS}) = \tilde{O}(R_2^{2/5}(f_{CS}))$ .*

Moreover, if [Open Problem 1](#) was resolved in affirmative, we could substitute the corresponding  $g$  instead of *FORRELATION* and [Theorem 5](#) would imply  $Q_2(g_{CS}) = \tilde{O}(R_2^{1/3+o(1)}(g_{CS}))$ .

The cheat sheet method also gives new separations between  $Q_2(f)$  and many of combinatorial complexity measures:  $Q_2(f) = \tilde{\Omega}(C^2(f))$ ,  $Q_2(f) = \tilde{\Omega}(\deg^2(f))$ , and  $Q_2(f) = \Omega(\widetilde{\deg}^{4-o(1)}(f))$ . Moreover, several of results proven via the pointer function method (for example, the  $Q_E(f) = \tilde{O}(R_2^{2/3}(f))$  separation) can be reproven via cheat sheets [Aaronson, Ben-David, and Kothari \[2016\]](#).

To show [Theorem 5, Aaronson, Ben-David, and Kothari \[ibid.\]](#) define the *cheat-sheet* function  $f_{CS}$  in a following way.  $f_{CS}$  has  $tN + 2^t tC \lceil \log N + 1 \rceil$  variables (for an appropriately chosen  $t = \Theta(\log N)$ ) which we denote  $x_{11}, \dots, x_{tN}, y_{11}, \dots, y_{2^t M}$  (where  $M = Ct \lceil \log N + 1 \rceil$ ). We interpret the blocks of variables  $x^{(i)} = (x_{i1}, \dots, x_{iN})$  as inputs to the function  $f$  and the blocks  $y^{(i)} = (y_{i1}, \dots, y_{iM})$  as descriptions for  $t$  certificates of function  $f$ , with the description containing both the set of variables  $S \subseteq [N]$  and the values that  $x_i, i \in S$  must take. (We refer to those blocks as *cheat-sheets*, as they allow to verify the values of  $f(x^{(1)}), \dots, f(x^{(t)})$  with less queries than it takes to compute them.)

We interpret the  $t$ -bit string  $s = s_1 \dots s_t, s_i = f(x^{(i)})$  as an index for the block  $y^{(s)}$ . We define that  $f_{CS} = 1$  if the block  $y^{(s)}$  contains certificates for  $f(x) = s_1, \dots, f(x) = s_t$  and the values of corresponding input variables in inputs  $x^{(1)}, \dots, x^{(t)}$  match the ones specified by the corresponding certificate. Otherwise,  $f_{CS} = 0$ .

To compute  $f_{CS}$  by a quantum algorithm, we proceed as follows:

1. compute  $f(x^{(1)}), \dots, f(x^{(t)})$ , repeating each computation  $O(\log t)$  times, to make the error probability at most  $1/(10t)$  for each  $f(x^{(i)})$  (then, the probability that all  $f(x^{(i)})$  are all simultaneously correct is at least  $9/10$ );
2. check whether the certificates in the block  $y^{(s)}$  are satisfied by inputs  $x^{(1)}, \dots, x^{(t)}$ , by using Grover's quantum search to search for a variable in one of  $x^{(i)}$  which does not match the corresponding certificate.

The complexity of the 1<sup>st</sup> stage is  $O(Qt \log t)$ . The complexity of the 2<sup>nd</sup> stage is  $O(\sqrt{Ct} \log N)$ , since we have to search among  $tC$  variables  $x_j^{(i)}$  ( $t$  certificates, each of which contains  $C$  variables), Grover's quantum search [Grover \[1996\]](#) allows to search among them by testing  $O(\sqrt{tC})$  possibilities, and testing each possibility requires reading  $O(\log N)$  variables in the block  $y^{(s)}$ . Thus, the overall complexity is  $\tilde{O}(Q + \sqrt{C})$  quantum queries.

Classically,  $Rt$  queries are required to solve  $t$  instances of  $f(x^{(i)})$ . Moreover, if the number of queries is substantially smaller (of an order  $o(Rt)$ ), then, with a high probability, most of  $f(x^{(i)})$  are not solved yet and, at that point, a classical algorithm cannot make use of certificate descriptions in  $y^{(j)}$  because there are too many possible  $y^{(j)}$ . This suggests that  $R_2(f_{CS}) = \Omega(Rt)$  and [Aaronson, Ben-David, and Kothari \[2016\]](#) show that this is indeed the case.

**5.2 Communication complexity.** The cheat sheet method has also found applications in a different domain, *communication complexity* [Kushilevitz and Nisan \[1997\]](#) and [Lee and Shraibman \[2007\]](#). In the standard model of communication complexity, we have two parties, Alice and Bob, who want to compute a function  $f(x, y)$ , with Alice holding the input  $x$  and Bob holding the input  $y$ . The task is to compute  $f(x, y)$  with the minimum amount of communication between Alice and Bob. Communication complexity has a number of applications, from designing efficient communication protocols for various tasks to proving lower bounds on other models of computation (for example, streaming algorithms).

If quantum communication is allowed, the communication complexity may decrease exponentially. Similarly to query complexity, let  $Q_2(f)$ ,  $Q_E(f)$  and  $R_2(f)$  denote the bounded-error quantum, exact quantum and bounded error randomized communication complexity of  $f$ . A partial function with an exponential gap between  $R_2(f)$  and  $Q_2(f)$  was first constructed by [Raz \[1999\]](#) in 1999. In a later work, it was shown that quantum protocols can be exponentially more efficient even if the quantum protocol is restricted to one message from [Klartag and Regev \[2011\]](#) but it is compared against randomized protocols that can send an arbitrary number of messages back and forth.

However, similarly to query complexity, quantum advantages for total functions have been much more limited, with the best known separation of  $Q(f) = O(\sqrt{R_2(f)})$  [Buhrman, Cleve, and Wigderson \[1999\]](#) and [Aaronson and Ambainis \[2005\]](#) for the set disjointness problem which is the natural communication counterpart of Grover's search. [Anshu, Belovs, Ben-David, Göös, Jain, Kothari, Lee, and Santha \[2016\]](#) have adapted the cheat sheet method to communication complexity, proving

**Theorem 7.** [Anshu, Belovs, Ben-David, Göös, Jain, Kothari, Lee, and Santha \[ibid.\]](#)

1. There is a total function  $f(x, y)$  with  $Q_2(f) = \tilde{O}(R_2^{2/5}(f))$ ;
2. There is a total function  $f(x, y)$  with  $Q_E(f) = \tilde{O}(R_2^{2/3}(f))$ ;

## 6 Quantum-classical separations on almost all inputs?

All known partial functions  $f(x_1, \dots, x_N)$  with a superpolynomial quantum advantage have the property that  $f$  takes one of values  $f = 0$  and  $f = 1$  on a very small subset of inputs. For example, for FORRELATION, the fraction of inputs with  $f = 1$  is exponentially small in the number of variables  $N$ . This had led to a following conjecture (known as a folklore since about 1999):

**Conjecture 1.** [Aaronson and Ambainis \[2014\]](#) Let  $\mathcal{Q}$  be a quantum algorithm that makes  $T$  queries and let  $\epsilon, \delta > 0$ . There is a deterministic algorithm with a number of queries

that is polynomial in  $T$ ,  $\frac{1}{\epsilon}$  and  $\frac{1}{\delta}$  and approximates the probability of  $\mathbb{Q}$  outputting 1 to an additive error  $\epsilon$  on at least  $1 - \delta$  fraction of all inputs.

For total function, this conjecture implies that quantum and deterministic complexity are polynomially equivalent in the setting of approximately computing  $f$ . That is, for a total function  $f$ , let  $D_\epsilon(f)$  and  $Q_\epsilon(f)$  be the smallest number of queries for a (deterministic or quantum) algorithm that outputs the correct answer on at least  $1 - \epsilon$  fraction of inputs  $(x_1, \dots, x_N)$ . Then, [Conjecture 1](#) implies that  $D_\epsilon(f)$  and  $Q_{\epsilon'}(f)$  are polynomially related, for all constant  $\epsilon, \epsilon'$  with  $\epsilon > \epsilon'$ .

There is a natural path towards proving [Conjecture 1](#). Due to [Lemma 1](#), [Conjecture 1](#) is implied by

**Conjecture 2.** [Aaronson and Ambainis \[2014\]](#) Let  $p(x_1, \dots, x_N)$  be a polynomial of degree  $2T$  which satisfies  $|p(x_1, \dots, x_N)| \leq 1$  for all  $x_1, \dots, x_N \in \{0, 1\}$  and let  $\epsilon, \delta > 0$ . There is a deterministic algorithm with a number of queries that is polynomial in  $T$ ,  $\frac{1}{\epsilon}$  and  $\frac{1}{\delta}$  and approximates  $p(x_1, \dots, x_N)$  to an additive error  $\epsilon$  on at least  $1 - \delta$  fraction of all inputs.

The natural way to design such a deterministic algorithm is by repeatedly choosing the variable  $x_i$  that has the biggest influence on the value of  $p$  (with the influence defined as  $\text{Inf}_i(p) = E_x[|p(x) - p(x^{\{i\}})|^2]$  with the expectation over a random choice of  $x \in \{0, 1\}^n$ ). To prove [Conjecture 2](#), it suffices to show

**Conjecture 3.** [Aaronson and Ambainis \[ibid.\]](#) Let  $p(x_1, \dots, x_N)$  be a polynomial of degree  $2T$  which satisfies  $|p(x_1, \dots, x_N)| \leq 1$  for all  $x_1, \dots, x_N \in \{0, 1\}$ . Assume that

$$E_{x \in \{0, 1\}^n} \left[ (p(x) - E[p(x)])^2 \right] \geq \epsilon.$$

Then, there is a variable  $i$  with  $\text{Inf}_i[p] \geq \left(\frac{\epsilon}{T}\right)^c$  for some constant  $c$ .

[Conjecture 3](#) connects with research in the analysis of Boolean functions. In particular, work of [Dinur, Friedgut, Kindler, and O'Donnell \[2006\]](#) implies a weaker form of the conjecture, with  $\text{Inf}_i[p] \geq \frac{\epsilon^3}{2^{\frac{3}{2}T}}$ . Improving it to  $\text{Inf}_i[p] \geq \left(\frac{\epsilon}{T}\right)^c$  is a challenging open problem which is interesting for both analysis of Boolean functions and quantum query complexity.

## 7 Structure of quantum speedups?

Another related question is: when can we achieve large quantum speedups? From the known examples of exponential and superexponential speedups for partial functions, we

can observe that they are typically achieved for problems with an algebraic structure. For example, [Simon \[1997\]](#) showed an exponential speedup for the following problem:

**Simon’s problem.** Let  $N = 2^n$ . We are promised that the input  $(x_0, \dots, x_{N-1})$  (where  $x_i \in [M]$ ) satisfies one of two promises:

- (a) the mapping  $i \rightarrow x_i$  is 2-to-1 with some  $z \in [N], z \neq 0$  such that  $x_y = x_{y \oplus z}$  for all  $y \in [N]$ , with  $\oplus$  denoting bitwise addition modulo 2;
- (b) the mapping  $i \rightarrow x_i$  is 1-1.

As shown by Simon,  $Q_2(f) = O(n)$  but  $R_2(f) = \Omega(2^{n/2})$ . However, randomly permuting inputs turns Simon’s problem into the problem of distinguishing whether  $i \rightarrow x_i$  is 2-1 or 1-1 for which it is known that  $Q_2(f) = \Theta(2^{n/3})$  but  $R_2(f) = \Theta(2^{n/2})$  [Brassard, Høyer, and Tapp \[1997\]](#) and [Aaronson and Ambainis \[2015\]](#), with the exponential quantum speedup disappearing. Similarly, permuting the input variables destroys the superexponential quantum speedup for the FORRELATION problem.

This leads to a question: can we show that quantum speedup is at most polynomial for any partial function that is symmetric with respect to permuting the input variables  $x_i$ ? A positive answer would imply that large quantum speedups require problems with a structure (typically, of algebraic nature) that disappears if inputs are permuted.

For the case when  $x_i$ ’s are binary, evaluating a partial symmetric function essentially requires counting the number of  $i : x_i = 1$  up to a certain precision (which is sufficient for distinguishing whether the input  $x = (x_1, \dots, x_N)$  satisfies  $f(x) = 0$  or  $f(x) = 1$ ). Quantum algorithms can count  $i : x_i = 1$  quadratically faster than classical algorithms [Brassard, Høyer, and Tapp \[1998\]](#) and it is easy to show that larger speedups cannot be obtained.

For non-binary inputs there are two possible ways of defining a “symmetric function”:

- (a)  $f : [M]^N \rightarrow \{0, 1\}$  is symmetric, if  $f(x_1, \dots, x_N) = f(x_{\pi(1)}, \dots, x_{\pi(N)})$  for any permutation  $\pi$  on  $\{1, 2, \dots, N\}$ ;
- (b)  $f : [M]^N \rightarrow \{0, 1\}$  is symmetric, if  $f(x_1, \dots, x_N) = f(\tau(x_{\pi(1)}), \dots, \tau(x_{\pi(N)}))$  for any permutations  $\pi$  on  $\{1, 2, \dots, N\}$  and  $\tau$  on  $\{1, 2, \dots, M\}$ .

For example, the property of being 1-1 or 2-1 is preserved both if  $x_1, \dots, x_N$  are permuted and if the values for  $x_1, \dots, x_N$  are permuted. Thus, it is symmetric in the second, stronger sense. Similarly, element distinctness (determining whether  $x_1, \dots, x_N$  are all distinct) and other natural properties are symmetric in the second sense. For such properties, we have

**Theorem 8.** *Assume that a partial function  $f : [M]^N \rightarrow \{0, 1\}$  is symmetric in the second sense. Then,  $R_2(f) = O(Q_2^7(f) \log^c Q_2(f))$ .*



It has been conjectured since about 2000 that a similar result also holds for  $f$  with a symmetry of the first type.

A related question has been studied by [Aaronson and Ben-David \[2016\]](#): given a total function  $f : \{0, 1\}^N \rightarrow \{0, 1\}$ , can we define a subproblem  $f_P$  ( $f$  restricted to some subset  $P \subseteq \{0, 1\}^N$ ) for which  $Q_2(f_P) = O(\log^c R_2(f_P))$ ?

For example, if  $f(x_1, \dots, x_N) = x_1 OR \dots OR x_N$ , then, for any restriction, the quantum advantage is at most quadratic. (An intuitive explanation is that computing OR is essentially equivalent to finding  $i : x_i = 1$  and, for search, the quantum advantage is quadratic whatever the number of  $i : x_i = 1$  is.) In contrast, both MAJORITY and PARITY can be restricted so that quantum advantage becomes exponential.

The next theorem gives a full characterization when superpolynomial speedups can be achieved:

**Theorem 9.** [Aaronson and Ambainis \[2014\]](#) *A promise  $P \subseteq \{0, 1\}^N$  with  $Q_2(f_P) = O(N^{o(1)})$  and  $R_2(f_P) = \Omega(N^{\Omega(1)})$  exists if and only if, for some  $c > 0$ , there are  $2^{N^c}$  inputs  $x \in \{0, 1\}^N$  with  $C_x(f) \geq N^c$ .*

## 8 From polynomials to quantum algorithms

As shown by [Lemma 1](#), a quantum algorithm that makes  $k$  queries can be converted into a polynomial of degree at most  $2k$ . In the opposite direction, the existence of a polynomial of degree  $2k$  does not imply the existence of a quantum algorithm that makes  $k$  queries. As mentioned in [Section 5.1](#), there is a total  $f$  with  $Q_2(f) = \Omega(\widetilde{deg}^{4-o(1)}(f))$  [Aaronson, Ben-David, and Kothari \[2016\]](#).

However, there is an interesting particular case in which polynomials and quantum algorithms are equivalent.

**Theorem 10** ([Aaronson, Ambainis, Iraids, Kokainis, and Smotrovs \[2016\]](#)). *Let  $f(x_1, \dots, x_N)$  be a partial Boolean function. Assume that there is a polynomial  $p(x_1, \dots, x_N)$  of degree 2 with the following properties:*

- *for any  $x_1, \dots, x_N \in \{0, 1\}$ ,  $0 \leq p(x_1, \dots, x_N) \leq 1$ ;*
- *if  $f(x_1, \dots, x_N) = 1$ ,  $p(x_1, \dots, x_N) \geq \frac{1}{2} + \delta$ ;*
- *if  $f(x_1, \dots, x_N) = 0$ ,  $p(x_1, \dots, x_N) \leq \frac{1}{2} - \delta$ .*

*Then,  $f(x_1, \dots, x_N)$  can be computed by a 1-query quantum algorithm with the probability of correct answer at least  $\frac{1}{2} + \frac{\delta}{3(2K+1)}$  where  $K$  is the Grothendieck's constant [Pisier \[2012\]](#) (for which it is known that  $1.5707\dots \leq K \leq 1.7822\dots$  [Braverman, K. Makarychev, Y. Makarychev, and Naor \[2011\]](#)).*

The main ideas for the transformation from a polynomial to a quantum algorithm are as follows:

1. For technical convenience, we assume that  $x_i$  are  $\{-1, 1\}$ -valued (instead of  $\{0, 1\}$ -valued). We start by transforming a polynomial  $p(x_1, \dots, x_N)$  into another polynomial

$$q(x_1, \dots, x_N, y_1, \dots, y_N) = \sum_{i,j} a_{i,j} x_i y_j$$

which satisfies  $q(x_1, \dots, x_N, x_1, \dots, x_N) = p(x_1, \dots, x_N)$  for all  $x_1, \dots, x_N \in \{-1, 1\}$  and  $|q(x_1, \dots, x_N, y_1, \dots, y_N)| \leq 1$  for all  $x_1, \dots, x_N, y_1, \dots, y_N \in \{-1, 1\}$ .

2. If the spectral norm  $\|A\|$  of the matrix  $A$  is small, then the polynomial  $q$  can be transformed into a quantum algorithm:

**Lemma 3.** *Let  $A = (a_{ij})_{i \in [N], j \in [M]}$  with  $\sqrt{NM} \|A\| \leq C$  and let*

$$q(x_1, \dots, x_N, y_1, \dots, y_M) = \sum_{i=1}^N \sum_{j=1}^M a_{ij} x_i y_j.$$

*Then, there is a quantum algorithm that makes 1 query to  $x_1, \dots, x_N, y_1, \dots, y_M$  and outputs 1 with probability*

$$r = \frac{1}{2} \left( 1 + \frac{q(x_1, \dots, x_N, y_1, \dots, y_M)}{C} \right).$$

The quantum algorithm consists of creating a combination of quantum states  $|\psi\rangle = \sum_{i=1}^N \frac{x_i}{\sqrt{N}} |i\rangle$  and  $|\phi\rangle = \sum_{j=1}^M \frac{y_j}{\sqrt{M}} |j\rangle$ , applying  $U = \sqrt{NM} \cdot A$  to  $|\phi\rangle$  and then using the SWAP test [Buhrman, Cleve, Watrous, and De Wolf \[2002\]](#) to estimate the inner product of  $|\psi\rangle$  and  $U|\phi\rangle$  which happens to be equal to the desired quantity  $\sum_{i=1}^N \sum_{j=1}^M a_{ij} x_i y_j$ .

If  $U$  is unitary, we can apply this procedure as described. If  $\|U\| = C > 1$ ,  $U$  is not unitary and cannot be applied directly. Instead, we design and apply a unitary transformation that is equal to  $\frac{1}{C}U$  on a certain subspace.

3. For the general case, a corollary of Grothendieck's inequality [Pisier \[2012\]](#), [Aaronson, Ambainis, Iraids, Kokainis, and Smotrovs \[2016\]](#), and [Arunachalam, Briët, and Palazuelos \[2017\]](#) implies that, if  $a_{ij}$  are such that  $|\sum_{i=1}^N \sum_{j=1}^M a_{ij} x_i y_j| \leq 1$  for all choices of  $x_i \in \{-1, 1\}$  and  $y_j \in \{-1, 1\}$ , there

exist  $\vec{u} = (u_i)_{i \in N}$  and  $\vec{v} = (v_j)_{j \in M}$  such that  $\|\vec{u}\| = 1$ ,  $\|\vec{v}\| = 1$ ,  $a_{ij} = b_{ij}u_iv_j$  for all  $i \in [N]$ ,  $j \in [M]$  and  $B = (b_{ij})_{i,j}$  satisfies  $\|B\| \leq K$ .

Then, we can perform a similar algorithm with quantum states  $|\psi\rangle = \sum_{i=1}^N u_i x_i |i\rangle$  and  $|\phi\rangle = \sum_{j=1}^M v_j y_j |j\rangle$ .

Following this work, it was shown [Arunachalam, Briët, and Palazuelos \[2017\]](#) that quantum algorithms are equivalent to polynomial representations by polynomials of a particular type. Namely, the accepting probability of a  $t$  query quantum algorithm is equal to a completely bounded form of degree  $2t$ . For  $t = 1$ , representations of  $f$  by a completely bounded forms are equivalent to representations by general polynomials (implying [Theorem 10](#)) but this does not hold for  $t \geq 2$ .

## References

- Scott Aaronson (2008). “Quantum certificate complexity”. *J. Comput. System Sci.* 74.3, pp. 313–322. MR: [2384077](#) (cit. on p. [3289](#)).
- Scott Aaronson and Andris Ambainis (2005). “Quantum search of spatial regions”. *Theory Comput.* 1, pp. 47–79. MR: [2322514](#) (cit. on p. [3297](#)).
- (2014). “The need for structure in quantum speedups”. *Theory Comput.* 10, pp. 133–166. MR: [3249097](#) (cit. on pp. [3297](#), [3298](#), [3300](#)).
  - (2015). “Forrelation: a problem that optimally separates quantum from classical computing”. In: *STOC’15—Proceedings of the 2015 ACM Symposium on Theory of Computing*. ACM, New York, pp. 307–316. MR: [3388209](#) (cit. on pp. [3285](#), [3289](#), [3291](#), [3292](#), [3299](#)).
- Scott Aaronson, Andris Ambainis, Jānis Iraids, Martins Kokainis, and Juris Smotrovs (2016). “Polynomials, quantum query complexity, and Grothendieck’s inequality”. In: *31st Conference on Computational Complexity*. Vol. 50. LIPIcs. Leibniz Int. Proc. Inform. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, Art. No. 25, 19. MR: [3540826](#) (cit. on pp. [3300](#), [3301](#)).
- Scott Aaronson and Shalev Ben-David (2016). “Sculpting quantum speedups”. In: *31st Conference on Computational Complexity*. Vol. 50. LIPIcs. Leibniz Int. Proc. Inform. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, Art. No. 26, 28. MR: [3540827](#) (cit. on p. [3300](#)).
- Scott Aaronson, Shalev Ben-David, and Robin Kothari (2016). “Separations in query complexity using cheat sheets”. In: *STOC’16—Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, pp. 863–876. MR: [3536620](#) (cit. on pp. [3285](#), [3295](#), [3296](#), [3300](#)).

- Scott Aaronson and Yaoyun Shi (2004). “Quantum lower bounds for the collision and the element distinctness problems”. *J. ACM* 51.4, pp. 595–605. MR: [2147849](#) (cit. on p. [3285](#)).
- Andris Ambainis (2004). “Quantum search algorithms”. *ACM SIGACT News* 35.2, pp. 22–35 (cit. on p. [3284](#)).
- (2007). “Quantum walk algorithm for element distinctness”. *SIAM J. Comput.* 37.1, pp. 210–239. MR: [2306290](#) (cit. on p. [3285](#)).
  - (2016). “Superlinear advantage for exact quantum algorithms”. *SIAM J. Comput.* 45.2, pp. 617–631. MR: [3490898](#) (cit. on p. [3293](#)).
- Andris Ambainis, Kaspars Balodis, Aleksandrs Belovs, Troy Lee, Miklos Santha, and Juris Smotrovs (2017). “Separations in query complexity based on pointer functions”. *J. ACM* 64.5, Art. 32, 24. MR: [3716887](#) (cit. on pp. [3293–3295](#)).
- Anurag Anshu, Aleksandrs Belovs, Shalev Ben-David, Mika Göös, Rahul Jain, Robin Kothari, Troy Lee, and Miklos Santha (2016). “Separations in communication complexity using cheat sheets and information complexity [extended abstract]”. In: *57th Annual IEEE Symposium on Foundations of Computer Science—FOCS 2016*. IEEE Computer Soc., Los Alamitos, CA, pp. 555–564. MR: [3631018](#) (cit. on p. [3297](#)).
- Srinivasan Arunachalam, Jop Briët, and Carlos Palazuelos (2017). “Quantum Query Algorithms are Completely Bounded Forms”. arXiv: [1711.07285](#) (cit. on pp. [3285](#), [3301](#), [3302](#)).
- Robert Beals, Harry Buhrman, Richard Cleve, Michele Mosca, and Ronald de Wolf (2001). “Quantum lower bounds by polynomials”. *J. ACM* 48.4, pp. 778–797. MR: [2144930](#) (cit. on pp. [3288](#), [3289](#), [3291](#), [3292](#)).
- Daniel J Bernstein, Stacey Jeffery, Tanja Lange, and Alexander Meurer (2013). “Quantum algorithms for the subset-sum problem”. In: *International Workshop on Post-Quantum Cryptography*. Springer, pp. 16–33 (cit. on p. [3285](#)).
- Gilles Brassard, Peter Høyer, and Alain Tapp (1997). “Quantum cryptanalysis of hash and claw-free functions”. *ACM Sigact News* 28.2, pp. 14–19 (cit. on p. [3299](#)).
- Gilles Brassard, Peter Høyer, and Alain Tapp (1998). “Quantum counting”. In: *Automata, languages and programming (Aalborg, 1998)*. Vol. 1443. Lecture Notes in Comput. Sci. Springer, Berlin, pp. 820–831. MR: [1683527](#) (cit. on p. [3299](#)).
- Mark Braverman, Konstantin Makarychev, Yuri Makarychev, and Assaf Naor (2011). “The Grothendieck constant is strictly smaller than Krivine’s bound”. In: *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science—FOCS 2011*. IEEE Computer Soc., Los Alamitos, CA, pp. 453–462. MR: [2932721](#) (cit. on p. [3300](#)).
- Harry Buhrman, Richard Cleve, John Watrous, and Ronald De Wolf (2002). “Quantum fingerprinting”. *Physical Review Letters* 87.16, p. 167902 (cit. on pp. [3290](#), [3301](#)).

- Harry Buhrman, Richard Cleve, and Avi Wigderson (1999). “Quantum vs. classical communication and computation”. In: *STOC '98 (Dallas, TX)*. ACM, New York, pp. 63–68. MR: [1731563](#) (cit. on p. [3297](#)).
- Harry Buhrman and Robert Špalek (2006). “[Quantum verification of matrix products](#)”. In: *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms*. ACM, New York, pp. 880–889. MR: [2373814](#) (cit. on p. [3285](#)).
- Harry Buhrman and Ronald de Wolf (2002). “[Complexity measures and decision tree complexity: a survey](#)”. *Theoret. Comput. Sci.* 288.1. Complexity and logic (Vienna, 1998), pp. 21–43. MR: [1934888](#) (cit. on p. [3284](#)).
- Sourav Chakraborty, Eldar Fischer, Arie Matsliah, and Ronald de Wolf (2010). “New results on quantum property testing”. In: *30th International Conference on Foundations of Software Technology and Theoretical Computer Science*. Vol. 8. LIPIcs. Leibniz Int. Proc. Inform. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, pp. 145–156. MR: [2853831](#) (cit. on p. [3284](#)).
- J Ignacio Cirac and Peter Zoller (2012). “Goals and opportunities in quantum simulation”. *Nature Physics* 8.4, pp. 264–266 (cit. on p. [3284](#)).
- Irit Dinur, Ehud Friedgut, Guy Kindler, and Ryan O’Donnell (2006). “[On the Fourier tails of bounded functions over the discrete cube \(extended abstract\)](#)”. In: *STOC’06: Proceedings of the 38th Annual ACM Symposium on Theory of Computing*. ACM, New York, pp. 437–446. MR: [2277169](#) (cit. on p. [3298](#)).
- IM Georgescu, Sahel Ashhab, and Franco Nori (2014). “Quantum simulation”. *Reviews of Modern Physics* 86.1, p. 153 (cit. on p. [3284](#)).
- Mika Göös, Toniann Pitassi, and Thomas Watson (2015). “Deterministic communication vs. partition number”. In: *2015 IEEE 56th Annual Symposium on Foundations of Computer Science—FOCS 2015*. IEEE Computer Soc., Los Alamitos, CA, pp. 1077–1088. MR: [3473358](#) (cit. on p. [3293](#)).
- Lov K. Grover (1996). “[A fast quantum mechanical algorithm for database search](#)”. In: *Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996)*. ACM, New York, pp. 212–219. MR: [1427516](#) (cit. on pp. [3284](#), [3296](#)).
- Stephen Jordan (n.d.). “[Quantum algorithm zoo](#)” (cit. on p. [3285](#)).
- Bo’az Klartag and Oded Regev (2011). “[Quantum one-way communication can be exponentially stronger than classical communication](#)”. In: *STOC’11—Proceedings of the 43rd ACM Symposium on Theory of Computing*. ACM, New York, pp. 31–40. MR: [2931952](#) (cit. on p. [3297](#)).
- Eyal Kushilevitz and Noam Nisan (1997). *Communication complexity*. Cambridge University Press, Cambridge, pp. xiv+189. MR: [1426129](#) (cit. on p. [3297](#)).

- Troy Lee and Adi Shraibman (2007). “Lower bounds in communication complexity”. *Found. Trends Theor. Comput. Sci.* 3.4, front matter, 263–399 (2009). MR: [2558900](#) (cit. on p. [3297](#)).
- Gatis Midrijanis (2004). “Exact quantum query complexity for total Boolean functions”. arXiv: [quant-ph/0403168](#) (cit. on p. [3292](#)).
- Sagnik Mukhopadhyay and Swagato Sanyal (2015). “Towards better separation between deterministic and randomized query complexity”. In: *35th IARCS Annual Conference on Foundations of Software Technology and Theoretical Computer Science*. Vol. 45. LIPIcs. Leibniz Int. Proc. Inform. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, pp. 206–220. MR: [3464466](#) (cit. on p. [3294](#)).
- Michael A. Nielsen and Isaac L. Chuang (2000). *Quantum computation and quantum information*. Cambridge University Press, Cambridge, pp. xxvi+676. MR: [1796805](#) (cit. on p. [3287](#)).
- Noam Nisan (1991). “CREW PRAMs and decision trees”. *SIAM J. Comput.* 20.6, pp. 999–1007. MR: [1135744](#) (cit. on pp. [3288](#), [3289](#), [3292](#)).
- Noam Nisan and Márió Szegedy (1994). “On the degree of Boolean functions as real polynomials”. *Comput. Complexity* 4.4. Special issue on circuit complexity (Barbados, 1992), pp. 301–313. MR: [1313531](#).
- Gilles Pisier (2012). “Grothendieck’s theorem, past and present”. *Bull. Amer. Math. Soc. (N.S.)* 49.2, pp. 237–323. MR: [2888168](#) (cit. on pp. [3300](#), [3301](#)).
- Ran Raz (1999). “Exponential separation of quantum and classical communication complexity”. In: *Annual ACM Symposium on Theory of Computing (Atlanta, GA, 1999)*. ACM, New York, pp. 358–367. MR: [1798056](#) (cit. on p. [3297](#)).
- Michael Saks and Avi Wigderson (1986). “Probabilistic Boolean decision trees and the complexity of evaluating game trees”. In: *Foundations of Computer Science, 1986., 27th Annual Symposium on*. IEEE, pp. 29–38 (cit. on pp. [3285](#), [3292](#)).
- Peter W. Shor (1997). “Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer”. *SIAM J. Comput.* 26.5, pp. 1484–1509. MR: [1471990](#) (cit. on p. [3284](#)).
- Daniel R. Simon (1997). “On the power of quantum computation”. *SIAM J. Comput.* 26.5, pp. 1474–1483. MR: [1471989](#) (cit. on pp. [3292](#), [3299](#)).

Received 2017-12-09.

ANDRIS AMBAINIS  
FACULTY OF COMPUTING  
UNIVERSITY OF LATVIA  
RAINA BULVĀRIS 19  
RĪGA, LV-1586  
LATVIA  
[andris.ambainis@lu.lv](mailto:andris.ambainis@lu.lv)



# APPROXIMATE NEAREST NEIGHBOR SEARCH IN HIGH DIMENSIONS

ALEXANDR ANDONI, PIOTR INDYK AND ILYA RAZENSHTYEN

## Abstract

The nearest neighbor problem is defined as follows: Given a set  $P$  of  $n$  points in some metric space  $(X, D)$ , build a data structure that, given any point  $q$ , returns a point in  $P$  that is closest to  $q$  (its “nearest neighbor” in  $P$ ). The data structure stores additional information about the set  $P$ , which is then used to find the nearest neighbor without computing all distances between  $q$  and  $P$ . The problem has a wide range of applications in machine learning, computer vision, databases and other fields.

To reduce the time needed to find nearest neighbors and the amount of memory used by the data structure, one can formulate the *approximate* nearest neighbor problem, where the goal is to return any point  $p' \in P$  such that the distance from  $q$  to  $p'$  is at most  $c \cdot \min_{p \in P} D(q, p)$ , for some  $c \geq 1$ . Over the last two decades many efficient solutions to this problem were developed. In this article we survey these developments, as well as their connections to questions in geometric functional analysis and combinatorial geometry.

## 1 Introduction

The *nearest neighbor* problem is defined as follows: Given a set  $P$  of  $n$  points in a metric space defined over a set  $X$  with distance function  $D$ , build a data structure<sup>1</sup> that, given any “query” point  $q \in X$ , returns its “nearest neighbor”  $\arg \min_{p \in P} D(q, p)$ . A particularly interesting and well-studied case is that of nearest neighbor in geometric spaces, where  $X = \mathbb{R}^d$  and the metric  $D$  is induced by some norm. The problem has a wide range of applications in machine learning, computer vision, databases and other fields, see [Shakhnarovich, Darrell, and Indyk \[2006\]](#) and [Andoni and Indyk \[2008\]](#) for an overview.

---

This research was supported by NSF and Simons Foundation.

MSC2010: primary 68W20; secondary 52A21, 46B09, 46B85, 68P05.

<sup>1</sup>See [Section 1.1](#) for a discussion about the computational model.



A simple solution to this problem would store the set  $P$  in memory, and then, given  $q$ , compute all distances  $\mathsf{D}(q, p)$  for  $p \in P$  and select the point  $p$  with the minimum distance. Its disadvantage is the computational cost: computing all  $n$  distances requires at least  $n$  operations. Since in many applications  $n$  can be as large as  $10^9$  (see e.g., [Sundaram, Turmukhametova, Satish, Mostak, Indyk, Madden, and Dubey \[2013\]](#)), it was necessary to develop faster methods that find the nearest neighbors without explicitly computing all distances from  $q$ . Those methods compute and store additional information about the set  $P$ , which is then used to find nearest neighbors more efficiently. To illustrate this idea, consider another simple solution for the case where  $X = \{0, 1\}^d$ . In this case, one could precompute and store in memory the answers to all  $2^d$  queries  $q \in X$ ; given  $q$ , one could then return its nearest neighbor by performing only a single memory lookup. Unfortunately, this approach requires memory of size  $2^d$ , which again is inefficient ( $d$  is at least  $10^3$  or higher in many applications).

The two aforementioned solutions can be viewed as extreme points in a tradeoff between the time to answer a query (“query time”) and the amount of memory used (“space”).<sup>2</sup> The study of this tradeoff dates back to the work of [Minsky and Papert \[1969, p. 222\]](#)], and has become one of the key topics in the field of *computational geometry* [Preparata and Shamos \[1985\]](#). During the 1970s and 1980s many efficient solutions have been discovered for the case when  $(X, \mathsf{D}) = (\mathbb{R}^d, \ell_2)$  and  $d$  is a constant independent of  $n$ . For example, for  $d = 2$ , one can construct a data structure using  $O(n)$  space with  $O(\log n)$  query time [Lipton and Tarjan \[1980\]](#). Unfortunately, as the dimension  $d$  increases, those data structures become less and less efficient. Specifically, it is known how construct data structures with  $O(d^{O(1)} \log n)$  query time, but using  $n^{O(d)}$  space ([Meiser \[1993\]](#), building on [Clarkson \[1988\]](#)).<sup>3</sup> Furthermore, there is evidence that data structures with query times of the form  $n^{1-\alpha} d^{O(1)}$  for some constant  $\alpha > 0$  might be difficult to construct efficiently.<sup>4</sup>

The search for efficient solutions to the nearest neighbor problem has led to the question whether better space/query time bounds could be obtained if the data structure was allowed to report *approximate* answers. In the *c-approximate nearest neighbor* problem, the data structure can report any point  $p' \in P$  within distance  $c \cdot \min_{p \in P} \mathsf{D}(q, p)$  from  $q$ ; the parameter  $c \geq 1$  is called “approximation factor”. The work of [Arya and Mount](#)

<sup>2</sup>There are other important data structure parameters, such as the time needed to construct it. For the sake of simplicity, we will mostly focus on query time and space.

<sup>3</sup>This exponential dependence on the dimension is due to the fact that those data structures compute and store the *Voronoi decomposition* of  $P$ , i.e., the decomposition of  $\mathbb{R}^d$  into cells such that all points in each cell have the same nearest neighbor in  $P$ . The combinatorial complexity of this decomposition could be as large as  $n^{\Omega(d)}$  [Carathéodory \[1911\]](#).

<sup>4</sup>If such a data structure could be constructed in polynomial time  $n^{O(1)}$ , then the Strong Exponential Time Hypothesis [Vassilevska Williams \[2018\]](#) would be false. This fact essentially follows from [R. Williams \[2005\]](#), see the discussion after Theorem 1 in [Ahle, Pagh, Razenshteyn, and Silvestri \[2016\]](#).

[1993] and Bern [1993] showed that allowing  $c > 1$  indeed leads to better data structures, although their solutions still retained exponential dependencies on  $d$  in the query time or space bounds Arya and Mount [1993] or required that the approximation factor  $c$  be polynomial in the dimension  $d$  Bern [1993]. These bounds have been substantially improved over the next few years, see e.g., Clarkson [1994], Chan [1998], Arya, Mount, Netanyahu, Silverman, and A. Y. Wu [1998], and Kleinberg [1997] and the references therein.

In this article we survey the “second wave” of approximate nearest neighbor data structures, whose query time and space bounds are polynomial in the dimension  $d$ .<sup>5</sup> At a high level, these data structures are obtained in two steps. In the first step, the approximate *nearest* neighbor problem is reduced to its “decision version”, termed approximate *near* neighbor (see e.g. Har-Peled, Indyk, and Motwani [2012]). The second step involves constructing a data structure for the latter problem. In this survey we focus mostly on the second step.

The approximate near neighbor problem is parameterized by an approximation factor  $c > 1$  as well as a “scale parameter”  $r > 0$ , and defined as follows.

**Definition 1.1** ( $((c, r)$ -Approximate Near Neighbor). *Given a set  $P$  of  $n$  points in a metric space  $(X, D)$ , build a data structure  $\mathcal{S}$  that, given any query point  $q \in X$  such that the metric ball  $B_D(q, r) = \{p \in X : D(p, q) \leq r\}$  contains a point in  $P$ ,  $\mathcal{S}$  returns any point in  $B_D(q, cr) \cap P$ .*

Note that the definition does not specify the behavior of the data structure if the ball  $B_D(q, r)$  does not contain any point in  $P$ . We omit the index  $D$  when it is clear from the context.

The above definition applies to algorithms that are *deterministic*, i.e., do not use random bits. However, most of the approximate near neighbor algorithms in the literature are *randomized*, i.e., generate and use random bits while constructing the data structure. In this case, the data structure  $\mathcal{S}$  is a random variable, selected uniformly at random from some distribution. This leads to the following generalization.

**Definition 1.2** ( $((c, r, \delta)$ -Approximate Near Neighbor). *Given a set  $P$  of  $n$  points in a metric space  $(X, D)$ , build a data structure  $\mathcal{S}$  that, given any query point  $q \in X$  such that  $B(q, r) \cap P \neq \emptyset$ ,*

$$\Pr[\mathcal{S} \text{ returns any point in } B(q, cr) \cap P] \geq 1 - \delta$$

The probability of failure  $\delta$  of the data structure can be reduced by independently repeating the process several times, i.e., creating several data structures. Therefore, in the

<sup>5</sup>Due to the lack of space, we will not cover several important related topics, such as data structures for point-sets with low intrinsic dimension Clarkson [2006], approximate furthest neighbor, approximate nearest line search Mahabadi [2014] and other variants of the problem.

rest of the survey we will set  $\delta$  to an arbitrary constant, say,  $1/3$ . We will use  $(c, r)$ -ANN to denote  $(c, r, 1/3)$ -Approximate Near Neighbor.

**1.1 Computational model.** For the purpose of this survey, a data structure of size  $M$  is an array  $A[1 \dots M]$  of numbers (“the memory”), together with an associated algorithm that, given a point  $q$ , returns a point in  $P$  as specified by the problem. The entries  $A[i]$  of  $A$  are called “memory cells”. Due to lack of space, we will not formally define other details of the computational model, in particular what an algorithm is, how to measure its running time, what is the range of the array elements  $A[i]$ , etc. There are several ways of defining these notions, and the material in this survey is relatively robust to the variations in the definitions. We note, however, that one way to formalize these notions is to restrict all numbers, including point coordinates, memory entries, etc, to rational numbers of the form  $a/b$ , where  $a \in \{-n^{O(1)} \dots n^{O(1)}\}$  and  $b = n^{O(1)}$ , and to define query time as the maximum number of memory cells accessed to answer any query  $q$ .

For an overview of these topics and formal definitions, the reader is referred to [Miltersen \[1999\]](#). For a discussion specifically geared towards mathematical audience, see [Fefferman and Klartag \[2009\]](#).

## 2 Data-independent approach

The first approach to the approximate near neighbor problem has been via data-independent data structures. These are data structures where the memory cells accessed by the query algorithm do not depend on the data set  $P$ , but only on  $q$  and (for randomized data structures) the random bits used to construct the data structure. In this section, we describe two methods for constructing such data structures, based on *oblivious dimension-reduction*, and on *randomized space partitions*. These methods give ANN data structures for the  $\ell_1$  and  $\ell_2$  spaces in particular.

**2.1 ANN via dimension reduction.** As described in the introduction, there exist ANN data structures with space and query time at most exponential in the dimension  $d$ . Since exponential space/time bounds are unaffordable for large  $d$ , a natural approach is to perform a *dimension reduction* beforehand, and then solve the problem in the lower, reduced dimension. The main ingredient of such an approach is a map  $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$  that preserves distances up to a  $c = 1 + \varepsilon$  factor, where  $k = O(\log n)$ . Then a space bound exponential in  $k$  becomes polynomial in  $n$ .

Such dimension-reducing maps  $f$  indeed exist for the  $\ell_2$  norm if we allow randomization, as first shown in the influential paper by Johnson and Lindenstrauss:

**Lemma 2.1** (Johnson and Lindenstrauss [1984]). Fix dimension  $d \geq 1$  and a “target” dimension  $k < d$ . Let  $A$  be the projection of  $\mathbb{R}^d$  to its  $k$ -dimensional subspace selected uniformly at random (with respect to the Haar measure), and define  $f : \mathbb{R}^d \rightarrow \mathbb{R}^k$  as  $f(x) = \frac{\sqrt{d}}{\sqrt{k}} Ax$ . Then, there is a universal constant  $C > 0$ , such that for any  $\varepsilon \in (0, 1/2)$ , and any  $x, y \in \mathbb{R}^d$ , we have that

$$\Pr_A \left[ \frac{\|f(x) - f(y)\|}{\|x - y\|} \in (1 - \varepsilon, 1 + \varepsilon) \right] \geq 1 - e^{-C\varepsilon^2 k}.$$

We can now apply this lemma, with  $k = O\left(\frac{\log n}{\varepsilon^2}\right)$ , to a set of points  $P$  to show that the map  $f$  has a  $(1 + \varepsilon)$  distortion on  $P$ , with probability at least  $2/3$ . Most importantly the map  $f$  is “oblivious”, i.e., it does not depend on  $P$ .

We now show how to use Lemma 2.1 to design a  $(1 + O(\varepsilon), r)$ -ANN data structure with the following guarantees.

**Theorem 2.2** (Indyk and Motwani [1998] and Har-Peled, Indyk, and Motwani [2012]). Fix  $\varepsilon \in (0, 1/2)$  and dimension  $d \geq 1$ . There is a  $(1 + O(\varepsilon), r)$ -ANN data structure over  $(\mathbb{R}^d, \ell_2)$  achieving  $Q = O(d \cdot \frac{\log n}{\varepsilon^2})$  query time, and  $S = n^{O(\log(1/\varepsilon)/\varepsilon^2)} + O(d(n + k))$  space. The time needed to build the data structure is  $O(S + ndk)$ .

*Proof sketch.* First, assume there is a  $(1 + \varepsilon, r)$ -ANN data structure  $\mathcal{Q}$  for the  $k$ -dimensional  $\ell_2$  space, achieving query time  $Q(n, k)$  and space bounded by  $S(n, k)$ . For  $k = O\left(\frac{\log n}{\varepsilon^2}\right)$ , we consider the map  $f$  from Lemma 2.1. For the dataset  $P$ , we compute  $f(P)$  and preprocess this set using  $\mathcal{Q}$  (with the scale parameter  $r(1 + \varepsilon)$ ). Then, for a query point  $q \in \mathbb{R}^d$ , we query the data structure  $\mathcal{Q}$  on  $f(q)$ . This algorithm works for a fixed dataset  $P$  and query  $q$  with  $5/6$  probability, by applying Lemma 2.1 to the points in the set  $P \cup \{q\}$ . The map  $f$  preserves all distances between  $P$  and  $q$  up to a factor of  $1 + \varepsilon$ .

We now construct  $\mathcal{Q}$  with space  $S(n, k) = n \cdot (1/\varepsilon)^{O(k)}$  and  $Q(n, k) = O(k)$ , which yields the stated bound for  $k = O\left(\frac{\log n}{\varepsilon^2}\right)$ . Given the scale parameter  $r$ , we discretize the space  $\mathbb{R}^k$  into cubes of sidelength  $\varepsilon r / \sqrt{k}$ , and consider the set  $S$  of cubes that intersect any ball  $B(p', r)$  where  $p' \in f(P)$ . Using standard estimates on the volume of  $\ell_2$  balls, one can prove that  $|S| \leq n \cdot (1/\varepsilon)^{O(k)}$ . The data structure then stores the set  $S$  in a dictionary data structure.<sup>6</sup> For a query  $f(q)$ , we just compute the cube that contains  $f(q)$ , and check whether it is contained in set  $S$  using the dictionary data structure. We note that there is an additional  $1 + \varepsilon$  factor loss from discretization since the diameter of a cube is  $\varepsilon r$ .  $\square$

<sup>6</sup>In the dictionary problem, we are given a set  $S$  of elements from a discrete universe  $U$ , and we need to answer queries of the form “given  $x$ , is  $x \in S$ ?”. This is a classic data structure problem and has many solutions. One concrete solution is via hashing Cormen, Leiserson, Rivest, and Stein [2001], which achieves space of  $O(|S|)$  words, each of  $O(\log |U|)$  bits, and query time of  $O(1)$  in expectation.

A similar approach was introduced [Kushilevitz, Ostrovsky, and Rabani \[2000\]](#) in the context of the Hamming space  $\{0, 1\}^d$ . An important difference is that there is no analog of [Lemma 2.1](#) for the Hamming space [Brinkman and M. Charikar \[2005\]](#).<sup>7</sup> Therefore, [Kushilevitz, Ostrovsky, and Rabani \[2000\]](#) introduce a weaker notion of randomized dimension reduction, which works only for a *fixed scale*  $r$ .

**Lemma 2.3** ([Kushilevitz, Ostrovsky, and Rabani \[ibid.\]](#)). *Fix the error parameter  $\varepsilon \in (0, 1/2)$ , dimension  $d \geq 1$ , and scale  $r \in [1, d]$ . For any  $k \geq 1$ , there exists a randomized map  $f : \{0, 1\}^d \rightarrow \{0, 1\}^k$  and an absolute constant  $C > 0$ , satisfying the following for any fixed  $x, y \in \{0, 1\}^d$ :*

- if  $\|x - y\|_1 \leq r$ , then  $\Pr_f[\|f(x) - f(y)\|_1 \leq k/2] \geq 1 - e^{-C\varepsilon^2 k}$ ;
- if  $\|x - y\|_1 \geq (1 + \varepsilon)r$ , then  $\Pr_f[\|f(x) - f(y)\|_1 > (1 + \varepsilon/2) \cdot k/2] \geq 1 - e^{-C\varepsilon^2 k}$ .

The map  $f$  can be constructed via a random projection over  $GF(2)$ . That is, take  $f(x) = Ax$ , where  $A$  is a  $k \times d$  matrix for  $k = O(\log(n)/\varepsilon^2)$ , with each entry being 1 with some fixed probability  $p$ , and zero otherwise. The probability  $p$  depends solely on  $r$ . The rest of the algorithm proceeds as before, with the exception that the “base” data structure  $\mathcal{Q}$  is particularly simple: just store the answer for any dimension-reduced query point  $f(q) \in \{0, 1\}^k$ . Since there are only  $2^k = n^{O(1/\varepsilon^2)}$  such possible queries, and computing  $f(q)$  takes  $O(dk)$  time, we get the following result.

**Theorem 2.4** ([Kushilevitz, Ostrovsky, and Rabani \[ibid.\]](#)). *Fix  $\varepsilon \in (0, 1/2)$  and dimension  $d \geq 1$ . There is a  $(1 + O(\varepsilon), r)$ -ANN data structure over  $(\{0, 1\}^d, \ell_1)$  using  $n^{O(1/\varepsilon^2)} + O(d(n + k))$  space and  $O(d \cdot \frac{\log n}{\varepsilon^2})$  query time.*

As a final remark, we note we cannot obtain improved space bounds by improving the dimension reduction lemmas [2.1](#) and [2.3](#). Indeed the above lemma are tight as proven in [Jayram and Woodruff \[2013\]](#). There was however work on improving the *run-time complexity* for computing a dimension reduction map, improving over the naïve bound of  $O(dk)$ ; see [Ailon and Chazelle \[2009\]](#), [Dasgupta, R. Kumar, and Sarlós \[2010\]](#), [Ailon and Liberty \[2013\]](#), [Krahmer and Ward \[2011\]](#), [Nelson, Price, and Wootters \[2014\]](#), and [Kane and Nelson \[2014\]](#).

**2.2 ANN via space partitions: Locality-Sensitive Hashing.** While dimension reduction yields ANN data structure with polynomial space, this is not enough in applications, where one desires space as close as possible to linear in  $n$ . This consideration led to the

---

<sup>7</sup>In fact, it has been shown that spaces for which analogs of [Lemma 2.1](#) hold are “almost” Hilbert spaces [Johnson and A. Naor \[2009\]](#).

following, alternative approach, which yields smaller space bounds, albeit at the cost of increasing the query time to something of the form  $n^\rho$  where  $\rho \in (0, 1)$ .

The new approach is based on randomized space partitions, and specifically on Locality-Sensitive Hashing, introduced in [Indyk and Motwani \[1998\]](#).

**Definition 2.5** (Locality-Sensitive Hashing (LSH)). *Fix a metric space  $(X, D)$ , scale  $r > 0$ , approximation  $c > 1$  and a set  $U$ . Then a distribution  $\mathcal{H}$  over maps  $h : X \rightarrow U$  is called  $(r, cr, p_1, p_2)$ -sensitive if the following holds for any  $x, y \in X$ :*

- if  $D(x, y) \leq r$ , then  $\Pr_h[h(x) = h(y)] \geq p_1$ ;
- if  $D(x, y) > cr$ , then  $\Pr_h[h(x) = h(y)] \leq p_2$ .

The distribution  $\mathcal{H}$  is called an LSH family, and has quality  $\rho = \rho(\mathcal{H}) = \frac{\log 1/p_1}{\log 1/p_2}$ .

In what follows we require an LSH family to have  $p_1 > p_2$ , which implies  $\rho < 1$ . Note that LSH mappings are also oblivious: the distribution  $\mathcal{H}$  does not depend on the point-set  $P$  or the query  $q$ .

Using LSH, [Indyk and Motwani \[ibid.\]](#) show how to obtain the following ANN data structure.

**Theorem 2.6** ([Indyk and Motwani \[ibid.\]](#)). *Fix a metric  $\mathfrak{M} = (X, D)$ , a scale  $r > 0$ , and approximation factor  $c > 1$ . Suppose the metric admits a  $(r, cr, p_1, p_2)$ -sensitive LSH family  $\mathcal{H}$ , where the map  $h(\cdot)$  can be stored in  $\sigma$  space, and, for given  $x$ , can be computed in  $\tau$  time; similarly, assume that computing distance  $D(x, y)$  takes  $O(\tau)$  time. Let  $\rho = \rho(\mathcal{H}) = \frac{\log 1/p_1}{\log 1/p_2}$ . Then there exists a  $(c, r)$ -ANN data structure over  $\mathfrak{M}$  achieving query time  $Q = O(n^\rho \cdot \tau \frac{\log_{1/p_2} n}{p_1})$  and space  $S = O(n^{1+\rho} \cdot \frac{1}{p_1} + n^\rho \frac{1}{p_1} \cdot \sigma \cdot \log_{1/p_2} n)$  (in addition to storing the original dataset  $P$ ). The time needed to build this data structure is  $O(S \cdot \tau)$ .*

While we describe some concrete LSH families later on, for now, one can think of the parameters  $\tau, \sigma$  as being proportional to the dimension of the space (although this is not always the case).

The overall idea of the algorithm is to use an LSH family as a pre-filter for the dataset  $P$ . In particular, for a random partition  $h$  from the family  $\mathcal{H}$ , the query point  $q$  will likely collide with its near neighbor (with probability at least  $p_1$ ), but with few points at a distance  $\geq cr$ , in expectation at most  $p_2 \cdot n$  of them. Below we show how an extension of this idea yields [Theorem 2.6](#).

*Proof sketch.* Given an LSH family  $\mathcal{H}$ , we can build a new, derived LSH family via a certain tensoring operation. In particular, for an integer  $k \geq 1$ , consider a new distribution  $\mathcal{G}_k$  over maps  $g : X \rightarrow U$ , where  $g(\cdot)$  is obtained by picking  $k$  i.i.d. functions  $h_1, \dots, h_k$

chosen from  $\mathcal{H}$  and setting  $g(x) = (h_1(x), h_2(x), \dots, h_k(x))$ . Then, if  $\mathcal{H}$  is  $(r, cr, p_1, p_2)$ -sensitive,  $\mathcal{G}_k$  is  $(r, cr, p_1^k, p_2^k)$ -sensitive. Note that the parameter  $\rho$  of the hash family does not change, i.e.,  $\rho(\mathcal{G}_k) = \rho(\mathcal{H})$ .

The entire ANN data structure is now composed of  $L$  dictionary data structures (e.g., hash tables discussed in the previous section), where  $L, k \geq 1$  are parameters to fix later. The  $i$ -th hash table is constructed as follows. Pick a map  $g_i$  uniformly at random from  $\mathcal{G}_k$ , and store the set  $g_i(P)$  in the dictionary structure. At the query time, we iterate over  $i = 1 \dots L$ . For a given  $i$ , we compute  $g_i(q)$ , and use the dictionary structure to obtain the set of “candidate” points  $Q_i = \{p \in P : g_i(p) = g_i(q)\}$ . For each candidate point we compute the distance from  $q$  to that point. The process is stopped when all  $Q_i$ ’s are processed, or when a point within distance  $cr$  to  $q$  is found, whichever happens first.

To analyze the success probability, we note that the dictionary structure  $i$  succeeds if  $p^* \in Q_i$ , where  $p^*$  is the assumed point at distance at most  $r$  from  $q$ . This happens with probability at least  $p_1^k$ . Thus, we can take  $L = O(1/p_1^k)$  such dictionary structures, and thus guarantee success with a constant probability.

The expected query time is  $O(L(k\tau + Ln \cdot p_2^k \cdot \tau))$ , which includes both the computation of the maps  $g_1(q), \dots, g_L(q)$  and the of distances to the candidates in sets  $Q_1, \dots, Q_L$ . We can now derive the value of  $k$  that minimizes the above, obtaining  $k = \lceil \log_{1/p_2} n \rceil \leq \log_{1/p_2} n + 1$ , and hence  $L = O(n^\rho / p_1)$ .

Finally, note that the space usage is  $O(Ln)$  for the dictionary structures, plus  $O(Lk\sigma)$  for the description of the maps.  $\square$

**2.3 Space partitions: LSH constructions.** [Theorem 2.6](#) assumes the existence of an LSH family  $\mathcal{H}$  with a parameter  $\rho < 1$ . In what follows we show a few examples of such families.

1. *Hamming space*  $\{0, 1\}^d$ , with  $\rho = 1/c$ . The distribution  $\mathcal{H}$  is simply a projection on a random coordinate  $i$ :  $\mathcal{H} = \{h_i : h_i(x) = x_i, i = 1, \dots, d\}$ . This family is  $(r, cr, 1 - r/d, 1 - cr/d)$ -sensitive, and hence  $\rho \leq 1/c$  [Indyk and Motwani \[1998\]](#).

This LSH scheme is near-optimal for the Hamming space, as described in [Section 2.4](#). We also note that, since  $\ell_2$  embeds isometrically into  $\ell_1$  (see [Section 6](#)), this result extends to  $\ell_2$  as well.

2. *Euclidean space*  $(\mathbb{R}^d, \ell_2)$ , with  $\rho < 1/c$ . In [Datar, Immorlica, Indyk, and Mirrokni \[2004\]](#), the authors introduced an LSH family which slightly improves over the above construction. It is based on random projections in  $\ell_2$ . In particular, define a random map  $h(x)$  as  $h(x) = \lfloor \frac{\langle x, g \rangle}{wr} + b \rfloor$ , where  $g$  is a random  $d$ -dimensional Gaussian vector,  $b \in [0, 1]$ , and  $w > 0$  is a fixed parameter. It can be shown that, for any fixed  $c > 1$ , there exists  $w > 0$  such that  $\rho < 1/c$ .

3. *Euclidean space*  $(\mathbb{R}^d, \ell_2)$ , with  $\rho \rightarrow 1/c^2$ . In [Andoni and Indyk \[2006\]](#), the authors showed an LSH family with a much better  $\rho$ , which later turned out to be optimal (see [Section 2.4](#)). At its core, the main idea is to partition the space into Euclidean balls.<sup>8</sup> It proceeds in two steps: 1) perform a random dimension reduction  $A$  to dimension  $t$  (a parameter), and 2) partition  $\mathbb{R}^t$  into balls. Since it is impossible to partition the space  $\mathbb{R}^t$  into balls precisely<sup>9</sup> when  $t \geq 2$ , instead one performs “ball carving”. The basic idea is to consider a sequence of randomly-centered balls  $B_1, B_2, \dots$ , each of radius  $wr$  for some parameter  $w > 1$ , and define the map  $h(x)$ , for  $x \in \mathbb{R}^d$ , to be the index  $i$  of the first ball  $B_i$  containing the point  $Ax$ . Since we want to cover an infinite space with balls of finite volume, the above procedure needs to be modified slightly to terminate in finite time. The modified procedure runs in time  $T = t^{O(t)}$ .

Overall, optimizing for  $w$ , one can obtain  $\rho = 1/c^2 + \frac{O(\log t)}{\sqrt{t}}$  which tends to  $1/c^2$  as  $t \rightarrow \infty$ . The time to hash is  $\tau = O(Tt + dt)$ , where  $T$  depends exponentially on the parameter  $t$ , i.e.,  $T = t^{\Theta(t)}$ . For the ANN data structure, the optimal choice is  $t = O(\log n)^{2/3}$ , resulting in  $\rho = 1/c^2 + \frac{O(\log \log n)}{(\log n)^{1/3}}$ .

The  $\ell_2$  LSH families can be extended to other  $\ell_p$ ’s. For  $p < 1$ , [Datar, Immorlica, Indyk, and Mirrokni \[2004\]](#) showed one can use method 2 as described above, but using  $p$ -stable distributions instead of Gaussians. See [Section 6](#) for other extensions for  $p > 1$ .

We remark that there is a number of other widely used LSH families, including min-hash [Broder \[1997\]](#) and [Broder, Glassman, Manasse, and Zweig \[1997\]](#) and simhash [M. S. Charikar \[2002\]](#), which apply to different notions of similarity between points. See [Andoni and Indyk \[2008\]](#) for an overview.

**2.4 Space partitions: impossibility results.** It is natural to explore the limits of LSH families and ask what is the best  $\rho$  one can obtain for a given metric space as a function of the approximation  $c > 1$ . In [Motwani, A. Naor, and Panigrahy \[2007\]](#) and [O’Donnell, Y. Wu, and Zhou \[2014\]](#), it was proven that the LSH families [Indyk and Motwani \[1998\]](#) and [Andoni and Indyk \[2006\]](#) from the previous section are near-optimal: for the Hamming space, we must have  $\rho \geq 1/c - o(1)$ , and for the Euclidean space,  $\rho \geq 1/c^2 - o(1)$ . Below is the formal statement from [O’Donnell, Y. Wu, and Zhou \[2014\]](#).

**Theorem 2.7.** *Fix dimension  $d \geq 1$  and approximation  $c \geq 1$ . Let  $\mathcal{H}$  be a  $(r, cr, p_1, p_2)$ -sensitive LSH family over the Hamming space, and suppose  $p_2 \geq 2^{-o(d)}$ . Then  $\rho \geq 1/c - o_d(1)$ .*

<sup>8</sup>In contrast, the above LSH family can be seen as partitioning the space into cubes: when considering the  $k$ -tensored family  $\mathcal{G} = \mathcal{H}^k$ , the resulting map  $g \in \mathcal{G}$  is equivalent to performing a random dimension reduction (by multiplying by a random  $k \times d$  Gaussian matrix), followed by discretization of the space into cubes.

<sup>9</sup>This is also termed *tessellation* of the space.



Note that the above theorem also immediately implies  $\rho \geq 1/c^2 - o(1)$  for the Euclidean space, by noting that  $\|x - y\|_1 = \|x - y\|_2^2$  for binary vectors  $x$  and  $y$ .

Finally, we remark that some condition on  $p_2$  is necessary, as there exists an LSH family with  $p_2 = 0$ ,  $p_1 = 2^{-O(d)}$  and hence  $\rho = 0$ . To obtain the latter, one can use the “ball carving” family of [Andoni and Indyk \[2006\]](#), where the balls have radius  $wr = cr/2$ . Note however that such a family results in query time that is at least exponential in  $d$ , which LSH algorithms are precisely designed to circumvent.

### 3 (More) Deterministic algorithms

A drawback of data structures described in the previous section is that they allow “false negatives”: with a controllable but non-zero probability, the data structure can report nothing even if the ball  $B(q, r)$  is non-empty. Although most of the data structures described in the literature have this property, it is possible to design algorithms with stronger guarantees, including deterministic ones.

The first step in this direction was an observation (already made in [Kushilevitz, Ostrovsky, and Rabani \[2000\]](#)) that for a finite metric  $(X, D)$  supported by  $(c, r)$ -ANN data structures, it is possible to construct a data structure that provides accurate answers to *all* queries  $q \in X$ . This is because one can construct and use  $O(\log |X|)$  independent data structures, reducing the probability of failure to  $\frac{1}{3^{|X|}}$ . By taking a union bound over all  $q \in X$ , the constructed data structure works, with probability at least  $2/3$ , for all queries  $X$ . Note that the space and query time bounds of the new data structure are  $O(\log |X|)$  times larger than the respective bounds for  $(c, r)$ -ANN. Unfortunately, the algorithm for constructing such data structures has still a non-zero failure probability, and no deterministic polynomial-time algorithm for this task is known.

The first deterministic polynomial-time algorithm for constructing a data structure that works for all queries  $q \in X$  appeared in [Indyk \[2000a\]](#). It was developed for  $d$ -dimensional Hamming spaces, and solved a  $(c, r)$ -ANN with an approximation factor  $c = 3 + \varepsilon$  for any  $\varepsilon > 0$ . The data structure had  $d(1/\varepsilon)^{O(1)}$  query time and used  $dn^{(1/\varepsilon)^{O(1)}}$  space. It relied on two components. The first component, “densification”, was a deterministic analog of the mapping in [Lemma 2.3](#), which was shown to hold with  $k = (d/\varepsilon)^{O(1)}$ . Retrospectively, the mapping can be viewed as being induced by an adjacency matrix of an *unbalanced expander* [Guruswami, Umans, and Vadhan \[2009\]](#).

**Definition 3.1** (Expander). An  $(r, \alpha)$ -unbalanced expander is a bipartite simple graph  $G = (U, V, E)$ ,  $|U| = d$ ,  $|V| = k$ , with left degree  $\Delta$  such that for any  $X \subset U$  with  $|X| \leq r$ , the set of neighbors  $N(X)$  of  $X$  has size  $|N(X)| \geq (1 - \alpha)\Delta|X|$ .

Given such a graph  $G$ , one can construct a mapping  $f = f_G : \{0, 1\}^d \rightarrow \Sigma^k$  for some finite alphabet  $\Sigma$  by letting  $f(x)_j$  to be the concatenation of all symbols  $x_i$  such

that  $(i, j) \in E$ . Let  $H(x, y)$  be the Hamming distance between  $x$  and  $y$ , i.e., the number of coordinates on which  $x$  and  $y$  differ. We have that:

- $H(f(x), f(y)) \leq \Delta H(x, y)$ , since each difference between  $a$  and  $b$  contributes to at most  $\Delta$  differences between  $f(x)$  and  $f(y)$ . In particular  $H(f(x), f(y)) \leq \Delta r(1 - \varepsilon)$  if  $H(x, y) \leq r(1 - \varepsilon)$ .
- if  $H(x, y) \geq r$ , then  $H(f(x), f(y)) \geq (1 - \alpha)\Delta r$  (from the expansion property).

Thus, setting  $\alpha = \varepsilon/2$  yields guarantees analogous to [Lemma 2.3](#), but using a deterministic mapping, and with coordinates of  $f(x)$  in  $\Sigma$ , not  $\{0, 1\}$ . To map into binary vectors, we further replace each symbol  $f(x)_j$  by  $C(f(x)_j)$ , where  $C : \Sigma \rightarrow \{0, 1\}^s$  is an *error-correcting code*, i.e., having the property that for any distinct  $a, b \in \Sigma$  we have  $H(C(a), C(b)) \in [s(1/2 - \varepsilon), s(1/2 + \varepsilon)]$ . We then use off-the-shelf constructions of expanders [Guruswami, Umans, and Vadhan \[ibid.\]](#) and codes [Guruswami, Rudra, and Sudan \[2014\]](#) to obtain the desired mapping  $g = C \circ f : \{0, 1\}^d \rightarrow \{0, 1\}^{ks}$ .

The second component partitions the coordinates of points  $g(x)$  into blocks  $S_1 \dots S_t$  of size  $\log(n)/\varepsilon^{O(1)}$  such that an analog of [Lemma 2.3](#) holds for all projections  $g(x)_{S_l}$  and  $g(y)_{S_l}$  where  $x, y \in P$ ,  $l = 1 \dots t$ . Such a partitioning can be shown to exist using the probabilistic method, and can be computed deterministically in time polynomial in  $n$  via the method of conditional probabilities. Unfortunately, this property does not extend to the case where one of the points (say,  $x$ ) is a query point from  $X - P$ . Nevertheless, by averaging, there must be at least one block  $S_l$  such that  $H(g(x)_{S_l}, g(y)_{S_l}) \leq H(g(x), g(y))/t$ , where  $y$  is the nearest neighbor of  $x$  in  $P$ . It can be then shown that an approximate near neighbor of  $g(x)_{S_l}$  in  $\{g(y)_{S_l} : y \in P\}$  is an approximate nearest neighbor of  $x$  in  $P$ . Finding the nearest neighbor in the space restricted to a single block  $S_l$  can be solved via exhaustive storage using  $n^{1/\varepsilon^{O(1)}}$  space, as in [Theorem 2.4](#).

Perhaps surprisingly, the above construction is the only known example of a polynomial-size deterministic approximate near neighbor data structure with a constant approximation factor. However, more progress has been shown for an “intermediary” problem, where the data structure avoids false negatives by reporting a special symbol  $\perp$ .

**Definition 3.2** ( $((c, r, \delta)$ -Approximate Near Neighbor Without False Negatives (ANNWFN)). *Given a set  $P$  of  $n$  points in a metric space  $(X, D)$ , build a data structure  $\mathcal{S}$  that, given any query point  $q \in X$  such that  $B(q, r) \cap P \neq \emptyset$ ,  $\mathcal{S}$  returns an element of  $(B(q, cr) \cap P) \cup \{\perp\}$ , and  $\Pr_{\mathcal{S}}[\mathcal{S} \text{ returns } \perp] \leq \delta$ .*

A  $(1 + \varepsilon, r, \delta)$ -ANNWFN data structure with bounds similar to those in [Theorem 2.4](#) was given in [Indyk \[2000a\]](#). It used densification and random block partitioning as described above. However, thanks to randomization, block partitioning could be assumed to hold even for the query point with high probability.

Obtaining “no false negatives” analogs of [Theorem 2.2](#) turned out to be more difficult. The first such data structure was presented in [Pagh \[2016\]](#), for the Hamming space, achieving query time of the form (roughly)  $dn^{1.38/c}$ . Building on that work, very recently, [Ahle \[2017\]](#) improved the bound to (roughly)  $dn^{1/c}$ , achieving the optimal runtime exponent.

In addition to variants of densification and random block partitioning, the latter algorithm uses a generalization of the space partitioning method from [Section 2.2](#), called *locality sensitive filtering*. Such objects can be constructed deterministically in time and space roughly exponential in the dimension. Unfortunately, random block partitioning leads to blocks whose length is larger than  $\log n$  by at least a (large) constant, which results in large (although polynomial) time and space bounds. To overcome this difficulty, [Ahle \[ibid.\]](#) shows how to combine filters constructed for dimension  $d$  to obtain a filter for dimension  $2d$ . This is achieved by using *splitters* [M. Naor, Schulman, and Srinivasan \[1995\]](#), which can be viewed as families of partitions of  $\{1 \dots 2d\}$  into pairs of sets  $(S_1, \overline{S_1}), (S_2, \overline{S_2}), \dots$  of size  $d$ , such that for any  $x, y$ , there is a pair  $(S_l, \overline{S_l})$  for which  $H(x_{S_l}, y_{S_l}) = H(x_{\overline{S_l}}, y_{\overline{S_l}}) \pm 1$ . The construction multiplies the space bound by a factor quadratic in  $d$ , which makes it possible to apply it a small but super-constant number of times to construct filters for (slightly) super-logarithmic dimension.

## 4 Data-dependent approach

In the earlier sections, we considered ANN data structures that are based on random and deterministic space partitions. The unifying feature of all of the above approaches is that the partitions used are independent of the dataset. This “data-independence” leads to certain barriers: for instance, the *best possible* LSH exponent is  $\rho \geq 1/c - o(1)$  for the  $\ell_1$  distance and  $\rho \geq 1/c^2 - o(1)$  for  $\ell_2$  (see [Section 2.4](#)). In this section, we show how to improve upon the above results significantly if one allows partitions to depend on the dataset.

This line of study has been developed in a sequence of recent results [Andoni, Indyk, H. L. Nguyễn, and Razenshteyn \[2014\]](#), [Andoni and Razenshteyn \[2015\]](#), and [Andoni, Laarhoven, Razenshteyn, and Waingarten \[2017\]](#). However, even before these works, the data-dependent approach had been very popular in practice (see, e.g., surveys [Wang, Shen, Song, and Ji \[2014\]](#) and [Wang, Liu, S. Kumar, and Chang \[2016\]](#)). Indeed, real-world datasets often have some implicit or explicit structure, thus it pays off to tailor space partitions to a dataset at hand. However, the theoretical results from [Andoni, Indyk, H. L. Nguyễn, and Razenshteyn \[2014\]](#), [Andoni and Razenshteyn \[2015\]](#), and [Andoni, Laarhoven, Razenshteyn, and Waingarten \[2017\]](#) improve upon data-independent partitions for *arbitrary* datasets. Thus, one must show that *any* set of  $n$  points has some structure that makes the ANN problem easier.

**4.1 The result.** In Andoni and Razenshteyn [2015] (improving upon Andoni, Indyk, H. L. Nguyen, and Razenshteyn [2014]), the following result has been shown.

**Theorem 4.1.** *For every  $c > 1$ , there exists a data structure for  $(c, r)$ -ANN over  $(\mathbb{R}^d, \ell_2)$  with space  $n^{1+\rho} + O(nd)$  and query time  $n^\rho + dn^{o(1)}$ , where*

$$\rho \leq \frac{1}{2c^2 - 1} + o(1).$$

This is much better than the best LSH-based data structure, which has  $\rho = \frac{1}{c^2} + o(1)$ . For instance, for  $c = 2$ , the above theorem improves the query time from  $n^{1/4+o(1)}$  to  $n^{1/7+o(1)}$ , while using less memory.

Next, we describe the new approach at a high level.

**4.2 Simplification of the problem.** Before describing new techniques, it will be convenient to introduce a few simplifications. First, we can assume that  $d = \log^{1+o(1)} n$ , by applying Lemma 2.1. Second, we can reduce the general ANN problem over  $(\mathbb{R}^d, \ell_2)$  to the *spherical* case: where dataset and queries lie on the unit sphere  $S^{d-1} \subset \mathbb{R}^d$  (see Razenshteyn [2017], pages 55–56). Both the dimension reduction and the reduction to the spherical case incur a negligible loss in the approximation<sup>10</sup>. After the reduction to the spherical case, the distance to the near neighbor  $r$  can be made to be any function of the number of points  $n$  that tends to zero as  $n \rightarrow \infty$  (for example,  $r = \frac{1}{\log \log n}$ ).

**4.3 Data-independent partitions for a sphere.** In light of the above discussion, we need to solve the  $(c, r)$ -ANN problem for  $S^{d-1}$ , where  $d = \log^{1+o(1)} n$  and  $r = o(1)$ . Even though the final data structure is based on data-dependent partitions, we start with developing a *data-independent* LSH scheme for the unit sphere, which will be later used as a building block.

The LSH scheme is parametrized by a number  $\eta > 0$ . Consider a sequence of i.i.d. samples from a standard  $d$ -dimensional Gaussian distribution  $N(0, 1)^d$ :  $g_1, g_2, \dots, g_t, \dots \in \mathbb{R}^d$ . The hash function  $h(x)$  of the point  $x \in S^{d-1}$  is then defined as  $\min_t \{t \geq 1 \mid \langle x, g_t \rangle \geq \eta\}$ . This LSH family gives the following exponent  $\rho$  for distances  $r$  and  $cr$ :

$$(1) \quad \rho = \frac{\log 1/p_1}{\log 1/p_2} = \frac{4 - c^2 r^2}{4 - r^2} \cdot \frac{1}{c^2} + \delta(r, c, \eta),$$

where  $\delta(r, c, \eta) > 0$  and  $\delta(r, c, \eta) \rightarrow 0$  as  $\eta \rightarrow \infty$ . Thus, the larger the value of the threshold  $\eta$  is, the more efficient the resulting LSH scheme is. At the same time,  $\eta$  affects the efficiency of hash functions. Indeed, one can show that with very high probability

<sup>10</sup>Approximation  $c$  reduces to approximation  $c - o(1)$ .

$\max_{x \in S^{d-1}} h(x) \leq e^{(1+o(1))\eta^2/2} \cdot d^{O(1)}$ , which bounds the hashing time as well as the number of Gaussian vectors to store.

Consider the expression (1) for the exponent  $\rho$  in more detail. If  $r = o(1)$ , then we obtain  $\rho = \frac{1}{c^2} + o(1)$ , which matches the guarantee of the best data-independent LSH for  $\ell_2$ . This is hardly surprising, since, as was mentioned above, the general ANN problem over  $\ell_2$  can be reduced to the  $(c, r)$ -ANN problem over the sphere for  $r = o(1)$ . If  $r \approx 2/c$ , then  $\rho$  is close to zero, and, indeed, the  $(c, 2/c)$ -ANN problem on the sphere is trivial (any point can serve as an answer to any valid query).

Between these two extremes, there is a point  $r \approx \frac{\sqrt{2}}{c}$  that is crucial for the subsequent discussion. Since the distance between a pair of random points on  $S^{d-1}$  is close to  $\sqrt{2}$  with high probability, the problem where  $r$  is *slightly* smaller than  $\frac{\sqrt{2}}{c}$  has the following interpretation: if one is guaranteed to have a data point within distance  $r$  from the query, find a data point that is a bit closer to the query than a typical point on the sphere. For  $r \approx \frac{\sqrt{2}}{c}$ , the Equation (1) gives exponent  $\rho \approx \frac{1}{2c^2-1}$ , which is significantly smaller than the bound  $\frac{1}{c^2}$  one is getting for  $r = o(1)$ . Later, using a certain data-dependent partitioning procedure, we will be able to reduce the *general* ANN problem on the sphere to this intermediate case of  $r \approx \frac{\sqrt{2}}{c}$ , thus obtaining the ANN data structure with the exponent  $\rho = \frac{1}{2c^2-1} + o(1)$ . This significantly improves upon the best possible LSH for  $\ell_2$  from Section 2, which yields  $\rho = \frac{1}{c^2} + o(1)$ .

**4.4 Data-dependent partitions.** We now describe at a high level how to obtain a data structure with space  $n^{1+\rho}$  and query time  $n^\rho$ , where  $\rho = \frac{1}{2c^2-1} + o(1)$ , for the  $(c, r)$ -ANN problem on the sphere for general  $r > 0$ . If  $r \geq \frac{\sqrt{2}}{c} - o(1)$ , then we can simply use the data-independent LSH described above. Now suppose  $r$  is nontrivially smaller than  $\frac{\sqrt{2}}{c}$ .

We start with finding and removing *dense low-diameter clusters*. More precisely, we repeatedly find a point  $u \in S^{d-1}$  such that  $|P \cap B(u, \sqrt{2} - \varepsilon)| \geq \tau n$ , where  $\varepsilon, \tau = o(1)$ , and set  $P := P \setminus B(u, \sqrt{2} - \varepsilon)$ . We stop when there are no more dense clusters remaining. Then we proceed with clusters and the remainder separately. Each cluster is enclosed in a ball of radius  $1 - \Omega(\varepsilon^2)$  and processed recursively. For the remainder, we sample one partition from the data-independent LSH family described above, apply it to the dataset, and process each resulting part of the dataset recursively. During the query stage, we (recursively) query the data structure for *every* cluster (note that the number of clusters is at most  $1/\tau$ ), and for the remainder we query (again, recursively) a part of the partition, where the query belongs to. Each step of the aforementioned procedure makes progress as follows. For clusters, we decrease the radius by a factor of  $1 - \Omega(\varepsilon^2)$ . It means that we come slightly closer to the ideal case of  $r \approx \frac{\sqrt{2}}{c}$ , and the instance corresponding to the cluster becomes easier. For the remainder, we use the fact that there are at most  $\tau n$  data

points closer than  $\sqrt{2} - \varepsilon$  to the query. Thus, when we apply the data-independent LSH, the expected number of data points in the same part as the query is at most  $(\tau + p_2)n$ , where  $p_2$  is the probability of collision under the LSH family for points at the distance  $\sqrt{2} - \varepsilon$ . We set  $\tau \ll p_2$ , thus the number of colliding data points is around  $p_2n$ . At the same time, the probability of collision with the near neighbor is at least  $p_1$ , where  $p_1$  corresponds to the distance  $r$ . Since  $r < \frac{\sqrt{2}}{c}$ , we obtain an effective exponent of at most  $\frac{1}{2c^2-1} + o(1)$ . Note that we need to keep extracting the clusters recursively to be able to apply the above reasoning about the remainder set in each step.

One omission in the above high-level description is that the clusters are contained in smaller *balls* rather than *spheres*. This is handled by partitioning balls into thin annuli and treating them as spheres (introducing negligible distortion).

**4.5 Time–space trade-off.** In [Andoni, Laarhoven, Razenshteyn, and Waingarten \[2017\]](#), [Theorem 4.1](#) has been extended to provide a smooth time–space trade-off for the ANN problem. Namely, it allows to decrease the query time at a cost of increasing the space and vice versa.

**Theorem 4.2.** *For every  $c > 1$  and every  $\rho_s, \rho_q$  such that*

$$(2) \quad c^2 \sqrt{\rho_q} + (c^2 - 1) \sqrt{\rho_s} \geq \sqrt{2c^2 - 1},$$

*there exists a data structure for  $(c, r)$ -ANN over  $(\mathbb{R}^d, \ell_2)$  with space  $n^{1+\rho_s+o(1)} + O(nd)$  and query time  $n^{\rho_q+o(1)} + dn^{o(1)}$ .*

The bound (2) interpolates between:

- The near-linear space regime:  $\rho_s = 0, \rho_q = \frac{2}{c^2} - \frac{1}{c^4}$ ;
- The “balanced” regime:  $\rho_s = \rho_q = \frac{1}{2c^2-1}$ , where it matches [Theorem 4.1](#);
- The fast queries regime:  $\rho_s = \left(\frac{c^2}{c^2-1}\right)^2, \rho_q = 0$ .

For example, for  $c = 2$ , one can obtain any of the following trade-offs: space  $n^{1+o(1)}$  and query time  $n^{7/16+o(1)}$ , space  $n^{8/7+o(1)}$  and query time  $n^{1/7+o(1)}$ , and space  $n^{16/9+o(1)}$  and query time  $n^{o(1)}$ .

[Theorem 4.2](#) significantly improves upon the previous ANN data structures in various regimes [Indyk and Motwani \[1998\]](#), [Kushilevitz, Ostrovsky, and Rabani \[2000\]](#), [Indyk \[2000b\]](#), [Panigrahy \[2006\]](#), and [Kapralov \[2015\]](#). For example, it improves the dependence on  $\varepsilon$  in [Theorem 2.2](#) from  $O(\log(1/\varepsilon)/\varepsilon^2)$  to  $O(1/\varepsilon^2)$ .

**4.6 Impossibility results.** Similarly to the data-independent case, it is natural to ask whether exponent  $\rho = \frac{1}{2c^2-1} + o(1)$  from [Theorem 4.1](#) is optimal for data-dependent space partitions. In [Andoni and Razenshteyn \[2016\]](#), it was shown that the above  $\rho$  is near-optimal in a properly formalized framework of data-dependent space partitions. This impossibility result can be seen as an extension of the results discussed in [Section 2.4](#).

Specifically, [Andoni and Razenshteyn \[ibid.\]](#) show that  $\rho \geq \frac{1}{2c^2-1}$ , where  $\rho = \frac{\log 1/p_1}{\log 1/p_2}$  for  $p_1$  and  $p_2$  being certain natural counterparts of the LSH collision probabilities for the data-dependent case, even when we allow the distribution on the partitions to depend on a dataset. This result holds under two further conditions. First, as in [Section 2.4](#), we need to assume that  $p_2$  is not too small.

The second condition is specific to the data-dependent case, necessary to address another necessary aspect of the space partition. For any dataset, where all the points are sufficiently well separated, we can build an “ideal” space partition, with  $\rho = 0$ , simply by considering its Voronoi diagram. However, this is obviously not a satisfactory space partition: it is algorithmically hard to compute fast where in the partition a fixed query point  $q$  falls to — in fact, it is precisely equivalent to the original nearest neighbor problem! Hence, to be able to prove a meaningful lower bound on  $\rho$ , we would need to restrict the space partitions to have low run-time complexity (e.g., for a given point  $q$ , we can compute the part where  $q$  lies in, in time  $n^{o(1)}$ ). This precise restriction is well beyond reach of the current techniques (it would require proving computational lower bounds). Instead, [Andoni and Razenshteyn \[ibid.\]](#) use a different, proxy restriction: they require that the *description complexity* of partitions is  $n^{1-\Omega(1)}$ . The latter restriction is equivalent to saying that the distribution of partitions (which may depend on the given dataset) is supported on a fixed (universal) family of partitions of the size  $2^{n^{1-\Omega(1)}}$ . This restriction, for instance, rules out the Voronoi diagram, since the latter has a description complexity of  $\Omega(n)$ . Furthermore, the description complexity of a randomized partition is a good proxy for the run-time complexity of a partition because in all the known constructions of random space partitions with a near-optimal  $\rho$ , the run-time complexity is at least the description complexity, which makes the requirement meaningful.

Overall, under the above two conditions, [Andoni and Razenshteyn \[ibid.\]](#) show that  $\rho \geq \frac{1}{2c^2-1} - o(1)$  for data-dependent random space partitions, and hence [Theorem 4.1](#) is essentially optimal in this framework.

**4.7 ANN for  $\ell_\infty$ .** In this subsection we will describe another type of data-dependent data structure, for the  $\ell_\infty$  norm. Historically, this was the first example of a data-dependent partitioning procedure used for ANN over high-dimensional spaces.

**Theorem 4.3** (Indyk [2001]). *For every  $0 < \varepsilon < 1$ , there exists a deterministic data structure for  $(c, 1)$ -ANN for  $(\mathbb{R}^d, \ell_\infty)$  with approximation  $c = O\left(\frac{\log \log d}{\varepsilon}\right)$ , space  $O(dn^{1+\varepsilon})$  and query time  $O(d \log n)$ .*

The algorithm relies on the following structural lemma.

**Lemma 4.4.** *Let  $P \subset \mathbb{R}^d$  be a set of  $n$  points and  $0 < \varepsilon < 1$ . Then:*

1. *Either there exists an  $\ell_\infty$ -ball of radius  $O\left(\frac{\log \log d}{\varepsilon}\right)$  that contains  $\Omega(n)$  points from  $P$ , or*
2. *There exists a “good” coordinate  $i \in \{1, 2, \dots, d\}$  and a threshold  $u \in \mathbb{R}$  such that for the sets  $A = \{p \in P \mid p_i < u - 1\}$ ,  $B = \{p \in P \mid u - 1 \leq p_i \leq u + 1\}$  and  $C = \{p \in P \mid p_i > u + 1\}$  one has:*

$$(3) \quad \left(\frac{|A| + |B|}{n}\right)^{1+\varepsilon} + \left(\frac{|B| + |C|}{n}\right)^{1+\varepsilon} \leq 1$$

*and  $|A|/n, |C|/n \geq \Omega(1/d)$ .*

Using this lemma, we can build the data structure for  $(c, 1)$ -ANN for  $(\mathbb{R}^d, \ell_\infty)$  recursively. If there exists a ball  $B(x, R)$  with  $R = O\left(\frac{\log \log d}{\varepsilon}\right)$  such that  $|P \cap B(x, R)| \geq \Omega(n)$  (Case 1), then we store  $x$  and  $R$  and continue partitioning  $P \setminus B(x, R)$  recursively. If there exists a good coordinate  $i \in \{1, 2, \dots, d\}$  and a threshold  $u \in \mathbb{R}$  (Case 2), then we define sets  $A, B, C$  as in the above lemma and partition  $A \cup B$  and  $B \cup C$  recursively. We stop as soon as we reach a set that consists of  $O(1)$  points.

The query procedure works as follows. Suppose there is a point in  $P$  within distance 1 from  $q$  (“the near neighbor”). If we are in Case 1, we check if the query point  $q$  lies in  $B(x, R + 1)$ . If it does, we return any data point from  $B(x, R)$ ; if not, we query the remainder recursively. On the other hand, if we are in Case 2, we query  $A \cup B$  if  $q_i \leq u$ , and  $B \cup C$  otherwise. In this case we recurse on the part which is guaranteed to contain a near neighbor.

Overall, we always return a point within distance  $O\left(\frac{\log \log d}{\varepsilon}\right)$ , and it is straightforward to bound the query time by bounding the depth of the tree. We obtain the space bound of  $O(dn^{1+\varepsilon})$  by using the property (3) to bound the number of times points that are replicated in the Case 2 nodes.

Surprisingly, the approximation  $O(\log \log d)$  turns out to be *optimal* in certain restricted models of computation Andoni, Croitoru, and Patrascu [2008] and Kapralov and Panigrahy [2012], including for the approach from Indyk [2001].



## 5 Closest pair

A problem closely related to ANN is the closest pair problem, which can be seen as an “offline” version of ANN. Here, we are given a set  $P$  of  $n$  points, and we need to find a pair  $p, q \in P$  of distinct points that minimize their distance.

A trivial solution is to compute the distance between all possible  $\binom{n}{2}$  pairs of points and take the one that minimizes the distance. However this procedure has quadratic running time. As for the nearest neighbor problem, there is evidence that for, say,  $d$ -dimensional  $\ell_2$  space, the closest pair problem cannot be solved in time  $n^{2-\alpha}d^{O(1)}$  for any constant  $\alpha > 0$ .

As with  $c$ -ANN, we focus on the approximate version of the problem. Furthermore, we consider the *decision version*, where we need to find a pair of points that are below a certain threshold  $r$ . The formal definition (for the randomized variant) follows.

**Definition 5.1** ( $(c, r)$ -approximate close pair problem, or  $(c, r)$ -CP). *Given a set of points  $P \subset X$  of size  $n$ , if there exist distinct points  $p^*, q^* \in X$  with  $D(p^*, q^*) \leq r$ , find a pair of distinct points  $p, q \in P$  such that  $D(p, q) \leq cr$ , with probability at least  $2/3$ .*

The  $(c, r)$ -CP problem is closely related to the  $(c, r)$ -ANN problem because we can solve the former using a data structure for the latter. In particular, one can run the following procedure: partition  $P$  into two sets  $A, B$  randomly; build  $(c, r)$ -ANN on the set  $A$ ; query every point  $q \in B$ . It is easy to see that one such run succeeds in solving a  $(c, r)$ -approximate close pair with probability at least  $1/2 \cdot 2/3$ . Repeating the procedure 3 times is enough to guarantee a success probability of  $2/3$ . If  $(c, r)$ -ANN under the desired metric can be solved with query time  $Q(n)$  and preprocessing time  $S(n)$ , we obtain a solution for  $(c, r)$ -CP running in time  $O(S(n) + nQ(n))$ . For example, applying the reduction from above for  $(\mathbb{R}^d, \ell_p)$  space for  $p \in \{1, 2\}$ , we immediately obtain an algorithm running in  $O(dn^{1+\rho})$  time, where  $\rho = \frac{1}{2c^p-1} + o(1)$  (Section 4).

Focusing on the case of  $\ell_2$ , and approximation  $c = 1 + \varepsilon$ , the above algorithm has runtime  $O(n^{2-4\varepsilon+O(\varepsilon^2)}d)$ . It turns out that, for the  $\ell_2$  norm, one can obtain algorithms with a better dependance on  $\varepsilon$ , for small  $\varepsilon$ . In particular, the line of work from Valiant [2015], Karppa, Kaski, and Kohonen [2016], and Alman, Chan, and R. Williams [2016] led to the following algorithm:

**Theorem 5.2** (Alman, Chan, and R. Williams [2016]). *Fix dimension  $d \geq 1$ ,  $r > 0$ , and  $\varepsilon \in (0, 1/2)$ . Then, for any set of  $n$  points in  $\mathbb{R}^d$ , one can solve the  $(1 + \varepsilon, r)$ -CP over  $\ell_2$  in time  $O(n^{2-\Omega(\varepsilon^{1/3}/\log(1/\varepsilon))} + dn)$ , with constant probability.*

Note that the running time bound in the above theorem is better than that obtained using LSH data structures, for small enough  $\varepsilon$ .

The main new technical ingredient is the *fast matrix multiplication* algorithm. In particular, suppose we want to multiply two matrices of size  $n \times m$  and  $m \times n$ . Doing so naïvely takes time  $O(n^2m)$ . Starting with the work of Strassen [1969], there has been substantial work to improve this run-time; see also V. V. Williams [2012]. Below we state the running time of a fast matrix multiplication algorithm due to Coppersmith [1982], which is most relevant for this section.

**Theorem 5.3 (Coppersmith [ibid.]).** *Fix  $n \geq 1$  and let  $m \geq 1$  be such that  $m \leq n^{0.172}$ . One can compute the product of two matrices of sizes  $n \times m$  and  $m \times n$  in  $O(n^2 \log^2 n)$  time.*

**5.1 Closest pair via matrix multiplication.** We now sketch the algorithm for the closest pair from Valiant [2015], which obtains  $O(n^{2-\Omega(\sqrt{\varepsilon})}d)$  time. The algorithm is best described in terms of *inner products*, as opposed to distances as before. In particular, suppose we have a set of points  $P \subset \mathbb{S}^d$  of unit norm, where all pairs of points have inner product in the range  $[-\theta, \theta]$ , except for one “special” pair that has inner product at least  $c\theta$ , for some scale  $\theta > 0$  and approximation  $c = 1 + \varepsilon$ . Now the problem is to find this special pair—we term this problem  $(c, \theta)$ -IP problem. We note that we can reduce  $(1 + \varepsilon, r)$ -CP over  $\ell_2$  to  $(1 + \Omega(\varepsilon), 1/2)$ -IP, by using the embedding of Schoenberg [1942], or Lemma 2.3 of Kushilevitz, Ostrovsky, and Rabani [2000].

A natural approach to the IP problem is to multiply two  $n \times d$  matrices: if we consider the matrix  $M$  where the rows are the points of  $P$ , then  $MM^t$  will have a large off-diagonal entry precisely for the special pair of points. This approach however requires at least  $n^2$  computation time, since even the output of  $MM^t$  has size  $n^2$ . Nevertheless, an extension of this approach gives a better run-time when  $c$  is very large (and hence  $\theta < 1/c$  very small, i.e., all points except for the special pair are near-orthogonal). In particular, partition randomly the vectors from  $P$  into  $n/g$  groups  $S_1, \dots, S_{n/g}$ , each of size  $O(g)$ . For each group  $i$ , we sum the vectors  $S_i$  with random signs, obtaining vectors  $v_i = \sum_{p_j \in S_i} \chi_j p_j$ , where  $p_j$  are the points in  $P$  and  $\chi_j$  are Rademacher random variables. Now the algorithm forms a matrix  $M$  with  $v_i$ ’s as rows, and computes  $MM^t$  using fast matrix multiplication (Theorem 5.3). The two special points are separated with probability  $1 - g/n$ . Conditioning on this event, without loss of generality, we can assume that they are in group 1 and 2 respectively. Then, it is easy to note that  $|(MM^t)_{12}| \approx \Theta(c \cdot \theta)$ , whereas, for  $(i, j) \neq (1, 2)$  and  $i \neq j$ , we have that  $|(MM^t)_{ij}| \approx O(g \cdot \theta)$  with constant probability. Hence, we can identify the special pair in the product  $MM^t$  as long as  $c \gg g$ , and yields runtime  $O(n^2/g^2)$ , i.e., a  $g^2 \ll c^2$  speed-up over the naïve algorithm (note that Theorem 5.3 requires that  $d < n^{0.172}$ ).

The above approach requires  $c$  to be very large, and hence the challenge is whether we can reduce the case of  $c = 1 + \varepsilon$  to the case of large  $c$ . Indeed, one method is

to use tensoring: for a fixed parameter  $k$  and any two vectors  $x, y \in \mathbb{R}^d$ , we consider  $x^{\otimes k}, y^{\otimes k} \in \mathbb{R}^{d^k}$ , for which  $\langle x^{\otimes k}, y^{\otimes k} \rangle = (\langle x, y \rangle)^k$ . Thus tensoring reduces the problem of  $(1 + \varepsilon, 1/2)$ -IP to  $((1 + \varepsilon)^k, 2^{-k})$ -IP, and hence we hope to use the above algorithm for  $c = (1 + \varepsilon)^k \approx e^{\varepsilon k}$ . If we use  $t = \zeta \ln n$ , for small constant  $\zeta$ , we obtain  $c = n^{\varepsilon \zeta}$ , and hence we obtain a speed-up of  $g^2 \approx c^2 = n^{2\varepsilon \zeta}$ . One caveat here is that, after tensoring the vectors, we obtain vectors of dimension  $d^k$ , which could be much larger than  $n$ —then even writing down such vectors would take  $\Omega(n^2)$  time. Yet, one can use a dimension reduction method, like [Lemma 2.1](#), to reduce dimension to  $O(\frac{\log n}{\theta^k}) = \tilde{O}(n^{\xi \ln 2})$ , which is enough to preserve all inner products up to additive, say,  $0.1 \cdot \theta^k$ . There are further details (e.g., we cannot afford to get high-dimensional vectors in the first place, even if we perform dimension-reduction), see [Valiant \[2015\]](#) and [Karppa, Kaski, and Kohonen \[2016\]](#) for more details.

The above algorithm yields a speed-up of the order of  $n^{O(\varepsilon)}$ , i.e., comparable to the speed-up via the LSH methods. To obtain a better speed-up, like in the [Theorem 5.2](#), one can replace the tensoring transformation with a more efficient one. Indeed, one can employ an *asymmetric* embedding  $f, g : \mathbb{R}^d \rightarrow \mathbb{R}^m$ , with the property that for any unit-norm vectors  $x, y$ , we have that  $\langle f(x), g(y) \rangle = p(\langle x, y \rangle)$ , where  $p(\cdot)$  is a polynomial of choice. In particular, we require a polynomial  $p(\cdot)$  that is small on the interval  $[-\theta, \theta]$ , as large as possible on  $[(1 + \varepsilon)\theta, 1]$ , and  $p(1)$  is bounded. Note that the tensoring operation implements such an embedding with  $p(a) = a^k$  and where  $f(x) = g(x) = x^{\otimes k}$ . However, there are more efficient polynomials: in fact, the optimal such polynomial is the Chebyshev polynomial. For example, for the degree- $k$  Chebyshev polynomial  $T_k(\cdot)$ , we have that  $T_k(1 + \varepsilon)/T_k(1) \approx e^{\sqrt{\varepsilon}k}$ , which is in contrast to the above polynomial  $p(a) = a^k$ , for which  $p(1 + \varepsilon)/p(1) \approx e^{\varepsilon k}$ .

Using the Chebyshev polynomials, one can obtain a runtime of  $n^{2-\Omega(\sqrt{\varepsilon})}$  for the IP and hence CP problem. To obtain the improved result from [Theorem 5.2](#), [Alman, Chan, and R. Williams \[2016\]](#) employ *randomized* polynomials, i.e., a distribution over polynomials where  $p(\cdot)$  is small/large only with a certain probability. Without going into further details, the theorem below states the existence of such polynomials, which are used to obtain  $n^{2-\Omega(\varepsilon^{1/3}/\log(1/\varepsilon))}$  run-time for the  $(1 + \varepsilon, r)$ -CP problem.

**Theorem 5.4** ([Alman, Chan, and R. Williams \[ibid.\]](#)). *Fix  $d \geq 1, \theta \geq 1, s \geq 1$ , and  $\varepsilon > 0$ . There exists a distribution over polynomials  $P : \{0, 1\}^d \rightarrow \mathbb{R}$  of degree  $O(\varepsilon^{-1/3} \log s)$ , such that we have the following for any  $x \in \{0, 1\}^d$ :*

- if  $\sum_{i=1}^d x_i \leq \theta$ , then  $|P(x)| \leq 1$  with probability at least  $1 - 1/s$ ;
- if  $\sum_{i=1}^d x_i \in (\theta, (1 + \varepsilon)\theta)$ , then  $|P(x)| > 1$  with probability at least  $1 - 1/s$ ;
- if  $\sum_{i=1}^d x_i > (1 + \varepsilon)\theta$ , then  $|P(x)| \geq s$  with probability at least  $1 - 1/s$ .

## 6 Extensions

In this section, we discuss several techniques that significantly extend the class of spaces which admit efficient ANN data structures.

**6.1 Metric embeddings.** So far, we have studied the ANN problem over the  $\ell_1$ ,  $\ell_2$  and  $\ell_\infty$  distances. A useful approach is to *embed* a metric of interest into  $\ell_1/\ell_2/\ell_\infty$  and use one of the data structures developed for the latter spaces.

### 6.1.1 Deterministic embeddings.

**Definition 6.1.** For metric spaces  $\mathfrak{M} = (X, D_X)$ ,  $\mathfrak{N} = (Y, D_Y)$  and for  $D \geq 1$ , we say that a map  $f: X \rightarrow Y$  is a bi-Lipschitz embedding with distortion  $D$  if there exists  $\lambda > 0$  such that for every  $x_1, x_2 \in X$  one has:

$$\lambda D_X(x_1, x_2) \leq D_Y(f(x_1), f(x_2)) \leq D \cdot \lambda D_X(x_1, x_2).$$

A bi-Lipschitz embedding of  $\mathfrak{M}$  into  $\mathfrak{N}$  with distortion  $D$  together with a data structure for  $(c, r)$ -ANN over  $\mathfrak{N}$  immediately implies a data structure for  $(cD, r')$ -ANN over  $\mathfrak{M}$ , where  $r' = \frac{r}{\lambda D}$ . However, space and query time of the resulting data structure depend crucially on the *computational efficiency* of the embedding, since, in particular, the query procedure requires evaluating the embedding on a query point.

As the following classic results show, any finite-dimensional normed or finite metric space can be embedded into finite-dimensional  $\ell_\infty$  with small distortion.

**Theorem 6.2** (Fréchet [1906] and Kuratowski [1935]). *If  $\mathfrak{M}$  is a finite metric space, which consists of  $N$  points, then  $\mathfrak{M}$  embeds into  $(\mathbb{R}^N, \ell_\infty)$  with distortion  $D = 1$  (isometrically).*

**Theorem 6.3** (see, e.g., Wojtaszczyk [1996]). *For every  $\varepsilon > 0$ , every normed space  $(\mathbb{R}^d, \|\cdot\|)$  embeds with distortion  $1 + \varepsilon$  into  $(\mathbb{R}^{d'}, \ell_\infty)$ , where  $d' = O(1/\varepsilon)^d$ , via a linear map.*

However, the utility of Theorems 6.2 and 6.3 in the context of the ANN problem is limited, since the required dimension of the target  $\ell_\infty$  space is very high (in particular, Theorem 6.3 gives a data structure with *exponential* dependence on the dimension). Moreover, even if we allow the distortion  $D$  of an embedding to be a large constant, the target dimension can not be improved much. As has been shown in Matoušek [1997], one needs at least  $N^{\Omega(1/D)}$ -dimensional  $\ell_\infty$  to “host” all the  $N$ -point metrics with distortion  $D$ . For  $d$ -dimensional norms, even as simple as  $\ell_2$ , the required dimension is  $2^{\Omega_D(d)}$  Figiel, Lindenstrauss, and Milman [1977] and Ball [1997].

More generally, (lower-dimensional)  $\ell_\infty$  turns out to be not so useful of a target space, and only a handful of efficient embeddings into  $\ell_\infty$  are known (for instance, such an embedding has been constructed in [Farach-Colton and Indyk \[1999\]](#) for the Hausdorff distance). Luckily, the situation drastically improves, if we allow *randomized* embeddings, see [Section 6.1.2](#) for the examples.

Instead of  $\ell_\infty$ , one can try to embed a metric of interest into  $\ell_1$  or  $\ell_2$ . Let us list a few cases, where such embeddings lead to efficient ANN data structures.

- Using the result from [Johnson and Schechtman \[1982\]](#), one can embed  $(\mathbb{R}^d, \ell_p)$  for  $1 < p \leq 2$  into  $(\mathbb{R}^{d'}, \ell_1)$  with distortion  $1 + \varepsilon$ , where  $d' = O(d/\varepsilon^2)$ . Moreover, the corresponding map is linear and hence efficient to store and apply. This reduction shows that the ANN problem over  $\ell_p$  for  $1 < p \leq 2$  is no harder than for the  $\ell_1$  case. However, later in this section we will show how to get a better ANN algorithm for the  $\ell_p$  case using a different embedding.
- For the Wasserstein-1 distance (a.k.a. the Earth-Mover distance in the computer science literature) between probability measures defined on  $\{1, 2, \dots, d\}^k$ , one can use the results from [M. S. Charikar \[2002\]](#), [Indyk and Thaper \[2003\]](#), and [A. Naor and Schechtman \[2007\]](#), to embed it into  $(\mathbb{R}^{d^{O(k)}}, \ell_1)$  with distortion  $O(k \log d)$ .
- The Levenshtein distance (a.k.a. edit distance) over the binary strings  $\{0, 1\}^d$  can be embedded into  $(\mathbb{R}^{d^{O(1)}}, \ell_1)$  with distortion  $2^{O(\sqrt{\log d \log \log d})}$  [Ostrovsky and Rabani \[2007\]](#).

Let us note that there exist generic results concerned with embeddings into  $\ell_1/\ell_2$  similar to [Theorem 6.2](#) and [Theorem 6.3](#).

**Theorem 6.4** ([Bourgain \[1985\]](#) and [Linial, London, and Rabinovich \[1995\]](#)). *Any  $N$ -point metric embeds into  $(\mathbb{R}^{O(\log N)}, \ell_2)$  with distortion  $O(\log N)$ .*

**Theorem 6.5** ([John \[1948\]](#) and [Ball \[1997\]](#)). *Any normed space  $(\mathbb{R}^d, \|\cdot\|)$  embeds into  $(\mathbb{R}^d, \ell_2)$  with distortion  $\sqrt{d}$  via a linear map.*

[Theorem 6.4](#) does not give an embedding efficient enough for the ANN applications: computing it in one point requires time  $\Omega(N)$ . At the same time, [Theorem 6.5](#) is efficient and, together with  $\ell_2$  data structures, gives an ANN data structure for a general  $d$ -dimensional norm with approximation  $O(\sqrt{d})$ .

Since the ANN problem is defined for two specific distance scales ( $r$  and  $cr$ ), we do not need the full power of bi-Lipschitz embeddings and sometimes can get away with weaker notions of embeddability. For example, the following theorem follows from the results of [Schoenberg \[1937\]](#).

In the theorem,  $\ell_2(\mathbb{N})$  denotes the space of infinite sequences  $(a_i)_{i=1}^\infty$  such that  $\sum_i |a_i|^2 < +\infty$  and the norm of the sequence  $\|a\|_2$  is equal to  $(\sum_i |a_i|^2)^{1/2}$ .

**Theorem 6.6.** *For every  $1 \leq p < 2$  and every  $d \geq 1$ , there exists a map  $f: \mathbb{R}^d \rightarrow \ell_2(\mathbb{N})$  such that for every  $x, y \in \mathbb{R}^d$ , one has:*

$$\|f(x) - f(y)\|_2^2 = \|x - y\|_p^p.$$

This embedding allows to use an ANN data structure for  $\ell_2$  with approximation  $c$  to get an ANN data structure for  $\ell_p$  with approximation  $c^{2/p}$ . However, for this we need to make the embedding computationally efficient. In particular, the target must be finite-dimensional. This can be done, see [H. Nguyễn \[2014\]](#) for details. As a result, for the  $\ell_p$  distance for  $1 \leq p < 2$ , we are able to get the result similar to the one given by [Theorem 4.2](#), where in (2)  $c^2$  is replaced with  $c^p$  everywhere.

**6.1.2 Randomized embeddings.** It would be highly desirable to utilize the fact that every metric embeds well into  $\ell_\infty$  (Theorems 6.2 and 6.3) together with the ANN data structure for  $\ell_\infty$  from [Section 4.7](#). However, as discussed above, spaces as simple as  $(\mathbb{R}^d, \ell_1)$  or  $(\mathbb{R}^d, \ell_2)$  require the target  $\ell_\infty$  to have  $2^{\Omega(d)}$  dimensions to be embedded with small distortion. It turns out, this can be remedied by allowing embeddings to be *randomized*. In what follows, we will consider the case of  $(\mathbb{R}^d, \ell_1)$ , and then generalize the construction to other metrics.

The randomized embedding of  $(\mathbb{R}^d, \ell_1)$  into  $(\mathbb{R}^d, \ell_\infty)$  is defined as follows: we generate  $d$  i.i.d. samples  $u_1, u_2, \dots, u_d$  from the exponential distribution with parameter 1, and then the embedding  $f$  maps a vector  $x \in \mathbb{R}^d$  into

$$\left( \frac{x_1}{u_1}, \frac{x_2}{u_2}, \dots, \frac{x_d}{u_d} \right).$$

Thus, the resulting embedding is linear. Besides that, it is extremely efficient to store ( $d$  numbers) and apply ( $O(d)$  time).

Let us now understand how  $\|f(x)\|_\infty$  is related to  $\|x\|_1$ . The analysis uses (implicitly) the *min-stability* property of the exponential distribution. One has for every  $t > 0$ :

$$\Pr_f[\|f(x)\|_\infty \leq t] = \prod_{i=1}^d \Pr\left[\frac{|x_i|}{u_i} \leq t\right] = \prod_{i=1}^d \Pr\left[u_i \geq \frac{|x_i|}{t}\right] = \prod_{i=1}^d e^{-|x_i|/t} = e^{-\|x\|_1/t}.$$

The random variable  $\|f(x)\|_\infty$  does not have a finite first moment, however its mode is in the point  $t = \|x\|_1$ , which allows us to use  $\|f(x)\|_\infty$  to estimate  $\|x\|_1$ . It is immediate

to show that for every  $\delta > 0$ , there exist  $C_1, C_2 > 1$  with  $C_1 = O(\log(1/\delta))$  and  $C_2 = O(1/\delta)$  such that for every  $x$ , one has:

$$(4) \quad \Pr_f \left[ \|f(x)\|_\infty \geq \frac{\|x\|_1}{C_1} \right] \geq 1 - \delta$$

and

$$(5) \quad \Pr_f [\|f(x)\|_\infty \leq C_2 \cdot \|x\|_1] \geq 1 - \delta$$

Thus, the map  $f$  has distortion  $O\left(\frac{\log(1/\delta)}{\delta}\right)$  with probability  $1 - \delta$ . However, unlike the deterministic case, the randomized guarantees (4) and (5) are not sufficient for the reduction between ANN data structures (if  $\delta \gg 1/n$ ). This is because the lower bound on  $\|f(x)\|_\infty$  must apply simultaneously to all “far” points. In order to obtain a desired reduction, we need to use slightly different parameters. Specifically, for  $0 < \varepsilon < 1$  one has:

$$\Pr_f \left[ \|f(x)\|_\infty \geq \Omega\left(\frac{\|x\|_1}{\log n}\right) \right] \geq 1 - \frac{1}{10n}$$

and

$$\Pr_f \left[ \|f(x)\|_\infty \leq O\left(\frac{\|x\|_1}{\varepsilon \cdot \log n}\right) \right] \geq n^{-\varepsilon}.$$

This allows us to reduce the  $(c/\varepsilon, r)$ -ANN problem over  $(\mathbb{R}^d, \ell_1)$  to  $n^{O(\varepsilon)}$  instances of the  $(c, r')$ -ANN problem over  $(\mathbb{R}^d, \ell_\infty)$ . Indeed, we sample  $n^{O(\varepsilon)}$  i.i.d. maps  $f_i$  as described above and solve the ANN problem over  $\ell_\infty$  on the image of  $f_i$ . Far points remain being far with probability  $1 - 1/10n$  each. Using the linearity of expectation and the Markov inequality, we observe that, with probability at least 0.9, *no* far point come close enough to the query point. At the same time, with probability at least  $n^{-\varepsilon}$ , the near neighbor does not move too far away, so, with high probability, at least one of the  $n^{O(\varepsilon)}$  data structures succeeds. This reduction is quite similar to the use of Locality-Sensitive Hashing in [Section 2.2](#).

As a result, we get an ANN data structure for  $(\mathbb{R}^d, \ell_1)$  with approximation  $O\left(\frac{\log \log d}{\varepsilon^2}\right)$ , query time  $O(dn^\varepsilon)$  and space  $O(dn^{1+\varepsilon})$ . This is worse than the best ANN data structure for  $\ell_1$  based on (data-dependent) space partitions. However, the technique we used is very versatile and generalizes easily to many other distances. The  $\ell_1$  embedding was first used in [Andoni, Indyk, and Krauthgamer \[2009\]](#). Later, it was generalized [Andoni \[2009\]](#) to  $\ell_p$  spaces for  $p \geq 1$ . To get such an embedding, one can divide every coordinate by the  $(1/p)$ -th power of an exponential random variable. Finally, in [Andoni, H. L. Nguyen, Nikolov, Razenshteyn, and Waingarten \[2017\]](#) the same technique has been shown to work for Orlicz norms and top- $k$  norms, which we define next.

**Definition 6.7.** Let  $\psi : [0; +\infty) \rightarrow [0; +\infty)$  be a non-negative monotone increasing convex function with  $\psi(0) = 0$ . Then, an Orlicz norm  $\|\cdot\|_\psi$  over  $\mathbb{R}^d$  is given by its unit ball  $K_\psi$ , defined as follows:

$$K_\psi = \left\{ x \in \mathbb{R}^d \mid \sum_{i=1}^d \psi(|x_i|) \leq 1 \right\}.$$

Clearly,  $\ell_p$  norm for  $p < \infty$  is Orlicz for  $\psi(t) = t^p$ .

**Definition 6.8.** For  $1 \leq k \leq d$ , we define the top- $k$  norm of a vector from  $\mathbb{R}^d$  as the sum of  $k$  largest absolute values of the coordinates.

The top-1 norm is simply  $\ell_\infty$ , while top- $d$  corresponds to  $\ell_1$ .

To embed an Orlicz norm  $\|\cdot\|_\psi$  into  $\ell_\infty$ , we divide the coordinates using a random variable  $X$  with the c.d.f.  $F_X(t) = \Pr[X \leq t] = 1 - e^{-\psi(t)}$ . To embed the top- $k$  norm, we use a truncated exponential distribution. All of the above embeddings introduce only a constant distortion.

Let us note that for the  $\ell_p$  norms one can achieve approximation  $2^{O(p)}$  [A. Naor and Rabani \[2006\]](#) and [Bartal and Gottlieb \[2015\]](#), which is an improvement upon the above  $O(\log \log d)$  bound if  $p$  is sufficiently small.

**6.2 ANN for direct sums.** In this section we describe a vast generalization of the ANN data structure for  $\ell_\infty$  from [Section 4.7](#). Namely, we will be able to handle *direct sums* of metric spaces.

**Definition 6.9.** Let  $M_1 = (X_1, D_1)$ ,  $M_2 = (X_2, D_2)$ , ...,  $M_k = (X_k, D_k)$  be metric spaces and let  $\|\cdot\|$  be a norm over  $\mathbb{R}^k$ . Then the  $\|\cdot\|$ -direct sum of  $M_1, M_2, \dots, M_k$  denoted by  $\left(\bigoplus_{i=1}^k M_i\right)_{\|\cdot\|}$  is a metric space defined as follows. The ground set is the Cartesian product  $X_1 \times X_2 \times \dots \times X_k$ . The distance function  $D$  is given by the following formula.

$$D((x_1, x_2, \dots, x_k), (y_1, y_2, \dots, y_k)) = \| (D_1(x_1, y_1), D_2(x_2, y_2), \dots, D_k(x_k, y_k)) \|.$$

It turns out that for many interesting norms  $\|\cdot\|$  the following holds. If for metrics  $M_1, M_2, \dots, M_k$  there exist efficient ANN data structures, then the same holds for  $\left(\bigoplus_{i=1}^k M_i\right)_{\|\cdot\|}$  (with a mild loss in the parameters).

The first result of this kind was shown in [Indyk \[2002\]<sup>11</sup>](#) for the case of  $\ell_\infty$ -direct sums. In what follows we denote by  $d$  the “complexity” of each metric  $M_i$ . That is, we assume

<sup>11</sup>In [Indyk \[2002\]](#), a slightly weaker version of [Theorem 6.10](#) has been stated. First, it assumed *deterministic* data structures for the spaces  $M_i$ . This is straightforward to address by boosting the probability of success



it takes  $O(d)$  time to compute the distance between two points, and that a point requires  $O(d)$  space to store.

**Theorem 6.10.** *Let  $c > 1$ ,  $r > 0$  and  $0 < \varepsilon < 1$ . Suppose that each  $M_i$  admits a  $(c, r)$ -ANN data structure for  $n$ -point sets with space  $n^{1+\rho}$  (in addition to storing the dataset) for some  $\rho \geq 0$  and query time  $Q(n)$ . Then, there exists a data structure for  $(c', r)$ -ANN over  $\left(\bigoplus_{i=1}^k M_i\right)_\infty$ , where  $c' = O\left(\frac{c \log \log n}{\varepsilon}\right)$ , the space is  $O(n^{1+\rho+\varepsilon})$  (in addition to storing the dataset), and the query time is  $Q(n) \cdot \log^{O(1)} n + O(dk \log n)$ .*

Informally speaking, compared to data structures for  $M_i$ , the data structure for  $\left(\bigoplus_i M_i\right)_\infty$  loses  $\frac{\log \log n}{\varepsilon}$  in approximation,  $n^\varepsilon$  in space, and  $\log^{O(1)} n$  in query time.

Later, the result of Indyk [2002] was significantly extended Indyk [2004], Andoni, Indyk, and Krauthgamer [2009], Andoni [2009], and Andoni, H. L. Nguyễn, Nikolov, Razenshteyn, and Waingarten [2017], to support  $\|\cdot\|$ -direct sums where  $\|\cdot\|$  is an  $\ell_p$  norm, an Orlicz norm, or a top- $k$  norm. The main insight is that we can use the randomized embeddings of various norms into  $\ell_\infty$  developed in Section 6.1.2, to reduce the case of  $\|\cdot\|$ -direct sums to the case of  $\ell_\infty$ -direct sums. Indeed, we described how to reduce the ANN problem over several classes of norms to  $n^\varepsilon$  instances of ANN over the  $\ell_\infty$  distance at a cost of losing  $O(1/\varepsilon)$  in the approximation. It is not hard to see that the exact same approach can be used to reduce the ANN problem over  $\left(\bigoplus_{i=1}^k M_i\right)_{\|\cdot\|}$  to  $n^\varepsilon$  instances of ANN over  $\left(\bigoplus_{i=1}^k M_i\right)_\infty$  also at a cost of losing  $O(1/\varepsilon)$  in approximation.

**6.3 Embeddings into direct sums.** As Section 6.2 shows, for a large class of norms  $\|\cdot\|$ , we can get an efficient ANN data structure for any  $\|\cdot\|$ -direct sum of metrics that admit efficient ANN data structures. This gives a natural approach to the ANN problem: embed a metric of interest into such a direct sum.

This approach has been successful in several settings. In Indyk [2002], the Fréchet distance between two sequences of points in a metric space is embedded into an  $\ell_\infty$ -direct

for data structures for  $M_i$  using repetition. Second, the resulting space bound Indyk [2002] was worse. An improvement to the space bound has been described in Appendix A of the arXiv version of Andoni, H. L. Nguyễn, Nikolov, Razenshteyn, and Waingarten [2017]. Finally, the paper Indyk [2002] assumes ANN data structures for  $M_i$  with a slightly stronger guarantee. Namely, each point is assigned a *priority* from 1 to  $n$ , and if the near neighbor has priority  $t$ , we must return a point with priority at most  $t$ . It is not hard to solve the version with priorities using a standard ANN data structure (with  $\log^{O(1)} n$  loss in space and query time). A naïve reduction builds an ANN data structure for points with priority at most  $t$  for every  $t$ . Then, we can run a binary search over the resulting priority. However, this gives a *linear* in  $n$  loss in space. To rectify this, we use a standard data structure technique: the decomposition of an interval into  $O(\log n)$  *dyadic intervals*, i.e., intervals of the form  $[2^k \cdot l + 1; 2^k \cdot (l + 1)]$  for integer  $k, l$ . Thus, we build an ANN data structure for every dyadic interval of priorities. This still gives  $O(n)$  ANN data structures, however, each data point participates in at most  $O(\log n)$  of them.

sums of Fréchet distances between shorter sequences. Together with [Theorem 6.10](#), this was used to obtain an ANN data structure for the Fréchet distance. In [Andoni, Indyk, and Krauthgamer \[2009\]](#), it is shown how to embed the Ulam metric (which is the edit distance between permutations of length  $d$ ) into  $\left(\bigoplus^d \left(\bigoplus^{O(\log d)} (\mathbb{R}^d, \ell_1)\right)\right)_{\ell_\infty}$  with a constant distortion which gives an ANN data structure with doubly-logarithmic approximation. At the same time, the Ulam distance requires distortion  $\Omega\left(\frac{\log d}{\log \log d}\right)$  to embed into  $\ell_1$  [Andoni and Krauthgamer \[2010\]](#). This shows that (lower-dimensional) direct sums form a strictly more “powerful” class of spaces than  $\ell_1$  or  $\ell_2$ . Finally, in [Andoni, H. L. Nguyen, Nikolov, Razenshteyn, and Waingarten \[2017\]](#), it is shown that *any* symmetric norm over  $\mathbb{R}^d$  is embeddable into  $\left(\bigoplus_{i=1}^{d^{O(1)}} \left(\bigoplus_{j=1}^d X_{ij}\right)_1\right)_\infty$  with constant distortion, where  $X_{ij}$  is  $\mathbb{R}^d$  equipped with the top- $j$  norm. Together with the results from [Section 6.1.2](#) and [Section 6.2](#), this gives an ANN algorithm with approximation  $(\log \log n)^{O(1)}$  for general *symmetric*<sup>12</sup> norms.

**6.4 ANN for general norms.** For *general*  $d$ -dimensional norms, the best known ANN data structure is obtained by combining [Theorem 6.5](#) with an efficient ANN data structure for  $\ell_2$  (for example, the one given by [Theorem 4.1](#)). This approach gives approximation  $O(\sqrt{d}/\varepsilon)$  for space  $d^{O(1)} \cdot n^{1+\varepsilon}$  and query time  $d^{O(1)} \cdot n^\varepsilon$  for every constant  $0 < \varepsilon < 1$ . Very recently, the approximation  $O(\sqrt{d}/\varepsilon)$  has been improved to  $O\left(\frac{\log d}{\varepsilon^2}\right)$  [Andoni, A. Naor, Nikolov, Razenshteyn, and Waingarten \[2017\]](#) for the same space and time bounds if one is willing to relax the model of computation to the *cell-probe model*, where the query procedure is charged for *memory accesses*, but any computation is free. This ANN data structure heavily builds on a recent geometric result from [A. Naor \[2017\]](#): a bi-Lipschitz embedding (see [Definition 6.1](#)) of the shortest-path metric of *any*  $N$ -node expander graph [Hoory, Linial, and Wigderson \[2006\]](#) into an *arbitrary*  $d$ -dimensional normed space must have distortion at least  $\Omega(\log_d N)$ . At a very high level, this non-embeddability result is used to claim that any large bounded-degree graph, which *does* embed into a normed space, can not be an expander, and hence it must have a sparse cut. The existence of the sparse cut is then used, via a duality argument, to build a (data-dependent) random space partition family for a general  $d$ -dimensional normed space. The latter family is used to obtain the final data structure.

This approach can be further extended for several norms of interest to obtain proper, *time-efficient* ANN data structures, with even better approximations. For instance, [Andoni, A. Naor, Nikolov, Razenshteyn, and Waingarten \[2017\]](#) show how to get ANN with approximation  $O(p)$  for the  $\ell_p$  norms, improving upon the bound  $2^{O(p)}$  from [A. Naor](#)

<sup>12</sup>Under permutations and negations of the coordinates.

and Rabani [2006] and Bartal and Gottlieb [2015]. Finally, for the Schatten- $p$  norms of matrices, defined as the  $\ell_p$  norm of the vector of singular values, one obtains approximation  $O(p)$  as well, while the previous best approximation was polynomial in the matrix size (by relating the Schatten- $p$  norm to the Frobenius norm).

**Acknowledgments.** The authors would like to thank Assaf Naor, Tal Wagner, Erik Waingarten and Fan Wei for many helpful comments.

## References

- Thomas D Ahle, Rasmus Pagh, Ilya Razenshteyn, and Francesco Silvestri (2016). “On the complexity of maximum inner product search”. In: *Proc. 8th 35th ACM Symposium on Principles of Database Systems (PODS)* (cit. on p. 3306).
- Thomas Dybdahl Ahle (2017). “Optimal Las Vegas Locality Sensitive Data Structures”. In: *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*. IEEE (cit. on p. 3316).
- Nir Ailon and Bernard Chazelle (2009). “The Fast Johnson–Lindenstrauss Transform and Approximate Nearest Neighbors”. *SIAM J. Comput.* 39.1, pp. 302–322 (cit. on p. 3310).
- Nir Ailon and Edo Liberty (2013). “An Almost Optimal Unrestricted Fast Johnson–Lindenstrauss Transform”. *ACM Transactions on Algorithms* 9.3, p. 21 (cit. on p. 3310).
- Josh Alman, Timothy M Chan, and Ryan Williams (2016). “Polynomial representations of threshold functions and algorithmic applications”. In: *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, pp. 467–476 (cit. on pp. 3322, 3324).
- Alexandr Andoni (2009). “Nearest neighbor search: the old, the new, and the impossible”. PhD thesis. Massachusetts Institute of Technology (cit. on pp. 3328, 3330).
- Alexandr Andoni, Dorian Croitoru, and Mihai Patrascu (2008). “Hardness of nearest neighbor under L-infinity”. In: *Proceedings of the 49th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, pp. 424–433 (cit. on p. 3321).
- Alexandr Andoni and Piotr Indyk (2006). “Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions”. In: *Proceedings of the 47th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, pp. 459–468 (cit. on pp. 3313, 3314).
- (2008). “Near-Optimal Hashing Algorithms for Approximate Nearest Neighbor in High Dimensions”. *Communications of the ACM* 51.1, p. 117 (cit. on pp. 3305, 3313).
- Alexandr Andoni, Piotr Indyk, and Robert Krauthgamer (2009). “Overcoming the  $\ell_1$  non-embeddability barrier: algorithms for product metrics”. In: *Proceedings of the twentieth*

- Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, pp. 865–874 (cit. on pp. [3328](#), [3330](#), [3331](#)).
- Alexandr Andoni, Piotr Indyk, Huy L Nguyễn, and Ilya Razenshteyn (2014). “Beyond locality-sensitive hashing”. In: *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, pp. 1018–1028 (cit. on pp. [3316](#), [3317](#)).
- Alexandr Andoni and Robert Krauthgamer (2010). “The computational hardness of estimating edit distance”. *SIAM Journal on Computing* 39.6, pp. 2398–2429 (cit. on p. [3331](#)).
- Alexandr Andoni, Thijs Laarhoven, Ilya Razenshteyn, and Erik Waingarten (2017). “Optimal hashing-based time-space trade-offs for approximate near neighbors”. In: *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, pp. 47–66 (cit. on pp. [3316](#), [3319](#)).
- Alexandr Andoni, Assaf Naor, Aleksandar Nikolov, Ilya Razenshteyn, and Erik Waingarten (2017). “Data-Dependent Hashing via Nonlinear Spectral Gaps”. *Manuscript* (cit. on p. [3331](#)).
- Alexandr Andoni, Huy L Nguyễn, Aleksandar Nikolov, Ilya Razenshteyn, and Erik Waingarten (2017). “Approximate near neighbors for general symmetric norms”. In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, pp. 902–913 (cit. on pp. [3328](#), [3330](#), [3331](#)).
- Alexandr Andoni and Ilya Razenshteyn (2015). “Optimal data-dependent hashing for approximate near neighbors”. In: *Proceedings of the forty-seventh annual ACM symposium on Theory of computing*. ACM, pp. 793–801 (cit. on pp. [3316](#), [3317](#)).
- (2016). “Tight Lower Bounds for Data-Dependent Locality-Sensitive Hashing”. In: *Proceedings of the 32nd International Symposium on Computational Geometry, SoCG 2016, June 14-18, 2016, Boston, MA, USA*, 9:1–9:11 (cit. on p. [3320](#)).
- Sunil Arya and David M Mount (1993). “Approximate Nearest Neighbor Queries in Fixed Dimensions.” In: *Proceedings of the Fourth Annual ACM-SIAM Symposium on Discrete Algorithms*. Vol. 93, pp. 271–280 (cit. on pp. [3306](#), [3307](#)).
- Sunil Arya, David M Mount, Nathan S Netanyahu, Ruth Silverman, and Angela Y Wu (1998). “An optimal algorithm for approximate nearest neighbor searching fixed dimensions”. *Journal of the ACM (JACM)* 45.6, pp. 891–923 (cit. on p. [3307](#)).
- Keith Ball (1997). “An elementary introduction to modern convex geometry”. *Flavors of geometry* 31, pp. 1–58 (cit. on pp. [3325](#), [3326](#)).
- Yair Bartal and Lee-Ad Gottlieb (2015). “Approximate Nearest Neighbor Search for  $\ell_p$ -Spaces ( $2 < p < \infty$ ) via Embeddings”. arXiv: [1512.01775](#) (cit. on pp. [3329](#), [3331](#), [3332](#)).
- Marshall Bern (1993). “Approximate closest-point queries in high dimensions”. *Information Processing Letters* 45.2, pp. 95–99 (cit. on pp. [3306](#), [3307](#)).

- Jean Bourgain (1985). “On Lipschitz embedding of finite metric spaces in Hilbert space”. *Israel Journal of Mathematics* 52.1, pp. 46–52 (cit. on p. 3326).
- Bo Brinkman and Moses Charikar (2005). “On the impossibility of dimension reduction in  $\ell_1$ ”. *Journal of the ACM (JACM)* 52.5, pp. 766–788 (cit. on p. 3310).
- Andrei Z Broder (1997). “On the resemblance and containment of documents”. In: *Compression and Complexity of Sequences 1997. Proceedings*. IEEE, pp. 21–29 (cit. on p. 3313).
- Andrei Z Broder, Steven C Glassman, Mark S Manasse, and Geoffrey Zweig (1997). “Syntactic clustering of the web”. *Computer Networks and ISDN Systems* 29.8-13, pp. 1157–1166 (cit. on p. 3313).
- Constantin Carathéodory (1911). “Über den Variabilitätsbereich der FourierÖschen Konstanten von positiven harmonischen Funktionen”. *Rendiconti Del Circolo Matematico di Palermo (1884-1940)* 32.1, pp. 193–217 (cit. on p. 3306).
- Timothy M Chan (1998). “Approximate nearest neighbor queries revisited”. *Discrete & Computational Geometry* 20.3, pp. 359–373 (cit. on p. 3307).
- Moses S Charikar (2002). “Similarity estimation techniques from rounding algorithms”. In: *Proceedings of the thirty-fourth annual ACM symposium on Theory of computing*. ACM, pp. 380–388 (cit. on pp. 3313, 3326).
- Kenneth L Clarkson (1988). “A randomized algorithm for closest-point queries”. *SIAM Journal on Computing* 17.4, pp. 830–847 (cit. on p. 3306).
- (1994). “An algorithm for approximate closest-point queries”. In: *Proceedings of the tenth annual symposium on Computational geometry*. ACM, pp. 160–164 (cit. on p. 3307).
- (2006). “Nearest-neighbor searching and metric space dimensions”. *Nearest-neighbor methods for learning and vision: theory and practice*, pp. 15–59 (cit. on p. 3307).
- Don Coppersmith (1982). “Rapid multiplication of rectangular matrices”. *SIAM Journal on Computing* 11.3, pp. 467–471 (cit. on p. 3323).
- Thomas H. Cormen, Charles E. Leiserson, Ronald L. Rivest, and Clifford Stein (2001). *Introduction to Algorithms*. 2nd. MIT Press (cit. on p. 3309).
- Anirban Dasgupta, Ravi Kumar, and Tamás Sarlós (2010). “A sparse johnson: Lindenstrauss transform”. In: *Proceedings of the forty-second ACM symposium on Theory of computing*. ACM, pp. 341–350 (cit. on p. 3310).
- Mayur Datar, Nicole Immorlica, Piotr Indyk, and Vahab S Mirrokni (2004). “Locality-sensitive hashing scheme based on p-stable distributions”. In: *Proceedings of the twentieth annual symposium on Computational geometry*. ACM, pp. 253–262 (cit. on pp. 3312, 3313).
- Martin Farach-Colton and Piotr Indyk (1999). “Approximate nearest neighbor algorithms for Hausdorff metrics via embeddings”. In: *Proceedings of the 40th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, pp. 171–179 (cit. on p. 3326).

- Charles Fefferman and Bo'az Klartag (2009). "Fitting a  $C^m$ -Smooth Function to Data I". *Annals of Mathematics*, pp. 315–346 (cit. on p. 3308).
- Tadeusz Figiel, Joram Lindenstrauss, and Vitali D Milman (1977). "The dimension of almost spherical sections of convex bodies". *Acta Mathematica* 139.1, pp. 53–94 (cit. on p. 3325).
- M Maurice Fréchet (1906). "Sur quelques points du calcul fonctionnel". *Rendiconti del Circolo Matematico di Palermo (1884-1940)* 22.1, pp. 1–72 (cit. on p. 3325).
- Venkatesan Guruswami, Atri Rudra, and Madhu Sudan (2014). "Essential coding theory" (cit. on p. 3315).
- Venkatesan Guruswami, Christopher Umans, and Salil Vadhan (2009). "Unbalanced expanders and randomness extractors from Parvaresh–Vardy codes". *Journal of the ACM (JACM)* 56.4, p. 20 (cit. on pp. 3314, 3315).
- Sariel Har-Peled, Piotr Indyk, and Rajeev Motwani (2012). "Approximate Nearest Neighbor: Towards Removing the Curse of Dimensionality." *Theory of computing* 8.1, pp. 321–350 (cit. on pp. 3307, 3309).
- Shlomo Hoory, Nathan Linial, and Avi Wigderson (2006). "Expander graphs and their applications". *Bulletin of the American Mathematical Society* 43.4, pp. 439–561 (cit. on p. 3331).
- Piotr Indyk (2000a). "Dimensionality reduction techniques for proximity problems". In: *Proceedings of the eleventh annual ACM-SIAM symposium on Discrete algorithms*. SIAM, pp. 371–378 (cit. on pp. 3314, 3315).
- (2000b). "High-dimensional computational geometry". PhD thesis. stanford university (cit. on p. 3319).
  - (2001). "On approximate nearest neighbors under  $\ell_\infty$  norm". *Journal of Computer and System Sciences* 63.4, pp. 627–638 (cit. on p. 3321).
  - (2002). "Approximate nearest neighbor algorithms for Fréchet distance via product metrics". In: *Proceedings of the eighteenth annual symposium on Computational geometry*. ACM, pp. 102–106 (cit. on pp. 3329, 3330).
  - (2004). "Approximate nearest neighbor under edit distance via product metrics". In: *Proceedings of the fifteenth annual ACM-SIAM symposium on Discrete algorithms*. SIAM, pp. 646–650 (cit. on p. 3330).
- Piotr Indyk and Rajeev Motwani (1998). "Approximate nearest neighbors: towards removing the curse of dimensionality". In: *Proceedings of the thirtieth annual ACM symposium on Theory of computing*. ACM, pp. 604–613 (cit. on pp. 3309, 3311–3313, 3319).
- Piotr Indyk and Nitin Thaper (2003). "Fast image retrieval via embeddings". *Workshop on Statistical and Computational Theories of Vision* (cit. on p. 3326).
- Thathachar S Jayram and David P Woodruff (2013). "Optimal bounds for Johnson-Lindenstrauss transforms and streaming problems with subconstant error". *ACM Transactions on Algorithms (TALG)* 9.3, p. 26 (cit. on p. 3310).

- Fritz John (1948). “Extremum Problems with Inequalities as Subsidiary Conditions”. In: *Studies and Essays Presented to R. Courant on his 60th Birthday, January 8, 1948*. Interscience Publishers, Inc., New York, N. Y., pp. 187–204 (cit. on p. 3326).
- William B Johnson and Joram Lindenstrauss (1984). “Extensions of Lipschitz mappings into a Hilbert space”. *Contemporary mathematics* 26.189–206, p. 1 (cit. on p. 3309).
- William B Johnson and Assaf Naor (2009). “The Johnson-Lindenstrauss lemma almost characterizes Hilbert space, but not quite”. In: *Proceedings of the twentieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, pp. 885–891 (cit. on p. 3310).
- William B Johnson and Gideon Schechtman (1982). “Embedding  $\ell_p^m$  into  $\ell_1^n$ ”. *Acta Mathematica* 149.1, pp. 71–85 (cit. on p. 3326).
- Daniel M. Kane and Jelani Nelson (2014). “Sparsifier Johnson-Lindenstrauss Transforms”. *J. ACM* 61.1, p. 4 (cit. on p. 3310).
- Michael Kapralov (2015). “Smooth tradeoffs between insert and query complexity in nearest neighbor search”. In: *Proceedings of the 34th ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*. ACM, pp. 329–342 (cit. on p. 3319).
- Michael Kapralov and Rina Panigrahy (2012). “NNS lower bounds via metric expansion for  $l(\infty)$  and EMD”. In: *International Colloquium on Automata, Languages, and Programming*. Springer, pp. 545–556 (cit. on p. 3321).
- Matti Karppa, Petteri Kaski, and Jukka Kohonen (2016). “A faster subquadratic algorithm for finding outlier correlations”. In: *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, pp. 1288–1305 (cit. on pp. 3322, 3324).
- Jon M Kleinberg (1997). “Two algorithms for nearest-neighbor search in high dimensions”. In: *Proceedings of the twenty-ninth annual ACM symposium on Theory of computing*. ACM, pp. 599–608 (cit. on p. 3307).
- Felix Krahmer and Rachel Ward (2011). “New and Improved Johnson-Lindenstrauss Embeddings via the Restricted Isometry Property”. *SIAM J. Math. Analysis* 43.3, pp. 1269–1281 (cit. on p. 3310).
- Casimir Kuratowski (1935). “Quelques problèmes concernant les espaces métriques non-séparables”. *Fundamenta Mathematicae* 25.1, pp. 534–545 (cit. on p. 3325).
- Eyal Kushilevitz, Rafail Ostrovsky, and Yuval Rabani (2000). “Efficient search for approximate nearest neighbor in high dimensional spaces”. *SIAM Journal on Computing* 30.2, pp. 457–474 (cit. on pp. 3310, 3314, 3319, 3323).
- Nathan Linial, Eran London, and Yuri Rabinovich (1995). “The geometry of graphs and some of its algorithmic applications”. *Combinatorica* 15.2, pp. 215–245 (cit. on p. 3326).
- Richard J Lipton and Robert E Tarjan (1980). “Applications of a Planar Separator Theorem.” *SIAM Journal on Computing* 9.3, pp. 615–627 (cit. on p. 3306).



- Sepideh Mahabadi (2014). “Approximate nearest line search in high dimensions”. In: *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, pp. 337–354 (cit. on p. 3307).
- Jiří Matoušek (1997). “On embedding expanders into  $\ell_p$  spaces”. *Israel Journal of Mathematics* 102.1, pp. 189–197 (cit. on p. 3325).
- Stefan Meiser (1993). “Point location in arrangements of hyperplanes”. *Information and Computation* 106.2, pp. 286–303 (cit. on p. 3306).
- Peter Bro Miltersen (1999). “Cell probe complexity—a survey”. In: *Proceedings of the 19th Conference on the Foundations of Software Technology and Theoretical Computer Science, Advances in Data Structures Workshop*, p. 2 (cit. on p. 3308).
- Marvin Minsky and Seymour A Papert (1969). *Perceptrons: An introduction to computational geometry*. MIT press (cit. on p. 3306).
- Rajeev Motwani, Assaf Naor, and Rina Panigrahy (2007). “Lower bounds on locality sensitive hashing”. *SIAM Journal on Discrete Mathematics* 21.4, pp. 930–935 (cit. on p. 3313).
- Assaf Naor (2017). “A Spectral Gap Precludes Low-Dimensional Embeddings”. In: *Proceedings of the 33rd International Symposium on Computational Geometry, SoCG 2017* (cit. on p. 3331).
- Assaf Naor and Yuval Rabani (2006). “On Approximate Nearest Neighbor Search in  $\ell_p$ ,  $p > 2$ ”. Manuscript (cit. on pp. 3329, 3331).
- Assaf Naor and Gideon Schechtman (2007). “Planar Earthmover is not in  $L_1$ ”. *SIAM Journal on Computing* 37.3, pp. 804–826 (cit. on p. 3326).
- Moni Naor, Leonard J Schulman, and Aravind Srinivasan (1995). “Splitters and near-optimal derandomization”. In: *Proceedings of the 36th Annual IEEE Symposium on Foundations of Computer Science*. IEEE, pp. 182–191 (cit. on p. 3316).
- Jelani Nelson, Eric Price, and Mary Wootters (2014). “[New constructions of RIP matrices with fast multiplication and fewer rows](#)”. In: *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pp. 1515–1528 (cit. on p. 3310).
- Huy L. Nguyễn (2014). “[Algorithms for High Dimensional Data](#)”. Princeton University Thesis (cit. on p. 3327).
- Ryan O’Donnell, Yi Wu, and Yuan Zhou (2014). “Optimal lower bounds for locality-sensitive hashing (except when  $q$  is tiny)”. *ACM Transactions on Computation Theory (TOCT)* 6.1, p. 5 (cit. on p. 3313).
- Rafail Ostrovsky and Yuval Rabani (2007). “Low distortion embeddings for edit distance”. *Journal of the ACM (JACM)* 54.5, p. 23 (cit. on p. 3326).
- Rasmus Pagh (2016). “Locality-sensitive hashing without false negatives”. In: *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, pp. 1–9 (cit. on p. 3316).



- Rina Panigrahy (2006). “Entropy based nearest neighbor search in high dimensions”. In: *Proceedings of the seventeenth annual ACM-SIAM symposium on Discrete algorithm*. SIAM, pp. 1186–1195 (cit. on p. 3319).
- Franco P Preparata and Michael Ian Shamos (1985). “Introduction”. In: *Computational Geometry*. Springer, pp. 1–35 (cit. on p. 3306).
- Ilya Razenshteyn (2017). “High-Dimensional Similarity Search and Sketching: Algorithms and Hardness”. PhD thesis. Massachusetts Institute of Technology (cit. on p. 3317).
- Isaac J. Schoenberg (1937). “On certain metric spaces arising from Euclidean spaces by a change of metric and their imbedding in Hilbert space”. *Annals of mathematics*, pp. 787–793 (cit. on p. 3326).
- (Mar. 1942). “Positive definite functions on spheres”. *Duke Math. J.* 9.1, pp. 96–108 (cit. on p. 3323).
- Gregory Shakhnarovich, Trevor Darrell, and Piotr Indyk (2006). *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice (Neural Information Processing)*. The MIT press (cit. on p. 3305).
- Volker Strassen (1969). “Gaussian elimination is not optimal”. *Numerische mathematik* 13.4, pp. 354–356 (cit. on p. 3323).
- Narayanan Sundaram, Aizana Turmukhametova, Nadathur Satish, Todd Mostak, Piotr Indyk, Samuel Madden, and Pradeep Dubey (2013). “Streaming similarity search over one billion tweets using parallel locality-sensitive hashing”. *Proceedings of the VLDB Endowment* 6.14, pp. 1930–1941 (cit. on p. 3306).
- Gregory Valiant (2015). “Finding correlations in subquadratic time, with applications to learning parities and the closest pair problem”. *Journal of the ACM (JACM)* 62.2, p. 13 (cit. on pp. 3322–3324).
- Virginia Vassilevska Williams (2018). “On some fine-grained questions in algorithms and complexity”. In: *Proc. of the International Congress of Mathematicians, Rio de Janeiro 2018*. Ed. by Boyan Sirakov, Paulo Ney de Souza, and Marcelo Viana. In this volume. World Scientific, pp. 3441–3482 (cit. on p. 3306).
- Jingdong Wang, Heng Tao Shen, Jingkuan Song, and Jianqiu Ji (2014). “Hashing for similarity search: A survey”. arXiv: 1408.2927 (cit. on p. 3316).
- Jun Wang, Wei Liu, Sanjiv Kumar, and Shih-Fu Chang (2016). “Learning to hash for indexing big data: a survey”. *Proceedings of the IEEE* 104.1, pp. 34–57 (cit. on p. 3316).
- Ryan Williams (2005). “A new algorithm for optimal 2-constraint satisfaction and its implications”. *Theoretical Computer Science* 348.2-3, pp. 357–365 (cit. on p. 3306).
- Virginia Vassilevska Williams (2012). “Multiplying matrices faster than Coppersmith-Winograd”. In: *Proceedings of the forty-fourth annual ACM symposium on Theory of computing*. ACM, pp. 887–898 (cit. on p. 3323).

Przemysław Wojtaszczyk (1996). *Banach spaces for analysts*. Vol. 25. Cambridge University Press (cit. on p. 3325).

Received 2017-12-01.

ALEXANDR ANDONI

[andoni@cs.columbia.edu](mailto:andoni@cs.columbia.edu)

PIOTR INDYK

[indyk@mit.edu](mailto:indyk@mit.edu)

ILYA RAZENSHTEYN

[ilyaraz@mit.edu](mailto:ilyaraz@mit.edu)



# GROUP, GRAPHS, ALGORITHMS: THE GRAPH ISOMORPHISM PROBLEM

LÁSZLÓ BABAI

## Abstract

Graph Isomorphism (GI) is one of a small number of natural algorithmic problems with unsettled complexity status in the P / NP theory: not expected to be NP-complete, yet not known to be solvable in polynomial time.

Arguably, the GI problem boils down to filling the gap between *symmetry* and *regularity*, the former being defined in terms of automorphisms, the latter in terms of equations satisfied by numerical parameters.

Recent progress on the complexity of GI relies on a combination of the *asymptotic theory of permutation groups* and asymptotic properties of highly regular combinatorial structures called *coherent configurations*. Group theory provides the tools to infer either global symmetry or global irregularity from local information, eliminating the symmetry/regularity gap in the relevant scenario; the resulting global structure is the subject of combinatorial analysis. These structural studies are melded in a *divide-and-conquer* algorithmic framework pioneered in the GI context by Eugene M. Luks (1980).

## 1 Introduction

We shall consider finite structures only; so the terms “graph” and “group” will refer to finite graphs and groups, respectively.

**1.1 Graphs, isomorphism, NP-intermediate status.** A *graph* is a set (the set of *vertices*) endowed with an irreflexive, symmetric binary relation called *adjacency*. Isomorphisms are adjacency-preserving bijections between the sets of vertices. The Graph Isomorphism (GI) problem asks to determine whether two given graphs are isomorphic.

It is known that graphs are *universal* among explicit finite structures in the sense that the isomorphism problem for explicit structures can be reduced in polynomial time to GI (in the sense of *Karp-reductions*<sup>1</sup>) [Hedrlín and Pultr \[1966\]](#) and [Miller \[1979\]](#). This

MSC2010: primary 68Q25; secondary 20B05, 05B30, 05C68, 05C60, 05E30.

<sup>1</sup>For basic concepts of complexity theory we refer to [Garey and Johnson \[1979\]](#).

makes GI a natural algorithmic problem. It is a *polynomial-time verifiable* problem: a candidate isomorphism is easily verified. This puts GI in the complexity class NP. Over time, increasingly strong conjectural evidence has been found that GI is not NP-complete, yet no polynomial-time algorithm is known to solve GI. This puts GI among the small number of natural NP-problems of potentially *intermediate complexity* (neither in P, nor NP-complete). (Another such problem is that of *factoring integers*, cf. [Section 11](#).) The interest in this status of GI was recognized at the dawn of the P / NP theory [Karp \[1972\]](#) and [Garey and Johnson \[1979\]](#).

**1.2 Brief history of the GI problem.** Combinatorial heuristics such as individualization and refinement (I/R) (see [Section 8](#)) have been used for the longest time to reduce the GI search space. It was shown that the “naive refinement” algorithm solves GI for almost all graphs in linear time [Babai, Erdős, and Selkow \[1980\]](#) and [Babai and Kucera \[1979\]](#). Efficient algorithms were found for special classes such as planar graphs [J. E. Hopcroft and Tarjan \[1972\]](#) and [J. E. Hopcroft and Wong \[1974\]](#). These algorithms exploited the combinatorial structure of the graphs concerned. However, combinatorial refinement methods alone cannot succeed in less than exponential time for the general GI problem, as shown in a seminal 1992 paper by [Cai, Fürer, and Immerman \[1992\]](#).

It has long been known that GI is equivalent to determining whether two vertices of a given graph belong to the same orbit of the automorphism group. Refinement procedures have been used to distinguish vertices, trying to refute *symmetry* by discovering *irregularity*. While this gives a first indication of the critical role of the gap between symmetry and regularity to GI, the CFI result shows the futility of trying to close this gap using combinatorial refinement heuristics alone. We use group theory to close a gap of this nature under particular circumstances (see [Theorem 5.3](#) and the paragraph preceding it). The relevant new group theoretic result, the “Unaffected Stabilizers Lemma,” is stated in [Theorem 6.2](#).

Elements of group theory were first introduced into the design of GI algorithms in 1979 [Babai \[1979\]](#). The *tower of groups* method described in that paper produced the following results. A *vertex-colored* graph has a “color” assigned to each vertex; isomorphisms preserve the colors by definition. The *multiplicity* of a color is the number of vertices of that color. The *adjacency matrix* of a graph with  $n$  vertices is the  $n \times n$   $(0, 1)$ -matrix whose  $(i, j)$ -entry is 1 if vertex  $i$  is adjacent to vertex  $j$ , and 0 otherwise. By the *eigenvalues of a graph* we mean the eigenvalues of its adjacency matrix.

**Theorem 1.1.** (a) [Babai \[1979\]](#) and [Furst, J. Hopcroft, and E. Luks \[1980\]](#) *Isomorphism of vertex-colored graphs of bounded color multiplicities can be tested in polynomial time.*  
 (b) [Babai, Grigoryev, and Mount \[1982\]](#) *Isomorphism of graphs with bounded eigenvalue multiplicities can be tested in polynomial time.*

It turns out that the CFI pairs of graphs, i. e., the pairs of graphs shown in [Cai, Fürer, and Immerman \[1992\]](#) to be hard to separate by combinatorial refinement, can be viewed as vertex-colored graphs with color multiplicity 4. This shows that elementary group theory (hardly more than the concept of cosets was used) was already capable of overcoming exponential barriers to combinatorial refinement methods. Modern extensions of the CFI result show that GI is hard for several more general refutation systems (see [Section 11](#)), putting GI in a somewhat paradoxical position in complexity theory (cf. [Section 11](#)).

In-depth use of group theory in the design of GI algorithms arrived with Luks's ground-breaking 1980 paper [E. M. Luks \[1982\]](#). We state the main result of that paper. Adjacent vertices of a graph are called *neighbors*; the *degree* of a vertex is the number of its neighbors.

**Theorem 1.2** (Luks, 1980). *Isomorphism of graphs of bounded degree can be tested in polynomial time.*

Luks's group theoretic method, combined with a combinatorial refinement result by [Zemlyachenko, Korneenko, and Tyshkevich \[1982\]](#), have lead to the *moderately exponential* complexity bound of

$$(1) \quad \exp(O(\sqrt{n \log n})),$$

where  $n$  denotes the number of vertices (Luks, 1983, cf. [Babai and E. M. Luks \[1983\]](#) and [Babai, Kantor, and E. M. Luks \[1983\]](#)). In spite of intermittent progress on important special cases, notably for *strongly regular graphs* [Spielman \[1996\]](#), [Chen, Sun, and Teng \[2013\]](#), [Babai and Wilmes \[2013\]](#), and [Babai, Chen, Sun, Teng, and Wilmes \[2013\]](#) and for *primitive coherent configurations* [Sun and Wilmes \[2015\]](#), Luks's bound (1) for the general case had not been improved until this author's recent announcement [Babai \[2016\]](#) of a quasipolynomial-time algorithm. A *quasipolynomial* function is a function of the form  $\exp(p(\log n))$  for some polynomial  $p$ . A quasipolynomial time bound is a bound of this form where  $n$  is the bit-length of the input; but if we take  $n$  to be the number of vertices of an input graph, the form of the bound will not be affected.

**Theorem 1.3** (B 2015). *Isomorphism of graphs can be tested in quasipolynomial time.*

In this paper we outline the main components of this result. For an introduction to the algorithmic theory of permutation groups we refer to the monograph [Seress \[2003\]](#).

**Disclaimer.** I should emphasize that the results discussed in this paper address the mathematical problem of the *asymptotic worst-case complexity* of GI and have little relevance to practical computation. A suite of remarkably efficient GI packages is available for practical GI testing; [McKay and Piperno \[2014\]](#) give a detailed comparison of methods and performance. These algorithms employ ingenious shortcuts to backtrack search. While

the worst-case performance of these heuristics seems to be exponential, this is increasingly difficult to demonstrate, cf. [Cai, Fürer, and Immerman \[1992\]](#), [Miyazaki \[1996\]](#), and [Neuen and Schweitzer \[2017\]](#).

## 2 The string isomorphism problem

We now define a generalization of the GI problem, introduced by [E. M. Luks \[1982\]](#).

Let  $\Omega$  be a finite set;  $\text{Sym}(\Omega)$  denotes the symmetric group acting on  $\Omega$ . Let  $\Sigma$  be finite alphabet. An  $\Omega$ -string (or just “string”) over  $\Sigma$  is a function  $x : \Omega \rightarrow \Sigma$ . There is a natural action  $x \mapsto x^\sigma$  of  $\text{Sym}(\Omega)$  on the set  $\Sigma^\Omega$  of strings ( $\sigma \in \text{Sym}(\Omega)$ ,  $x \in \Sigma^\Omega$ ). We say that  $\sigma \in \text{Sym}(\Omega)$  is a *G-isomorphism* between the strings  $x$  and  $y$  if  $\sigma \in G$  and  $x^\sigma = y$ . The strings  $x$  and  $y$  are *G-isomorphic*, denoted  $x \cong_G y$ , if such a  $\sigma$  exists. The *String Isomorphism (SI) problem* asks, given  $G$ ,  $x$ , and  $y$ , does  $x \cong_G y$  hold? We refer to  $G$  as the *ambient group*; it is given by a list of generators.

Luks pointed out [E. M. Luks \[ibid.\]](#) that GI reduces to SI by encoding each graph  $X$  by the characteristic function  $f_X$  of its adjacency relation,  $f_X : \binom{\Omega}{2} \rightarrow \{0, 1\}$ , where  $\binom{\Omega}{2}$  denotes the set of unordered pairs of elements of  $\Omega$ . So  $f_X$  is an  $\binom{\Omega}{2}$ -string over the alphabet  $\{0, 1\}$ . The pertinent ambient group is  $\text{Sym}(\Omega)^{(2)}$ , the induced action of  $\text{Sym}(\Omega)$  on the set  $\binom{\Omega}{2}$ . It is easy to see that two graphs are isomorphic if and only if the corresponding  $\binom{\Omega}{2}$ -strings are  $\text{Sym}(\Omega)^{(2)}$ -isomorphic. The actual result we shall discuss concerns the complexity of SI [Babai \[2016\]](#).

**Theorem 2.1** (B 2015). *String isomorphism can be tested in quasipolynomial time.*

[Theorem 1.3](#) is then a corollary. The previous best bound for SI was  $\exp(\tilde{O}(n^{1/2}))$ , where  $n = |\Omega|$  is the length of the strings in question [Babai \[1983\]](#) (cf. [Babai, Kantor, and E. M. Luks \[1983\]](#)). (The tilde hides a polylogarithmic factor.)

Luks also observed that several other problems of computational group theory are polynomial-time equivalent to SI (under Karp-reductions), including the coset intersection, double coset membership, and ‘centralizer in coset’ problems. Given two subgroups  $G, H$  of the symmetric group  $S_n$  and two elements  $\sigma, \pi \in S_n$ , the *Coset Intersection* problem asks whether  $G\sigma \cap H\pi \neq \emptyset$ ; the *double coset membership* problem asks whether  $\sigma \in G\pi H$ , and the *centralizer in coset* problem asks whether there exists an element in the coset  $G\sigma$  that commutes with  $\pi$ . As a consequence, these problems, too, can be solved in quasipolynomial time.

The advantage of approaching GI through the SI problem is that SI permits recursion on the ambient group. This was Luks’s core idea.

### 3 Divide-and-Conquer

In the theory of algorithms, the term “Divide-and-Conquer” refers to recursive procedures that reduce an instance of a computational problem to a moderate number of significantly smaller instances. If our input has size  $n$ , we shall consider instances of size  $\leq 0.9n$  to be “significantly smaller.” Let  $q(n)$  be the number of such smaller instances to which our input is reduced; we refer to  $q(n)$  as the *multiplicative cost* of the reduction. If  $f(n)$  denotes the worst-case cost of processing an input of size  $n$ , this leads to the following recurrence (ignoring the additive cost of assembling all information from the smaller instances, which will typically not affect the cost estimate).

$$(2) \quad f(n) \leq q(n)f(0.9n)$$

Assuming that  $q(n)$  is monotone, this gives the bound  $f(n) \leq q(n)^{O(\log n)}$ , so if  $q(n)$  is quasipolynomially bounded then so is  $f(n)$ . Therefore our goal will be to significantly reduce the problem size at a quasipolynomial multiplicative cost.

### 4 Large primitive permutation groups

Not only did Luks point out that GI reduces to SI, but he also showed that (i) the SI problem for groups with restricted structure can be used to solve the GI problem for certain classes of graphs; and that (ii) SI can be solved efficiently under such structural constraints. The issue of relevance here is bounding the order of primitive permutation groups under structural constraints.

A *permutation group* acting on the set  $\Omega$  (the *permutation domain*) is a subgroup  $G \leq \text{Sym}(\Omega)$ . (The “ $\leq$ ” sign stands for “subgroup.”) The *degree* of  $G$  is  $|\Omega|$ . The set  $x^G = \{x^\sigma \mid \sigma \in G\}$  is the  $G$ -*orbit* of  $x$ ; the orbit has *length*  $|x^G|$ . We say that  $G$  is *transitive* if  $x^G = \Omega$  for some (and therefore any)  $x \in \Omega$ . A transitive group  $G \leq \text{Sym}(\Omega)$  is *primitive* if  $|\Omega| \geq 2$  and there is no nontrivial  $G$ -invariant equivalence relation on  $\Omega$ .

In 1982, Pálffy [1982] and Wolf [1982] showed that primitive *solvable* groups of degree  $n$  have order  $\leq n^c$  where  $c \approx 3.243$ . It turns out that the critical structural parameter of a group for polynomial bounds on the order of its primitive permutation representations is its “thickness.”

**Definition 4.1.** The *thickness*<sup>2</sup>  $\theta(G)$  of a group  $G$  is the largest  $t$  such that the alternating group  $A_t$  is involved in  $G$  as a quotient of a subgroup.

The following result characterizes those hereditary classes of groups (classes that are closed under subgroups and quotients) which have only small primitive permutation representations.

<sup>2</sup>The term “thickness” was coined in Babai [2014].



**Theorem 4.2** (B, Cameron, Pálffy, 1982). *If  $G$  is a primitive permutation group of degree  $n$  and thickness  $t$  then  $|G| = n^{O(t)}$ .*

This result first appeared in Babai, Cameron, and Pálffy [1982]; here it is stated with an improved exponent due to Pyber [1990]. Refined versions were subsequently obtained by Liebeck, Shalev, Maróti; see Liebeck and Shalev [2003, Sec. 3] for a survey of those developments. We note that while the initial motivation for Theorem 4.2 came from the GI problem, the result also found applications in other areas, such as the theory of profinite groups Borovik, Pyber, and Shalev [1996].

E. M. Luks [1982] introduced a group theoretic divide-and-conquer technique to attack the SI problem. Luks's method, combined with the above bounds, yields the following.

**Corollary 4.3.** *The SI problem can be solved in polynomial time if the ambient group is solvable or more generally, if it has bounded thickness.*

Let  $G$  be the stabilizer of an edge in the automorphism group of a connected graph in which every vertex has degree  $\leq k$ . It is easy to see that every composition factor of  $G$  is a subgroup of the symmetric group  $S_{k-1}$ . In particular,  $\theta(G) \leq k-1$  and therefore the SI problem can be solved in polynomial time for such  $G$  as the ambient group. This fact is at the heart<sup>3</sup> of the proof of Theorem 1.2.

While Theorem 4.2 is helpful for groups with small thickness, our interest is in the general case. Luks's technique for SI works in quasipolynomial time as long as the primitive groups involved in the ambient group have quasipolynomially bounded orders. In 1981, building on the then expected completion of the classification of the finite simple groups (CFSG), Cameron [1981] gave a precise characterization of primitive groups of large order. The socle of a group is the product of its minimal normal subgroups. It is known that the socle of a primitive permutation group is a direct product of isomorphic simple groups. For a permutation group  $T \leq \text{Sym}(\Delta)$ , the *product action* of the direct power  $T^k$  on the Cartesian power  $\Delta^k$  is the independent action of each copy of  $T$  on the corresponding coordinate. Wreath product in addition permutes the coordinates by some group “on the top.” For a permutation group  $G \leq \text{Sym}(\Omega)$  we denote by  $G^{(t)}$  the induced action of  $G$  on the set  $\binom{\Omega}{t}$  of unordered  $t$ -tuples of elements of  $\Omega$ .

**Definition 4.4.**  $G \leq S_n$  is a *Cameron group* with parameters  $s, t \geq 1$  and  $k \geq \max(2t + 1, 5)$  if we have  $n = \binom{k}{t}^s$ , the socle of  $G$  is isomorphic to  $A_k^s$  and acts as  $(A_k^{(t)})^s$  in the product action, and  $(A_k^{(t)})^s \leq G \leq S_k^{(t)} \wr S_s$  (wreath product, product action), moreover the induced action  $G \rightarrow S_s$  on the direct factors of the socle is transitive.

---

<sup>3</sup>Theorem 4.2 was not available to Luks at the time; he used a further layer of recurrence so a weaker group-theoretic result was sufficient for his analysis E. M. Luks [1982].

**Theorem 4.5** (Cameron 1981). *For  $n \geq 25$ , if  $G \leq S_n$  has order  $|G| \geq n^{1+\log_2 n}$  then  $G$  is a Cameron group.*

This sharp version of Cameron's theorem [Cameron \[ibid.\]](#) is due to [Maróti \[2002\]](#).

## 5 Luks's method and the bottleneck

In attacking the SI problem, Luks applies a combination of the following two types of recursive operations to the ambient group.

- Descend to a subgroup.
- Process orbits one by one.

Orbit-by-orbit processing leads to ultra-efficient (linear-time) recurrence. Descent to a subgroup  $H \leq G$  incurs a heavy penalty, namely, a multiplicative cost of  $|G : H|$ , so this can only be used to replace the ambient group with a subgroup of small index, and to compensate for the multiplicative cost, such a step needs to lead to significantly reduced problem size. Small primitive groups acting on a minimal system of imprimitivity (system of maximal blocks of a  $G$ -invariant equivalence relation) provide such an opportunity; the orbits of the kernel of the action of such a primitive group have length  $\leq n/2$ , hence orbit-by-orbit processing reduces the problem to significantly smaller instances.

Using [Theorem 4.5](#) we can identify the bottleneck for Luks's method.

**Definition 5.1.** We say that a group  $G$  has a *giant quotient of degree  $m$*  if  $G$  has an epimorphism onto  $S_m$  or  $A_m$ .

**Proposition 5.2.** *For any constant  $C \geq 1$  one can use Luks recurrence for the SI problem to achieve one of the following at a multiplicative cost of  $n^{O(\log n)}$ .*

- (a) *Significantly reduce the problem size.*
- (b) *Reduce the ambient group to a transitive group with a giant quotient of degree  $\geq C \log n$ .*

Our work addresses case (b), the bottleneck situation. The goal is to either confirm or effectively break the symmetry represented by the giant quotient. This inserts another layer of recurrence into Luks's framework: significant reduction of  $m$ , the degree of the giant quotient.

More specifically, let  $G \leq \text{Sym}(\Omega)$  be our ambient group and  $x, y : \Omega \rightarrow \Sigma$  be two strings of which we wish to determine the  $G$ -isomorphisms. Let, further,  $\varphi : G \rightarrow H$  be an epimorphism where  $\text{Alt}(\Gamma) \leq H \leq \text{Sym}(\Gamma)$  for some large set  $\Gamma$ , where  $\text{Alt}(\Gamma)$  denotes the alternating group (even permutations of  $\Gamma$ ). Let  $m = |\Gamma|$  and let  $P(x) =$

$\varphi(\text{Aut}_G(x)) \leq \text{Sym}(\Gamma)$ ; define  $P(y)$  analogously. We say that a group  $K \leq \text{Sym}(\Psi)$  is a *giant* on  $\Psi$  if  $\text{Alt}(\Psi) \leq K \leq \text{Sym}(\Psi)$ .

**Theorem 5.3** (Canonical obstruction to symmetry). *Either  $P(x)$  acts as a giant on a  $P$ -orbit of length  $\geq 0.9m$ , or there exists a  $P(x)$ -invariant canonical  $k$ -ary relational structure  $\mathcal{X}(x)$  on  $\Gamma$  with  $k = O(\log n)$  such that  $\mathcal{X}(x)$  has symmetry defect  $> 0.1$ . Moreover, in each case, we can find, via efficient Luks recurrences, an effective representation of the stated objects.*

We explain the concepts involved in this statement.

By ‘efficient Luks recurrence’ we mean a sequence of Luks operations that significantly reduces the problem size at a multiplicative cost of  $n^{O(\log n)}$ .

In the first case, ‘effective representation’ means we can find a subgroup  $M \leq \text{Aut}_G(x)$  such that  $\varphi(M)$  has a large orbit on which it acts as a giant. Note that  $\text{Aut}_G(x)$  is not known; in fact, determining  $\text{Aut}_G(x)$  is equivalent to the SI problem.

We need to explain the second case. A  $k$ -ary relation on a set  $\Gamma$  is a subset of the Cartesian power  $\Gamma^k$ . A  $k$ -ary relational structure on  $\Gamma$  is a pair  $\mathcal{X} = (\Gamma, \mathcal{R})$  where  $\mathcal{R} = (R_1, \dots, R_r)$  is a list of  $k$ -ary relations  $R_i$  on  $\Gamma$ . ‘Effective representation’ of  $\mathcal{X}$  simply means listing each  $R_i$ . We may assume the  $R_i$  are disjoint, so the total length of the lists is  $\leq m^k$ .

We say that the *symmetry defect* of  $\mathcal{X}$  is  $\geq \alpha$  if every orbit of  $\text{Aut}(\mathcal{X})$  on which  $\text{Aut}(\mathcal{X})$  acts as a giant has size  $\leq (1 - \alpha)m$ .

*Canonicity* of the  $x \mapsto \mathcal{X}(x)$  assignment means this construction is a *functor* from the category of  $G$ -isomorphisms of strings in the set  $\{x, y\}$  (two objects) to the category of isomorphisms of  $k$ -ary relational structures on  $\Gamma$ , so every  $G$ -isomorphism  $\beta_1 \rightarrow \beta_2$  ( $\beta_i \in \{x, y\}$ ) induces an isomorphism  $\mathcal{X}(\beta_1) \rightarrow \mathcal{X}(\beta_2)$ .

The two cases listed in [Theorem 5.3](#) are mutually exclusive by the definition of symmetry defect. The result provides a *constructive obstruction* to certain type of very large symmetry (small symmetry defect); the structure  $\mathcal{X}$  has sufficient *irregularity* to preclude such large symmetry. This is the sense in which, under our special circumstances, we have been able to close a *symmetry vs. regularity gap* (see [Section 1](#)), a key step toward [Theorem 2.1](#).

## 6 Unaffected Stabilizers Lemma

In this section we state a group theoretic result, [Theorem 6.2](#) (a), that is our main mathematical (non-algorithmic) tool for the proof of [Theorem 5.3](#).

For a group  $G \leq \text{Sym}(\Omega)$  and  $x \in \Omega$ , the *stabilizer* of  $x$  in  $G$  is the subgroup  $G_x = \{\sigma \in G \mid x^\sigma = x\}$ . For  $\Delta \subseteq \Omega$ , the *pointwise stabilizer* of  $\Delta$  is the subgroup  $G_{(\Delta)} = \bigcap_{x \in \Delta} G_x$ .

For a group  $G$  and a set  $\Gamma$  we say that the action  $\varphi : G \rightarrow \text{Sym}(\Gamma)$  is a *giant representation* of  $G$  (or a *giant homomorphism*) if the image  $\varphi(G)$  is a giant, i. e.,  $\varphi(G) \geq \text{Alt}(\Omega)$ . We now define our central new concept.

**Definition 6.1** (Affected). Let  $\Omega$  and  $\Gamma$  be sets,  $G \leq \text{Sym}(\Omega)$ , and let  $\varphi : G \rightarrow \text{Sym}(\Gamma)$  be a giant representation. We say that  $x \in \Omega$  is *affected* by  $\varphi$  if the  $\varphi$ -image of the stabilizer  $G_x$  is not a giant, i. e.,  $\varphi(G_x) \not\geq \text{Alt}(\Gamma)$ .

We note that if  $x \in \Omega$  is affected then every element of the orbit  $x^G$  is affected. So we can speak of *affected orbits*.

**Theorem 6.2.** Let  $G \leq \text{Sym}(\Omega)$  be a permutation group of degree  $n = |\Omega|$  and  $\varphi : G \rightarrow S_k$  a giant representation, i. e.,  $\varphi(G) \geq A_k$ . Let  $U \subseteq \Omega$  denote the set of elements of  $\Omega$  not affected by  $\varphi$ . Then the following hold.

- (a) (Unaffected Stabilizers Lemma) Assume  $k > \max\{8, 2 + \log_2 n\}$ . Then  $\varphi$  restricted to  $G_{(U)}$ , the pointwise stabilizer of  $U$ , is still a giant representation, i. e.,  $\varphi(G_{(U)}) \geq A_k$ . In particular,  $U \neq \Omega$  (at least one element is affected).
- (b) (Affected Orbit Lemma) Assume  $k \geq 5$ . If  $\Delta$  is an affected  $G$ -orbit, i. e.,  $\Delta \cap U = \emptyset$ , then  $\ker(\varphi)$  is not transitive on  $\Delta$ ; in fact, each orbit of  $\ker(\varphi)$  in  $\Delta$  has length  $\leq |\Delta|/k$ .

The affected/unaffected dichotomy underlies the core “local certificates” algorithm (Section 7).

Part (b) is an easy exercise; its significance is that it permits efficient Luks reductions on affected orbits.

Part (a) is the central result mentioned. The proof of part (a) builds on the O’Nan–Scott–Aschbacher characterization of primitive permutation groups (L. L. Scott [1980] and Aschbacher and L. Scott [1985], cf. Dixon and Mortimer [1996, Thm. 4.1A]) and depends on the classification of Finite Simple Groups (CFSG)<sup>4</sup> through Schreier’s Hypothesis (a consequence of CFSG) that asserts that the outer automorphism group of every finite simple group is solvable.

Note that part (a) is counter-intuitive: it asserts that if the stabilizer of each  $x \in U$  maps onto  $A_k$  or  $S_k$  then even the intersection of these stabilizers maps onto  $A_k$  or  $S_k$ .

The condition  $k > 2 + \log_2 n$  in part (a) is tight. In fact, there are infinitely many examples with  $k = 2 + \log_2 n$  which have *no affected points*, as shown by the example of a semidirect product  $\mathbb{Z}_2^{k-2} \rtimes A_k \leq \text{AGL}(k-2, 2)$  for even  $k$ , acting on  $n = 2^{k-2}$  elements.

<sup>4</sup>A less tight version of the lemma, still sufficient for the quasipolynomial claim, was recently proved by Pyber [2016] without the CFSG.

## 7 Local certificates

In this section we describe our core algorithmic result. The goal is to categorize ordered  $k$ -tuples of  $\Gamma$ , setting the stage for a combinatorial analysis of the resulting  $k$ -ary relational structure. The method requires the construction of global automorphisms from local information; our key tool is the Unaffected Stabilizers Lemma.

We consider the Luks bottleneck situation. The input is a transitive group  $G \leq \text{Sym}(\Omega)$ , a giant representation  $\varphi : G \rightarrow \text{Alt}(\Gamma)$ , and two strings  $x, y : \Omega \rightarrow \Sigma$ . We write  $n = |\Omega|$  and  $m = |\Gamma|$ . We fix a number  $k > 2 + \log_2 n$  (but not much greater, e. g.,  $k = 3 + \lfloor \log_2 n \rfloor$ ) and assume  $m \geq 10k$ . Subsets  $T \subset \Gamma$  of size  $|T| = k$  will be referred to as “test sets.”

If  $L \leq G$  then  $L$  also acts on  $\Gamma$  via  $\varphi$  so for a test set  $T$  we can speak of the setwise stabilizer of  $T$  in  $L$ ; we write  $L_T$  for this subgroup.

We say that  $T$  is  $L$ -invariant if  $L_T = L$ . We write  $\psi_T : G_T \rightarrow \text{Sym}(T)$  for the map that restricts the domain of  $\varphi$  to  $G_T$  and the codomain to  $\text{Sym}(T)$ . The group  $G_T$  can be computed in polynomial time as  $G_T = \varphi^{-1}(\text{Sym}(\Gamma)_T)$ . Our focus is the (unknown) group  $P(T) := \psi_T(\text{Aut}_{G_T}(x))$ .

**Definition 7.1** (Fullness). Let  $T$  be a test set. We say that  $T$  is *full* with respect to the input string  $x$  if  $P(T) \geq \text{Alt}(T)$ , i. e., the  $G$ -automorphisms of  $x$  induce a giant on  $T$ .

We consider the problem of deciding whether a given test set is full and compute useful certificates of either outcome. We show that this question can efficiently (in time  $k! n^{O(1)}$ ) be reduced to the String Isomorphism problem on inputs of size  $\leq n/k$ .

**Certificate of non-fullness.** We certify non-fullness of the test set  $T$  by computing a permutation group  $M(T) \leq \text{Sym}(T)$  such that (i)  $M(T) \not\geq \text{Alt}(T)$  and (ii)  $M(T) \geq P(T)$  ( $M(T)$  is guaranteed to contain the projection of the  $G$ -automorphism group of  $x$ ). Such an “encasing group”  $M(T)$  can be thought of as a constructive refutation of fullness.

**Certificate of fullness.** We certify fullness of the test set  $T$  by computing a permutation group  $K(T) \leq \text{Sym}(\Omega)$  such that (i)  $K(T) \leq \text{Aut}_{G_T}(x)$  and (ii)  $\psi_T(K(T)) \geq \text{Alt}(T)$ . Note that  $K(T) \leq P(T)$ , so  $K(T)$  represents a polynomial-time verifiable proof of fullness of  $T$ .

Our ability to find  $K(T)$ , the certificate of fullness, may be surprising because it means that from a local start (that may take only a small segment of  $x$  into account), we have to build up global automorphisms (automorphisms of the full string  $x$ ). Our ability to do so critically depends on the “Unaffected Stabilizers Lemma” (Theorem 6.2 (a)).

**Theorem 7.2** (Local certificates). *Let  $T \subseteq \Gamma$  where  $|T| = k$  is a test set. Assume  $\max\{8, 2 + \log_2 n\} < k \leq m/10$  (where  $m = |\Gamma|$ ). By making  $\leq k! n^2$  calls to SI problems on domains of size  $\leq n/k$  and performing  $k! n^{O(1)}$  computation we can decide whether  $T$  is full and*

- (a) if  $T$  is full, find a certificate  $K(T) \leq \text{Aut}_G(\mathfrak{x})$  of fullness
- (b) if  $T$  is not full, find a certificate  $M(T) \leq \text{Sym}(T)$  of non-fullness.

To aggregate the local certificates, first we consider the group  $F$  generated by the fullness certificates. If the support of  $\varphi(F) \leq \text{Sym}(\Gamma)$  has at least  $m/10$  elements then the structure of  $\varphi(F)$  suffices for the proof of [Theorem 5.3](#). In the alternative, non-fullness certificates dominate. In this case a slight extension of [Theorem 7.2](#) is needed, to encase not only the group  $\psi_T(\text{Aut}_{G_T}(\mathfrak{x}))$  but also the images of the cosets  $\text{Iso}_{G_{T,T'}}(\beta_1, \beta_2)$  for all pairs  $T, T'$  of test sets and all choices of  $\beta_1, \beta_2 \in \{\mathfrak{x}, \mathfrak{y}\}$ . The result will be two classifications of the ordered  $k$ -tuples of  $\Gamma$ , one associated with  $\mathfrak{x}$ , the other with  $\mathfrak{y}$ , yielding the canonical assignment  $\mathfrak{x} \mapsto \mathfrak{X}(\mathfrak{x})$  and  $\mathfrak{y} \mapsto \mathfrak{X}(\mathfrak{y})$ .

## 8 Individualization and refinement

We consider  $k$ -ary *partition structures*  $\mathfrak{X} = (\Gamma, \mathfrak{R})$  where  $\mathfrak{R} = (R_1, \dots, R_r)$  is a partition of  $\Gamma^k$ . We think of such a structure as a *coloring*  $c : \Gamma^k \rightarrow \{1, \dots, r\}$  where  $c(\vec{x}) = i$  if  $\vec{x} \in R_i$  ( $\vec{x} \in \Gamma^k$ ). We also write  $\mathfrak{X} = (\Gamma, c)$  instead of  $\mathfrak{X} = (\Gamma, \mathfrak{R})$ . A *refinement* of a coloring  $c$  is a coloring  $c'$  such that  $(\forall \vec{x}, \vec{y} \in \Gamma^k)(c'(\vec{x}) = c'(\vec{y}) \implies c(\vec{x}) = c(\vec{y}))$ .

An assignment  $\mathfrak{X} \mapsto \mathfrak{X}'$  is *canonical* if it is defined by a functor between categories of isomorphisms of structures.

By a *binary configuration* we mean a binary partition structure  $\mathfrak{X} = (\Gamma, c)$  such that

- (i)  $(\forall x, y, z \in \Gamma)(c(x, y) = c(z, z) \implies x = y)$  and
- (ii)  $(\forall x, y \in \Gamma)(c(x, y) \text{ determines } c(y, x))$ .

The *Weisfeiler–Leman* canonical refinement process (WL) [Weisfeiler \[1968\]](#) and *On construction and identification of graphs* [\[1976\]](#) takes a binary configuration and with every pair  $(x, y) \in \Gamma^2$  associates the list  $c'(x, y) = (c(x, y), d_{i,j}(x, y) \mid i, j = 1, \dots, r)$  where  $d_i(x, y) = |\{z \in \Gamma \mid c(x, z) = i, c(z, y) = j\}|$ . This is clearly a canonical refinement.

Let  $\mathfrak{X} = (\Gamma, c)$  be a  $k$ -ary partition structure. We assign colors to the elements by setting  $c(x) = c(x, \dots, x)$ . *Individualizing* an element  $x \in \Gamma$  means assigning it a special color, thereby introducing irregularity. This irregularity propagates via canonical refinement, reducing the isomorphism search space. Let  $\mathfrak{X}_x$  denote  $\mathfrak{X}$  with  $x \in \Gamma$  individualized. Then  $\mathfrak{X} \cong \mathfrak{Y} \iff (\exists y \in \Gamma)(\mathfrak{X}_x \cong \mathfrak{Y}_y)$ . So progress comes at a multiplicative cost of  $m = |\Gamma|$ . The multiplicative cost of individualizing  $t$  points is  $n^t$ , so we need  $t \leq \text{polylog}$  for a quasipolynomial complexity bound.

## 9 Coherent configurations

The stable configurations of the WL process (where no proper refinement is obtained) are called *coherent configurations*. This concept goes back to [Schur \[1933\]](#) who abstracted its axiom from the *orbital configurations* of permutation groups. An *orbital* of  $G \leq \text{Sym}(\Omega)$  is an orbit of the induced action of  $G$  on  $\Omega \times \Omega$ . Let  $\mathfrak{X}(G)$  denote the configuration on  $\Omega$  with the orbitals as the relations. This configuration is clearly coherent, but there are many coherent configurations that do not arise this way. For  $v \geq 2k + 1$ , the *Johnson scheme*  $\mathfrak{J}(v, k)$  has  $\binom{v}{k}$  vertices; it is defined as the orbital configuration of the group  $S_v^{(k)}$  (induced action of  $S_v$  on unordered  $k$ -tuples).

A coherent configuration is *homogeneous* if every point has the same color. A homogeneous configuration is *primitive* if  $|\Gamma| \geq 2$  and each off-diagonal color (relation) is a (strongly) connected (directed) graph. We note that the orbital configuration  $\mathfrak{X}(G)$  of a permutation group  $G$  is homogeneous iff  $G$  is transitive and  $\mathfrak{X}(G)$  is primitive iff  $G$  is primitive. The *rank* of a configuration is the number of colors, so for  $|\Gamma| \geq 2$  the rank is at least 2. The only rank-2 configuration is the *clique*; its automorphism group is  $\text{Sym}(\Gamma)$ . The Johnson scheme  $\mathfrak{J}(v, k)$  has rank  $k + 1$ .

The WL process and its natural  $k$ -ary generalization play a key role in the combinatorial analysis of the  $k$ -ary relational structures handed down by the Local Certificates algorithm.

## 10 Combinatorial partitioning

Recall that we have a giant homomorphism  $\varphi : G \rightarrow \text{Sym}(\Gamma)$  for some ‘ideal domain’  $\Gamma$  and we are given a canonical  $k$ -ary partition structure  $\mathfrak{X}(x) = (\Gamma, c_x)$  with symmetry defect  $\geq 0.1$  where  $x$  is the input string. Here  $k = O(\log n)$  where  $n = |\Omega|$  is the size of our original domain. Recall that our recursive goal is to significantly reduce the size of the ideal domain at moderate multiplicative cost. Ideally we would like to achieve this by finding a *good canonical coloring* of  $\Gamma$  (no color has multiplicity greater than  $0.9m$ ) or a *good equipartition*, i. e., a nontrivial canonical equipartition of the dominant ( $> 0.9m$ ) vertex-color class.

This goal cannot be achieved because of the resilience of the Johnson schemes to canonical partitioning.

**Proposition 10.1** (Resilience of Johnson schemes). *The multiplicative cost of a good canonical coloring or a good canonical equipartition of the Johnson scheme  $\mathfrak{J}(v, t)$  is  $\geq (4t)^{v/(4t)}$ .*

The proof shows that if we pay less than exponential multiplicative cost then our Johnson scheme is simply reduced to a slightly smaller Johnson scheme.

Note that  $t = 2$  is an interesting case, largely responsible for the lack of progress over the  $\exp(\tilde{O}(\sqrt{n}))$  bound for a long time.

The good news is that in a sense, the Johnson schemes are the only obstacles.

So our modified goal will be to find either (a) a good canonical coloring, or (b) a good canonical equipartition, or (c) a canonically embedded Johnson scheme on a dominant vertex-color class. In item (c), canonical embedding means a functor from the isomorphisms of the input structures  $\mathcal{X}$  to the isomorphisms of the secondary structures whose vertex set is a dominant vertex-color class in  $\Gamma$  (under a canonical coloring).

We achieve this goal in two stages: first we go from  $k$ -ary to binary (Design Lemma) and then from binary to the desired goal (Split-or-Johnson).

**Theorem 10.2** (Design lemma). *Let  $\mathcal{X} = (\Gamma, c)$  be a  $k$ -ary partition structure with  $m = |\Gamma|$  elements,  $2 \leq k \leq m/2$ , and symmetry defect  $\geq 0.1$ . Then in time  $m^{O(k)}$  we can find a sequence  $S$  of at most  $k - 1$  vertices such that after individualizing each element of  $S$  we can either find*

- (a) *a good canonical coloring of  $\Gamma$ , or*
- (b) *a good canonical equipartition of  $\Gamma$ , or*
- (c) *a good canonically embedded primitive coherent configuration of rank  $\geq 3$ .*

Here canonicity is relative to the arbitrary choice of the sequence  $S$ .

Outcomes (a) and (b) allow for efficient Luks reduction. Case (c) requires further processing.

**Theorem 10.3** (Split-or-Johnson). *Given a primitive coherent configuration  $\mathcal{X} = (\Gamma, c)$  of rank  $\geq 3$ , at quasipolynomial multiplicative cost we can find either*

- (a) *a good canonical coloring of  $\Gamma$ , or*
- (b) *a good canonical equipartition of  $\Gamma$ , or*
- (c) *a good canonically embedded nontrivial Johnson scheme.*

Here canonicity is relative to the arbitrary choices made that resulted in the multiplicative cost. The trivial Johnson schemes are the cliques  $\mathfrak{J}(v, 1)$ .

Outcomes (a) and (b) again allow for efficient Luks reduction. Outcome (c) provides even greater efficiency. Assume the canonically embedded Johnson scheme is  $\mathfrak{J}(m', t)$ ; so  $m \geq \binom{m'}{t} \geq \binom{m'}{2}$  and therefore  $m' < 1 + \sqrt{2m}$ . Now  $\text{Aut}(\mathfrak{J}(m', t)) \cong S_{m'}$ , so we can replace  $\Gamma$  by a set  $\Gamma'$  of size  $m' = O(\sqrt{m})$ , a dramatic reduction of the size of the ideal domain.

**Overall algorithm.** We follow Luks's algorithm until we hit a bottleneck, at which time an "ideal domain"  $\Gamma$  arises and our recursive goal becomes to significantly reduce the size of



the ideal domain. First we use our central group theoretic algorithm (“Local certificates”), based on the “Unaffected Stabilizers Lemma,” to construct a canonical structure on  $\Gamma$  of logarithmic arity and with non-negligible symmetry defect. Then we use our combinatorial partitioning algorithms to achieve the desired reduction. Once  $\Gamma$  itself becomes very small (polylogarithmic), we can individualize all of its elements, yielding a significant reduction of  $n$ , the size of the input string.

## 11 Paradoxes of Graph Isomorphism

GI is perceived to be an “easy” computational problem. As discussed in the Introduction (see “Disclaimer”), it is efficiently solved in practice. It is also provably easy on average. Our result shows it has rather low worst-case time complexity. In comparison, the problem of factoring integers is perceived to be “hard” – the assumption that it is hard, not only in the worst case but even of average, is the basis of the RSA cryptosystem and many other cryptographic applications. Yet, by common measures used in structural complexity theory, GI seems harder than factoring. The decision version of the factorization problem is in  $\text{NP} \cap \text{coNP}$ ; this is not known to be the case for GI. Factoring is solvable in polynomial time in the quantum computation model; no quantum advantage has been found (in spite of significant effort) for GI. Most remarkable is the series of recent hardness results for GI in proof complexity, inspired by the CFI result. It turns out that in commonly studied hierarchies of semialgebraic and algebraic proof systems, isomorphism of certain pairs of graphs cannot be refuted on levels lower than  $cn$  for some constant  $c > 0$  (where  $n$  is the number of vertices), corresponding to refutation proofs of exponential length in these systems [Atserias and Maneva \[2013\]](#), [O’Donnell, Wright, Wu, and Zhou \[2014\]](#), and [Berkholz and Grohe \[2015\]](#). (Cf. [Atserias and Ochremiak \[2017\]](#) for an overview of these and related systems.)

## 12 Open problems

**Complexity theory.** It is not known whether GI belongs to  $\text{coNP}$ . On the other hand, it is also not known whether  $\text{P}$  has logspace reductions to GI. This is equivalent to a logspace reduction of the *circuit value problem* (CVP) to GI. The CVP takes a Boolean circuit and an input to the circuit and asks to evaluate the circuit. Such a reduction would be viewed as strong evidence against the existence of an efficient parallel algorithm for GI.

While GI is universal over isomorphism problems for explicit structures, there are interesting classes of isomorphism problems for non-explicit structures that are also not expected to be  $\text{NP}$ -complete (based on strong evidence from the theory of interactive proofs), yet cannot currently be solved in less than exponential time. Perhaps the simplest among

them is the *code equivalence problem* that asks, given two subspaces  $U$  and  $V$  of  $\mathbb{F}^n$  for some finite field  $\mathbb{F}$ , is there a permutation  $\sigma \in S_n$  such that  $U^\sigma = V$ ? Here  $\sigma$  acts on  $\mathbb{F}^n$  by permuting the coordinates.

Can GI be solved in quasipolynomial time and *polynomial space*? (Luks)

Can *canonical forms* of graphs be constructed in quasipolynomial time? (Cf. Babai and E. M. Luks [1983].)

Can isomorphism of hypergraphs be decided in time, quasipolynomial in the number of vertices and *polynomial in the number of edges*?

**Combinatorics.** The author’s decades-old project to find combinatorial relaxations of Cameron’s [Theorem 4.5](#) has seen major progress recently, made by PhD students. *Cameron schemes* are the orbital configurations of Cameron groups ([Definition 4.4](#)). Let us say that a primitive coherent configuration is a *non-Cameron PCC* if it is not a Cameron scheme. The author has circulated various versions of the following conjectures for some time.

**Conjecture 12.1.** There exists a polynomial  $p$  such that the following hold. Let  $\mathcal{X}$  be a non-Cameron PCC with  $n$  vertices. Let  $G = \text{Aut}(\mathcal{X})$ . Then

- (a)  $\theta(\text{Aut}(\mathcal{X})) \leq p(\log n)$  (where  $\theta$  denotes the thickness, [Definition 4.1](#))  
(polylogarithmically bounded thickness)
- (b)  $|G| \leq \exp(p(\log n))$  (quasipolynomially bounded order)

Part (a) obviously follows from part (b). Regarding (b), for non-Cameron PCCs, an upper bound  $|G| \leq \exp(\tilde{O}(\sqrt{n}))$  was proved in Babai [1981] in 1981. After no progress for three and a half decades, in a recent *tour de force* of combinatorial reasoning, Sun and Wilmes reduced this upper bound to  $\exp(\tilde{O}(n^{1/3}))$ , building a new combinatorial structure theory of primitive coherent configurations along the way. The weaker Conjecture (a) has been confirmed for rank-3 configurations (essentially, strongly regular graphs) in Babai [2014] (2014). Overcoming an array of technical obstacles through a powerful combination of structural and spectral theory, Kivva [2017] very recently confirmed (a) for rank-4 configurations. These are major steps, and raise the hope of further progress, although the technical challenges seem daunting.

## References

- M. Aschbacher and L. Scott (1985). “Maximal subgroups of finite groups”. *J. Algebra* 92.1, pp. 44–80. MR: [772471](#) (cit. on p. 3345).
- Albert Atserias and Elitza Maneva (2013). “Sherali-Adams relaxations and indistinguishability in counting logics”. *SIAM J. Comput.* 42.1, pp. 112–137. MR: [3033123](#) (cit. on p. 3350).

- Albert Atserias and Joanna Ochremiak (2017). “[Proof complexity meets algebra](#)”. In: *44th International Colloquium on Automata, Languages, and Programming*. Vol. 80. LIPIcs. Leibniz Int. Proc. Inform. Schloss Dagstuhl. Leibniz-Zent. Inform., Wadern, Art. No. 110, 14. arXiv: [1711.07320](#). MR: [3685850](#) (cit. on p. [3350](#)).
- L. Babai, P. J. Cameron, and P. P. Pálffy (1982). “[On the orders of primitive groups with restricted nonabelian composition factors](#)”. *J. Algebra* 79.1, pp. 161–168. MR: [679977](#) (cit. on p. [3342](#)).
- László Babai (1979). “[Monte-Carlo algorithms in graph isomorphism testing](#)”. *Université de Montréal Technical Report, DMS* 79-10, p. 42 (cit. on p. [3338](#)).
- (1981). “[On the order of uniprimitive permutation groups](#)”. *Ann. of Math.* (2) 113.3, pp. 553–568. MR: [621016](#) (cit. on p. [3351](#)).
  - (1983). “Permutation Groups, Coherent Configurations, and Graph Isomorphism”. PhD thesis. D. Sc. Thesis (Hungarian), Hungarian Academy of Sciences (cit. on p. [3340](#)).
  - (2014). “On the automorphism groups of strongly regular graphs I”. In: *ITCS’14—Proceedings of the 2014 Conference on Innovations in Theoretical Computer Science*. ACM, New York, pp. 359–368. MR: [3359489](#) (cit. on pp. [3341](#), [3351](#)).
  - (2016). “[Graph isomorphism in quasipolynomial time \[extended abstract\]](#)”. In: *STOC’16—Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, New York, pp. 684–697. arXiv: [1512.03547](#). MR: [3536606](#) (cit. on pp. [3339](#), [3340](#)).
- László Babai, Xi Chen, Xiaorui Sun, Shang-Hua Teng, and John Wilmes (2013). “Faster canonical forms for strongly regular graphs”. In: *Foundations of Computer Science (FOCS), 2013 IEEE 54th Annual Symposium on*. IEEE, pp. 157–166 (cit. on p. [3339](#)).
- László Babai, Paul Erdős, and Stanley M. Selkow (1980). “[Random graph isomorphism](#)”. *SIAM J. Comput.* 9.3, pp. 628–635. MR: [584517](#) (cit. on p. [3338](#)).
- László Babai, D Yu Grigoryev, and David M Mount (1982). “Isomorphism of graphs with bounded eigenvalue multiplicity”. In: *Proceedings of the fourteenth annual ACM symposium on Theory of computing*. ACM, pp. 310–324 (cit. on p. [3338](#)).
- László Babai, William M Kantor, and Eugene M Luks (1983). “Computational complexity and the classification of finite simple groups”. In: *Foundations of Computer Science, 1983., 24th Annual Symposium on*. IEEE, pp. 162–171 (cit. on pp. [3339](#), [3340](#)).
- László Babai and Ludik Kucera (1979). “Canonical labelling of graphs in linear average time”. In: *Foundations of Computer Science, 1979., 20th Annual Symposium on*. IEEE, pp. 39–46 (cit. on p. [3338](#)).
- László Babai and Eugene M Luks (1983). “Canonical labeling of graphs”. In: *Proceedings of the fifteenth annual ACM symposium on Theory of computing*. ACM, pp. 171–183 (cit. on pp. [3339](#), [3351](#)).

- László Babai and John Wilmes (2013). “Quasipolynomial-time canonical form for Steiner designs”. In: *STOC’13—Proceedings of the 2013 ACM Symposium on Theory of Computing*. ACM, New York, pp. 261–270. MR: [3210787](#) (cit. on p. [3339](#)).
- Christoph Berkholz and Martin Grohe (2015). “Limitations of algebraic approaches to graph isomorphism testing”. In: *Automata, languages, and programming. Part I*. Vol. 9134. Lecture Notes in Comput. Sci. Springer, Heidelberg, pp. 155–166. arXiv: [1502.05912](#). MR: [3382436](#) (cit. on p. [3350](#)).
- Alexandre V. Borovik, Laszlo Pyber, and Aner Shalev (1996). “Maximal subgroups in finite and profinite groups”. *Trans. Amer. Math. Soc.* 348.9, pp. 3745–3761. MR: [1360222](#) (cit. on p. [3342](#)).
- Jin-Yi Cai, Martin Fürer, and Neil Immerman (1992). “An optimal lower bound on the number of variables for graph identification”. *Combinatorica* 12.4, pp. 389–410. MR: [1194730](#) (cit. on pp. [3338](#)–[3340](#)).
- Peter J. Cameron (1981). “Finite permutation groups and finite simple groups”. *Bull. London Math. Soc.* 13.1, pp. 1–22. MR: [599634](#) (cit. on pp. [3342](#), [3343](#)).
- Xi Chen, Xiaorui Sun, and Shang-Hua Teng (2013). “Multi-stage design for quasipolynomial-time isomorphism testing of Steiner 2-systems”. In: *STOC’13—Proceedings of the 2013 ACM Symposium on Theory of Computing*. ACM, New York, pp. 271–280. MR: [3210788](#) (cit. on p. [3339](#)).
- John D. Dixon and Brian Mortimer (1996). *Permutation groups*. Vol. 163. Graduate Texts in Mathematics. Springer-Verlag, New York, pp. xii+346. MR: [1409812](#) (cit. on p. [3345](#)).
- Merrick Furst, John Hopcroft, and Eugene Luks (1980). “Polynomial-time algorithms for permutation groups”. In: *21st Annual Symposium on Foundations of Computer Science (Syracuse, N.Y., 1980)*. IEEE, New York, pp. 36–41. MR: [596045](#) (cit. on p. [3338](#)).
- Michael R. Garey and David S. Johnson (1979). *Computers and intractability*. A guide to the theory of NP-completeness, A Series of Books in the Mathematical Sciences. W. H. Freeman and Co., San Francisco, Calif., pp. x+338. MR: [519066](#) (cit. on pp. [3337](#), [3338](#)).
- Z. Hedrlín and A. Pultr (1966). “On full embeddings of categories of algebras”. *Illinois J. Math.* 10, pp. 392–406. MR: [0191858](#) (cit. on p. [3337](#)).
- J. E. Hopcroft and R. E. Tarjan (1972). “Isomorphism of planar graphs”, pp. 131–152, 187–212. MR: [0403302](#) (cit. on p. [3338](#)).
- J. E. Hopcroft and J. K. Wong (1974). “Linear time algorithm for isomorphism of planar graphs: preliminary report”, pp. 172–184. MR: [0433964](#) (cit. on p. [3338](#)).
- Richard M. Karp (1972). “Reducibility among combinatorial problems”, pp. 85–103. MR: [0378476](#) (cit. on p. [3338](#)).

- Bohdan Kivva (2017). “On the automorphism groups of distance-regular graphs and rank-4 primitive coherent configurations”. Manuscript, University of Chicago (cit. on p. 3351).
- Martin W. Liebeck and Aner Shalev (2003). “Bases of primitive permutation groups”. In: *Groups, combinatorics & geometry (Durham, 2001)*. World Sci. Publ., River Edge, NJ, pp. 147–154. MR: 1994965 (cit. on p. 3342).
- Eugene M. Luks (1982). “Isomorphism of graphs of bounded valence can be tested in polynomial time”. *J. Comput. System Sci.* 25.1, pp. 42–65. MR: 685360 (cit. on pp. 3339, 3340, 3342).
- Attila Maróti (2002). “On the orders of primitive groups”. *J. Algebra* 258.2, pp. 631–640. MR: 1943938 (cit. on p. 3343).
- Brendan D. McKay and Adolfo Piperno (2014). “Practical graph isomorphism, II”. *J. Symbolic Comput.* 60, pp. 94–112. arXiv: 1301.1493. MR: 3131381 (cit. on p. 3339).
- Gary L. Miller (1979). “Graph isomorphism, general remarks”. *J. Comput. System Sci.* 18.2, pp. 128–142. MR: 532172 (cit. on p. 3337).
- Takunari Miyazaki (1996). *Luks’s reduction of Graph isomorphism to code equivalence*. Comment on The Math Forum (cit. on p. 3340).
- Daniel Neuen and Pascal Schweitzer (May 2017). “An exponential lower bound for Individualization-Refinement algorithms for Graph Isomorphism”. arXiv: 1705.03283 (cit. on p. 3340).
- Ryan O’Donnell, John Wright, Chenggang Wu, and Yuan Zhou (2014). “Hardness of robust graph isomorphism, Lasserre gaps, and asymmetry of random graphs”. In: *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. ACM, New York, pp. 1659–1677. MR: 3376480 (cit. on p. 3350).
- On construction and identification of graphs* (1976). Lecture Notes in Mathematics, Vol. 558. With contributions by A. Lehman, G. M. Adelson-Velsky, V. Arlazarov, I. Faragev, A. Uskov, I. Zuev, M. Rosenfeld and B. Weisfeiler, Edited by Boris Weisfeiler. Springer-Verlag, Berlin-New York, pp. xiv+237. MR: 0543783 (cit. on p. 3347).
- P. P. Pálffy (1982). “A polynomial bound for the orders of primitive solvable groups”. *J. Algebra* 77.1, pp. 127–137. MR: 665168 (cit. on p. 3341).
- László Pyber (1990). Unpublished (cit. on p. 3342).
- (May 2016). “A CFSG-free analysis of Babai’s quasipolynomial GI-algorithm”. arXiv: 1605.08266 (cit. on p. 3345).
- Issai Schur (1933). “Zur Theorie der einfach transitiven Permutationsgruppen”. *Sitzungsb. Preuss. Akad. Wiss.* Pp. 598–623 (cit. on p. 3348).
- Leonard L. Scott (1980). “Representations in characteristic  $p$ ”. In: *The Santa Cruz Conference on Finite Groups (Univ. California, Santa Cruz, Calif., 1979)*. Vol. 37. Proc. Sympos. Pure Math. Amer. Math. Soc., Providence, R.I., pp. 319–331. MR: 604599 (cit. on p. 3345).

- Ákos Seress (2003). *Permutation group algorithms*. Vol. 152. Cambridge Tracts in Mathematics. Cambridge University Press, Cambridge, pp. x+264. MR: [1970241](#) (cit. on p. [3339](#)).
- Daniel A. Spielman (1996). “Faster isomorphism testing of strongly regular graphs”. In: *Proceedings of the Twenty-eighth Annual ACM Symposium on the Theory of Computing (Philadelphia, PA, 1996)*. ACM, New York, pp. 576–584. MR: [1427556](#) (cit. on p. [3339](#)).
- Xiaorui Sun and John Wilmes (2015). “Faster canonical forms for primitive coherent configurations (extended abstract)”. In: *STOC’15—Proceedings of the 2015 ACM Symposium on Theory of Computing*. ACM, New York, pp. 693–702. MR: [3388249](#) (cit. on p. [3339](#)).
- Boris Weisfeiler (1968). “A reduction of a graph to a canonical form and an algebra arising during this reduction”. *Nauchno-Tekhnicheskaya Informatsiya* 9, pp. 12–16 (cit. on p. [3347](#)).
- Thomas R. Wolf (1982). “Solvable and nilpotent subgroups of  $GL(n, q^m)$ ”. *Canad. J. Math.* 34.5, pp. 1097–1111. MR: [675682](#) (cit. on p. [3341](#)).
- V. N. Zemlyachenko, N. M. Korneenko, and R. I. Tyshkevich (1982). “The graph isomorphism problem”. *Zap. Nauchn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* 118. The theory of the complexity of computations, I, pp. 83–158, 215. MR: [659084](#) (cit. on p. [3339](#)).

Received 2017-12-11.

LÁSZLÓ BABAI  
UNIVERSITY OF CHICAGO  
[laci@cs.uchicago.edu](mailto:laci@cs.uchicago.edu)



# DELEGATING COMPUTATION VIA NO-SIGNALING STRATEGIES

Yael Tauman Kalai

## Abstract

Efficient verification of computation, also known as *delegation of computation*, is one of the most fundamental notions in computer science, and in particular it lies at the heart of the P vs. NP question.

This article contains a high level overview of the evolution of proofs in computer science, and shows how this evolution is instrumental to solving the problem of delegating computation. We highlight a curious connection between the problem of delegating computation and the notion of *no-signaling strategies* from quantum physics.

## 1 Introduction

The problem of delegating computation considers the setting where a computationally weak device (the client) wishes to offload his computations to a powerful device (the server). Such a client may not trust the server, and would therefore want the server to accompany the result of each computation with an *easy-to-verify proof* of correctness. Clearly, the time it takes to verify such a proof should be significantly lower than the time it takes to do the computation from scratch, since otherwise there is no point of delegating this computation to begin with. At the same time, it is desirable that the time it takes to generate a proof is not too high (i.e., not significantly higher than doing the computation) since otherwise it will be too costly to delegate this computation.

Efficient delegation carries significance to applications. In many cases, computation today is asymmetric, where lightweight computations are done locally, and large computational tasks are performed off-site (e.g. by a cloud server). In addition, complex computations are often delegated to powerful (possibly untrusted) hardware. The ability to verify that the computation is carried out correctly without investing significant computational resources is obviously useful in these situations. The applicability of delegation schemes goes even further. For example, efficient verification of computation is used today as a



building block in one of the prominent, and widely used, crypto currencies [Ben-Sasson, Chiesa, Garman, Green, Miers, Tromer, and Virza \[2014\]](#).

Aside from its practical value, efficient verification of computation is one of the most fundamental notions in computer science. Indeed, one of the most basic computational objects in computer science is the complexity class NP, which is defined as the class of problems whose computation can be verified “efficiently”, where an “efficient” computation is defined as one that takes polynomial time.

Unfortunately, the power of the complexity class NP is still unknown, and the question of whether  $NP \neq P$ , which is arguably *the* most important open problem in computer science, remains open.<sup>1</sup> Thus, in a sense, we don’t even know if verifying a proof is easier than finding a proof from scratch! Moreover, even under the widely believed assumption that  $NP \neq P$ , it is widely believed that general  $T$ -time computations do not have a proof that is verifiable in time significantly smaller than  $T$ . This seems to pose an insurmountable barrier to the problem of delegating general purpose computations.

We overcome this barrier by abandoning the traditional (thousands-year-old) notion of a proof being a piece of text, and instead utilize a beautiful and instrumental line of work, motivated by cryptography, where various alternative proof models were proposed and studied. These proof models include *interactive proofs*, *multi-prover interactive proofs* and *probabilistically checkable proofs*, which we elaborate on below.

Jumping ahead, in this article we show how this line of work, together with the use of cryptography and an intriguing connection to *no-signaling strategies* from quantum physics, can be used to construct secure delegation schemes.

**Interactive Proofs.** The notion of interactive proofs was defined by [Goldwasser, Micali, and Rivest \[1988\]](#). Their goal was to construct *zero-knowledge proofs*, which intuitively are proofs that reveal no information, beyond the validity of the statement being proven. Goldwasser et. al. noticed that such a notion is not achievable using traditional proofs, and hence they introduced a new proof model, which they called *interactive proofs*.

In contrast to a traditional proof, an interactive proof is an interactive process between a prover and an efficient (i.e., polynomial time) verifier. Interestingly (and oddly), the verifier is allowed to toss coins, and let these coin tosses determine the questions he asks the prover. Importantly, whereas traditionally, it is required that false statements do not have a valid proof, here we require that a (cheating) prover cannot convince the verifier of the correctness of any false statement, except with very small probability (over the verifier’s coin tosses). We denote the class of all languages that have an interactive proof by IP. Note that allowing the verifier to be randomized is crucial, since otherwise, the

---

<sup>1</sup>The complexity class P consists of the class of problems that can be computed in polynomial time.

prover can predict the verifier's questions and can send a single document emulating the entire interaction. Thus without the use of randomness we would get  $IP = NP$ .

Interestingly, with the use of randomness, it seems that the class  $IP$  is significantly more powerful than the class  $NP$ . The celebrated results of [Lund, Fortnow, Karloff, and Nisan \[1992\]](#) and [Shamir \[1992\]](#) prove that  $IP = PSPACE$ , where  $PSPACE$  is the class of all languages that can be computed by a Turing machine that uses polynomial space and with arbitrarily long runtime. This class of  $PSPACE$  is believed to be significantly larger than  $NP$ , and hence the  $IP$  proof system seems to be very powerful. Looking into the  $IP = PSPACE$  theorem more closely, it says that any computation that takes time  $T$  and space  $S$  has an interactive proof, where the verifier runs in time proportional to  $S$  (and the length of the statement being proven). However, the runtime of the prover is significantly higher than  $T$ . In the original works of [Lund, Fortnow, Karloff, and Nisan \[1992\]](#) and [Shamir \[1992\]](#) the runtime of the prover was proportional to  $2^{S^2}$ , and thus is super-polynomial (in  $T$ ) even for log-space computations (i.e., computations that take space  $O(\log T)$ ).<sup>2</sup>

**1.1 Our Goal: Doubly-Efficient Proofs.** The computation delegation challenge requires not only efficiently verifiable proofs but in actuality *doubly efficiently verifiable proofs* [Goldwasser, Kalai, and G. N. Rothblum \[2008\]](#). Such proofs require the complexity of the verifier to be efficient *without paying* a noticeable penalty in increasing the provers running time. This is in contrast to the results of the early 90's, that were focused on the question of which computations have an “easy to verify proof” of correctness, without putting any restriction on the runtime of the prover.

In the most basic setting, in a *delegation scheme* a prover  $P$  proves to a verifier  $V$  the correctness of an arbitrary time  $T$  computation. Our most basic goal is to construct a delegation scheme, with the following three properties.

1. Verifying a proof should be easier than running the computation from scratch, and in particular should take time significantly less than  $T$ . Otherwise, the weak device will simply run the computation on its own in the first place.
2. Proving the correctness of a computation should not be “much harder” than running the computation, and in particular should take time at most  $\text{poly}(T)$  (for some polynomial  $\text{poly}$ ). Indeed, if proving requires say an exponential blowup in runtime, then even powerful devices will not be able to prove the correctness of computations.

---

<sup>2</sup>We emphasize that almost all (natural) computations require space at least  $\log T$ , since even holding an index (or pointer) to a location in the computation tableau requires space  $\log T$ .

3. It is impossible to prove the correctness of a false statement (except with very small probability). Otherwise, these proofs will have no meaning. This latter requirement is known as *soundness*.

Unfortunately, achieving these three properties (and even achieving only properties (1) and (3)) simultaneously is widely believed to be impossible. This follows from the fact that  $\text{IP} \subseteq \text{PSPACE}$ , which implies that delegating computations that require large space (say, space proportional to the runtime), is simply impossible! Nevertheless, one can still consider delegating certain limited classes of computations.

Goldwasser et al. [Goldwasser, Kalai, and G. N. Rothblum \[2008\]](#) constructed such a delegation scheme for computations that are computable in “low depth”. Intuitively, low depth computations correspond to computations that are highly parallelizable. In [Goldwasser, Kalai, and G. N. Rothblum \[ibid.\]](#) it was shown how to delegate any time  $T$  and depth  $D$  computation, where the runtime of the prover is proportional to  $D$  (and the instance size), and the runtime of the prover is  $\text{poly}(T)$ . Thus, for low-depth computations, this is a doubly-efficient interactive proof. Moreover, it is quite simple and (almost) efficient enough to use in practice. Indeed, many systems based on the [Goldwasser, Kalai, and G. N. Rothblum \[ibid.\]](#) blueprint were implemented (for example, [Cormode, Mitzenmacher, and Thaler \[2012\]](#), [Thaler, M. Roberts, Mitzenmacher, and Pfister \[2012\]](#), [Thaler \[2013\]](#), [Vu, Setty, Blumberg, and Walfish \[2013\]](#), [Blumberg, Thaler, Vu, and Walfish \[2014\]](#), [Wahby, Howald, S. J. Garg, Shelat, and Walfish \[2016\]](#), [Wahby, Ji, Blumberg, Shelat, Thaler, Walfish, and Wies \[2017\]](#), and [Zhang, Genkin, Katz, Papadopoulos, and Papamanthou \[2017\]](#)), with the goal of using these systems in our day-to-day lives.

The following fundamental problem remains open: Does the  $\text{IP} = \text{PSPACE}$  theorem hold if we restrict the prover in the interactive proof to be efficient? Namely, does every  $T$ -time  $S$ -space computation has an interactive proof where the verifier runs in time proportional to the space  $S$  (and the statement length), and the prover runs in time  $\text{poly}(T)$ ?

Significant progress was recently made by [Reingold, G. N. Rothblum, and R. D. Rothblum \[2016\]](#), who proved that for every constant  $\epsilon > 0$ , every  $T$ -time  $S$ -space computation have an interactive proof where the verifier runs in time proportional to  $S \cdot T^\epsilon$  (and the statement length), and the prover runs in time  $\text{poly}(T)$ . But the fundamental problem above remains a very interesting open problem.

In the rest of this article, we focus on the general problem of delegating *any*  $T$ -time computation. As we saw, in order to achieve this goal we must depart from the interactive proof model, since this proof model is not powerful enough.

**Multi-Prover Interactive Proofs.** The notion of multi-prover interactive proofs was defined by [Ben-Or, Goldwasser, Kilian, and Wigderson \[1988\]](#). This notion, similarly to the interactive proof notion, was defined with a cryptographic goal in mind.

Shortly after Goldwasser et. al. [Goldwasser, Micali, and Rivest \[1988\]](#) introduced the notions of interactive proofs and zero-knowledge proofs, [Goldreich, Micali, and Wigderson \[1987\]](#) showed that every interactive proof can be made zero-knowledge, assuming the existence of a one-way function (a function that is easy to compute but hard to invert). The goal of Ben-Or et. al. [Ben-Or, Goldwasser, Kilian, and Wigderson \[1988\]](#) was to construct an *information theoretic* zero-knowledge proof, without relying on any computational assumptions. This is believed to be impossible in the interactive proof model, which led them to define the *multi-prover interactive proof* (MIP) model.

In this model, the verifier interacts with two (or more) provers. Importantly, it is assumed that these two provers do not communicate during the protocol. Intuitively, this can be enforced by placing the two provers in different rooms (without any connection to the outside world).

Beyond enabling the construction of information theoretic zero-knowledge proofs, this proof model was proven to be extremely powerful. It was proven by [Babai, Fortnow, and Lund \[1991\]](#) that any  $T$ -time computation can be proved to be correct, using a two-prover interactive proof, where the verifier sends a single query to each prover, and each prover replies with an answer. The queries and answers consist of only  $\text{polylog}(T)$  bits, and the runtime of the verifier is  $n \cdot \text{polylog}(T)$ , where  $n$  is the length of the input. Moreover, the runtime of the provers is  $\text{poly}(T)$ , as desired.

In addition, it was shown that the above holds also for *non-deterministic computations*. In other words, it was shown that any proof of length  $T$  can be converted to a 2-prover interactive proof as above where the two queries and two answers are of length  $\text{polylog}(T)$ . In the language of complexity theory, [Babai, Fortnow, and Lund \[ibid.\]](#) proved that  $\text{MIP} = \text{NEXP}$ .<sup>3</sup> Intuitively, the reason this model is so powerful is that it is hard to cheat in a “consistent” manner. Indeed, known 2-prover interactive proof systems consist of a bunch of cross examinations (or consistency checks).

Thus, if we were willing to assume the existence of two non-communicating provers, then we could use these results from the early 90’s to construct a delegation scheme, where the client interacts with two servers, and soundness is ensured as long as these two servers do not interact during the proof process. However, we do not want to make such an assumption, since in many applications (such as for crypto-currencies) this is not a realistic assumption, and for other applications (such as cloud computing) the non-communicating assumption may be too strong, or at the very least simply expensive.

Nevertheless, we show how cryptography can be used to emulate two (or more) non-communicating provers using a single prover.

---

<sup>3</sup>We slightly abuse notation, and throughout this article we denote by MIP the class of all languages that have a multi-prover interactive proof (see [Definition 5](#)), and we also denote by MIP any specific multi-prover interactive proof system.

**Probabilistically Checkable Proofs.** Shortly after this MIP model was introduced, it was noticed that this model is equivalent to the fascinating notion of *probabilistically checkable proofs* (PCP's), which are (non-interactive) proofs that can be verified by reading only a few of their bits. It was observed by Fortnow et. al. [Fortnow, Rompel, and Sipser \[1994\]](#) that any MIP can be trivially converted into a PCP, by writing down for each prover the answers to all the possible queries of the verifier. Since there are known MIP schemes where the length of each query (and each answer) is  $O(\log T)$  the number of possible queries is at most  $\text{poly}(T)$ , and the size of each answer is at most  $O(\log T)$ . Thus, this entire list of queries and answers is of length at most  $\text{poly}(T)$ . Hence, one can verify this proof by running the verifier and sampling a few queries (one for each prover), and reading only the answers corresponding to these queries.

Since this observation, there has been a beautiful line of work (eg., [Feige, Goldwasser, Lovász, Safra, and Szegedy \[1991\]](#), [Babai, Fortnow, Levin, and Szegedy \[1991\]](#), [Arora and Safra \[1992\]](#), and [Arora, Lund, Motwani, Sudan, and Szegedy \[1998\]](#)), culminating with the remarkable PCP theorem that says that any proof of length  $T$  can be converted into a probabilistically checkable one, of length  $\text{poly}(T)$ , where the verifier needs to read only *three* bits of the proof in order to be convinced that the statement is true with constant probability, and this soundness probability can be amplified by repetition. Moreover, to verify the correctness of the proof the verifier only needs to do a single polynomial time (in the statement size) computation, which is independent of the answers, followed by a single  $\text{polylog}(T)$ -time computation.

Probabilistically checkable proofs seem very relevant to the problem of delegating computation, since verifying a PCP can be done very efficiently (reading only a few bits of the proof). However, the length of the PCP is  $\text{poly}(T)$ , and thus even communicating (and storing) this proof is too expensive. If communication and storage were free then indeed PCPs would yield a delegation scheme.

To summarize, despite the beautiful evolution of proofs in computer science starting from the late 80's, it seems that this tremendous progress still does not solve our problem of delegating computation: PCPs require storing a long proof (as long as the computation at hand), multi-prover interactive proofs require assuming two non-communicating provers, and interactive proofs are not general enough to delegate all computations (only bounded space computations). Moreover, as we mentioned, constructing a doubly-efficient interactive proofs for all bounded space computations remains an open problem. Finally, we mention that interactive proofs require many rounds of interaction between the prover and the verifier, and one of the major goals of delegating computation is to obtain non-interactive solutions.

Therefore, in the context of delegating computation, this line of work suffers from significant limitations. Somewhat surprisingly, it has been shown that cryptography can be used to remove many of these limitations.

**1.2 Cryptography to the Rescue.** It turns out that cryptography can be used to convert any PCP or MIP scheme into a delegation scheme. At first, the use of cryptography may seem quite surprising, since the problem at hand does not seem related to cryptography in any way, since we are not concerned with privacy, only in proving correctness. Nevertheless, we show how using cryptography one can shrink a long PCP into a short one, and how one can simulate a multi-prover interactive proof via a single prover. To this end, we need to relax the soundness condition, to *computational soundness*.

**Computational Soundness.** Rather than requiring that it is impossible to prove the validity of a false statement, we require that it is “*practically impossible*” to prove the validity of a false statement. More specifically, we require that it is impossible to prove a false statement only for *computationally-bounded* (e.g., polynomial time) cheating provers. Yet, a computationally all powerful cheating prover may be able to cheat. Honest provers are also required to be efficient (i.e., computationally bounded), in keeping with the philosophy that security should hold against adversaries who are at least as powerful as honest parties. Such proof systems are also known in the literature as *argument systems* Brassard, Chaum, and Crépeau [1988] or *computationally sound proofs* Micali [1994] (as opposed to *statistically sound* proofs that ensure that even a computationally unbounded cheating prover cannot convince a verifier to accept a false statement).

Typically, computational soundness relies on a computational hardness assumption, such as the assumption that it is hard to factor large composite numbers (known as the Factoring Assumption). In this case the soundness guarantee is that if a cheating prover can convince the verifier to accept a false statement (with high probability), then this prover can be used to break the Factoring Assumption. Most of the work in the literature on delegating computation, considers the setting of computational soundness, where we require soundness to hold only against cheating provers who cannot break some underlying cryptographic assumption (such as the Factoring Assumption).

Very loosely speaking, the literature on computation delegation can be partitioned into three categories. The first constructs delegation schemes from any PCP scheme by using the notion of *collision resistant hash functions* to “shrink” the long PCP. The second constructs delegation schemes from any MIP scheme by using cryptography to emulate the many (non-communicating) provers using a single prover. The third uses the notion of *obfuscation* to construct a delegation scheme directly (without using the beautiful evolution

of proofs in computer science, summarized above). In this article we focus on the second category. In what follows, we slightly elaborate on the line of work in the first category, and due to lack of space, we do not elaborate on the works in the third category.

**Delegation from PCP schemes.** Kilian [1992] showed how to use a *collision resistant hash function* to convert any PCP scheme into a 4-message delegation scheme for any deterministic (or even non-deterministic) computation.

Many of the applications where a delegation scheme is used (such as in crypto-currencies) require the proof to be *non-interactive*. A non-interactive delegation scheme consists of public parameters (generated honestly by the verifier). These public parameters are used to generate proofs that consist of a *single* message, and soundness holds even if a (cheating) prover chooses the statement to be proved as a function of the public parameters.

Micali [1994] proved that a similar approach to the one by Kilian, yields a non-interactive delegation scheme in the so called “Random Oracle Model” Bellare and Rogaway [1993]. Specifically, his scheme uses a hash function, and security is proven assuming the adversary only makes black-box use of this hash function. However, the Random Oracle Model is known to be insecure in general, and there are examples of schemes that are secure in the Random Oracle Model, yet are known to be insecure when the random oracle is replaced with any (succinct) hash function Canetti, Goldreich, and Halevi [2004], Barak [2001], and Goldwasser and Kalai [2003].

Since this seminal work of Micali, there has been a long line of followup works (eg., Groth [2010], Lipmaa [2012], Damgård, Faust, and Hazay [2012], Gennaro, Gentry, Parno, and Raykova [2013], Bitansky, Chiesa, Ishai, Ostrovsky, and Paneth [2013], Bitansky, Canetti, Chiesa, and Tromer [2013], and Bitansky, Canetti, Chiesa, Goldwasser, H. Lin, Rubinstein, and Tromer [2014]), constructing a delegation scheme without resorting to the Random Oracle Model. However, these delegation schemes were proven secure under very strong and non-standard “knowledge assumptions”. Knowledge assumptions are different from standard complexity assumptions, and (similarly to the Random Oracle Model) they restrict the class of adversaries considered to those which compute things in a certain way.<sup>4</sup>

**Our focus.** In this article we focus on the second line of work, which constructs a non-interactive delegation scheme based on a *standard* cryptographic assumption. This line of work is based on a curious connection, noted in Kalai, Raz, and R. D. Rothblum [2013, 2014], between the problem of delegating computation and the concept of *no-signaling strategies* from quantum physics.

---

<sup>4</sup>For example, the Knowledge-of-Exponent assumption Damgård [1992] assumes that any adversary that given  $(g, h)$  computes  $(g^z, h^z)$ , must do so by “first” computing  $z$  and then computing  $(g^z, h^z)$ .

The starting point is an elegant method introduced by Biehl et. al. [Biehl, Meyer, and Wetzel \[1999\]](#), for converting any MIP into a 1-round delegation scheme. In what follows, for the sake of simplicity, we describe this method using a *fully homomorphic encryption scheme*, though weaker primitives (such a computational private retrieval scheme) are known to suffice. A fully homomorphic encryption scheme is a secure encryption scheme that allows to do computations on encrypted data (we refer the reader to [Section 2.3](#), and to [Definition 8](#) for the precise definition). Starting from the breakthrough work of [Gentry \[2009\]](#) and of [Brakerski and Vaikuntanathan \[2011\]](#), such homomorphic encryption schemes were constructed based on the Learning with Error Assumption, which is a standard and well established cryptographic assumption.

**The Biehl, Meyer, and Wetzel [1999] method.** Loosely speaking, the [Biehl, Meyer, and Wetzel \[ibid.\]](#) method takes any MIP scheme and converts it into the following 1-round delegation scheme: The verifier of the delegation scheme computes all the queries for the MIP provers, and sends all these queries to a (single) prover, each encrypted using a different (freshly generated) key corresponding to an FHE scheme. The prover who receives all these encrypted queries, computes for each of the MIP provers its response homomorphically, underneath the layer of the FHE encryption.

This method was considered to be a heuristic, since no proof of soundness was given. The intuition for why this heuristic was believed to be sound is that when a cheating prover answers each of the queries, the other queries are encrypted using different (independently generated) keys, and hence these other queries are completely hidden. Surprisingly, despite this intuition, Dwork et. al. [Dwork, Langberg, Naor, Nissim, and Reingold \[2004\]](#) and Dodis et. al. [Dodis, Halevi, R. D. Rothblum, and Wichs \[2016\]](#) showed that this heuristic, in general, is insecure. Intuitively the reason is that the soundness of the MIP is ensured only against cheating provers that answer each query *locally*, only as a function of the corresponding query. In this delegation scheme a cheating prover is not restricted to use local strategies. Rather the security of the FHE scheme ensures that each answer (provided by a cheating prover) does not “signal” information about the other queries, since if it did then we could use this prover to break the security of the FHE scheme.

However, there are strategies that are neither signaling nor local. Such strategies are known in the quantum literature as *no-signaling* strategies (and are formally defined in [Section 3.2](#)). The intuition above suggests that these no-signaling strategies are useless. However, in the quantum literature it is well known that this is not necessarily the case.

In a series of work, starting from [Kalai, Raz, and R. D. Rothblum \[2013, 2014\]](#), it was proven that if the underlying MIP is sound against (statistically) no-signaling strategies, then the delegation scheme resulting from the [Biehl, Meyer, and Wetzel \[1999\]](#) heuristic



is sound. Moreover, these works constructed for any  $T$ -time (deterministic) computation an MIP with (statistical) no-signaling soundness, with communication complexity  $\text{polylog}(T)$ , and where the runtime of the verifier is  $n \cdot \text{polylog}(T)$ . This led to the first 1-round delegation scheme for arbitrary (deterministic) computations based on standard cryptographic assumptions. Moreover, these works were later generalized to include RAM computations [Kalai and Paneth \[2015\]](#), non-adaptive delegation (i.e., 1-round delegation with adaptive soundness) [Brakerski, Holmgren, and Kalai \[2017\]](#), and even generalized to non-deterministic space-bounded computations.

As opposed to the previous line of work, where anyone can verify the proof since all that is needed for verification is the public parameters and the proof, in this line of work the proofs are *privately verifiable*, meaning that in order to verify the proof one needs to know a “secret state” generated together with the public parameters.<sup>5</sup>

In a very recent work, [Paneth and G. N. Rothblum \[2017\]](#) provide a blue-print that generalizes the approach taken in this line of work, to obtain publicly verifiable delegation schemes. However, currently we do not know how to realize this blue-print based on standard cryptographic assumptions.

We mention that the third line of work, that constructs delegation schemes based on obfuscation (e.g., [Canetti, Holmgren, Jain, and Vaikuntanathan \[2015\]](#), [Koppula, Lewko, and Waters \[2015\]](#), [Bitansky, S. Garg, H. Lin, Pass, and Telang \[2015\]](#), [Canetti and Holmgren \[2016\]](#), [Ananth, Chen, Chung, H. Lin, and W. Lin \[2016\]](#), and [Chen, Chow, Chung, Lai, W. Lin, and Zhou \[2016\]](#)), achieve public verifiable delegation schemes for deterministic computations. However, known constructions of obfuscation are built on shaky grounds, and are not known to be secure based on standard assumptions.<sup>6</sup>

The question of constructing a *publicly verifiable* 1-round delegation scheme for general computations under standard assumptions remains a fascinating open question. In addition, the question of constructing a 1-round delegation scheme for general *non-deterministic* computations (beyond space-bounded computations) under standard assumptions (and even under obfuscation type assumptions) remains a fascinating open question.

## 2 Preliminaries

We model efficient algorithms as probabilistic polynomial time (PPT) algorithms, formally modeled as Turing machines. We denote by  $\text{DTIME}(T)$  the class of all the languages that can be computed by a *deterministic* Turing machine that on input  $x$  runs in time  $T(|x|)$  (i.e., terminates within  $T(|x|)$  steps). We denote by  $\text{NTIME}(T)$  the class of all the languages

<sup>5</sup>Indeed, the secret keys of the FHE scheme are needed in order to decrypt the answers and verify correctness.

<sup>6</sup>We mention that these schemes are also not non-interactive, in the sense that soundness holds only if the false statement does not depend on the public parameters.

that can be computed by a *non-deterministic* Turing machine that on input  $x$  runs in time  $T(|x|)$ .

Throughout this article we use  $\lambda$  to denote the security parameter. This value determines the security level of our schemes. Taking larger values of  $\lambda$  results with better security, though the prover(s) and verifier run in time polynomial in  $\lambda$ , and thus the efficiency of the scheme degrades as we increase  $\lambda$ . The prover(s) and the verifier take as input  $1^\lambda$ , and the reason we give  $\lambda$  in unary is since we allow our algorithms to run in polynomial time, and we want to allow them to run in time polynomial in  $\lambda$ .

**Definition 1.** A function  $v : \mathbb{N} \rightarrow \mathbb{N}$  is said to be negligible if for every polynomial  $p : \mathbb{N} \rightarrow \mathbb{N}$ , there exists a constant  $c > 0$  such that for every  $\lambda > c$  it holds that  $v(\lambda) \leq \frac{1}{p(\lambda)}$ .

For a distribution  $\mathcal{Q}$ , we denote by  $a \leftarrow \mathcal{Q}$  a random variable distributed according to  $\mathcal{Q}$  (independently of all other random variables).

**Definition 2.** Two distribution ensembles  $\{\mathcal{X}_\lambda\}_{\lambda \in \mathbb{N}}$  and  $\{\mathcal{Y}_\lambda\}_{\lambda \in \mathbb{N}}$  are said to be computationally indistinguishable if for every PPT distinguisher  $\mathcal{D}$ ,

$$\left| \Pr_{x \leftarrow \mathcal{X}_\lambda} [\mathcal{D}(x) = 1] - \Pr_{y \leftarrow \mathcal{Y}_\lambda} [\mathcal{D}(y) = 1] \right| = \text{negl}(\lambda).$$

They are said to be statistically indistinguishable if the above holds for every (even computationally unbounded) distinguisher  $\mathcal{D}$ .

**2.1 Delegation Schemes.** In what follows, we define the notion of a 1-round delegation scheme and a non-interactive delegation scheme. We require that the first message sent by the verifier does not depend on the statement to be proven. In the literature, this is often not explicitly required, and we add this requirement to the definition since our constructions achieve this desirable property. We define delegation schemes for non-deterministic languages, though we emphasize that this includes also deterministic languages, since any deterministic computation can be thought of as a non-deterministic one where the non-deterministic advice is empty.

**Definition 3.** Fix any  $T : \mathbb{N} \rightarrow \mathbb{N}$  and any  $L \in \text{NTIME}(T)$ . A 1-round delegation scheme  $(P, V)$  for the language  $L$ , has the following properties.

1. **Structure:** The algorithm  $V$  can be partitioned into two PPT algorithms  $V = (V_1, V_2)$ , where  $V_1$  is a PPT algorithm that generates parameters  $(\text{pp}, \text{st}) \leftarrow V(1^\lambda)$ . To prove that  $x \in L$ , upon receiving  $\text{pp}$  and  $x$ , the prover  $P$  runs in time  $\text{poly}(\lambda, T(|x|))$  and computes  $\text{pf} \leftarrow P(x, \text{pp})$ . The algorithm  $V_2$  takes as input  $(x, \text{pf}, \text{st})$  and outputs a bit, indicating whether he accepts or rejects the proof  $\text{pf}$  with respect to the public parameters corresponding to his secret state  $\text{st}$ .

2. **Completeness:** For every security parameter  $1^\lambda$ , and every  $x \in L$  such that  $|x| \leq 2^\lambda$ ,

$$\Pr[V_2(x, \text{pf}, \text{st}) = 1] = 1$$

where the probability is over  $(\text{pp}, \text{st}) \leftarrow V_1(1^\lambda)$  and over  $\text{pf} \leftarrow P(x, \text{pp})$ .

3. **Soundness:** For every PPT (cheating) prover  $P^* = (P_1^*, P_2^*)$ ,

$$\Pr[V_2(x, \text{pf}, \text{st}) = 1 \wedge (x \notin L)] = \text{negl}(\lambda)$$

where the probability is over  $x \leftarrow P_1^*(1^\lambda)$ , over  $(\text{pp}, \text{st}) \leftarrow V_1(1^\lambda)$  and over  $\text{pf} \leftarrow P_2^*(x, \text{pp})$ .

4. **Efficiency:** The communication complexity is  $\text{poly}(\lambda, \log T(|x|))$ . The honest verifier runs in time  $|x| \cdot \text{polylog}(T(|x|)) + \text{poly}(\lambda, \log T(|x|))$ , and the honest prover runs in time  $\text{poly}(\lambda, T(|x|))$  (given non-deterministic advice for  $x \in L$ ).

**Definition 4.** A non-interactive delegation scheme for a language  $L \in \text{NTIME}(T)$ , has the same properties as a 1-round delegation scheme except that the soundness condition is replaced with the following adaptive soundness condition:

**Adaptive Soundness:** For every PPT (cheating) prover  $P^*$ ,

$$\Pr[V_2(x, \text{pf}, \text{st}) = 1 \wedge (x \notin L)] = \text{negl}(\lambda)$$

where the probability is over  $(\text{pp}, \text{st}) \leftarrow V_1(1^\lambda)$  and over  $(x, \text{pf}) \leftarrow P^*(\text{pp})$ .

**2.2 Multi-Prover Interactive Proofs.** In what follows, we define the notion of a multi-prover interactive proof (MIP). Let  $L$  be a language. In a 1-round  $k$ -prover interactive proof,  $k = k(\lambda)$  provers,  $P_1, \dots, P_k$ , try to convince a (probabilistic) verifier  $V$ , that  $x \in L$ . The input  $x$  is known to all parties.

In the traditional works on MIP, it was required that the verifier's runtime on input  $(1^\lambda, x)$  is at most  $\text{poly}(|x|, \lambda)$  and the honest provers' runtime could be unbounded.<sup>7</sup> We change these efficiency requirements to align with the requirements of a delegation scheme. In particular, we require that if  $L \in \text{NTIME}(T)$  then the runtime of the verifier is at most  $|x| \cdot \text{polylog}(T(|x|)) + \text{poly}(\lambda, \log T(|x|))$  and the runtime of the (honest) provers is at most  $\text{poly}(\lambda, T(|x|))$  (assuming they are given the non-deterministic advice for  $x \in L$ ).

The proof consists of only one round. Given a security parameter  $1^\lambda$  (which determines the soundness), and a random string, the verifier generates  $k = k(\lambda)$  queries,  $q_1, \dots, q_k$ , one for each prover, and sends them to the  $k$  provers. Each prover responds with an

<sup>7</sup>To be precise, the traditional definition does not even include a security parameter. The verifier is required to run in time  $\text{poly}(|x|)$  and soundness is required to hold with constant probability (say  $1/2$ ).

answer that depends only on its own individual query. That is, the provers on input  $x$  (and associated non-deterministic advice) respond with answers  $a_1, \dots, a_k$ , where for every  $i$  we have  $a_i \leftarrow P_i(x, q_i)$ . Finally, the verifier decides whether to accept or reject based on the answers that it receives (as well as the input  $x$  and the random string).

**Definition 5.** Fix any  $T : \mathbb{N} \rightarrow \mathbb{N}$  and any  $L \in \text{NTIME}(T)$ . We say that  $(V, P_1, \dots, P_k)$  is a one-round  $k$ -prover interactive proof system (MIP) for  $L$  if the following properties are satisfied:

1. **Structure:** The verifier consists of two PPT algorithms,  $V = (V_1, V_2)$ , where  $(q_1, \dots, q_k, \text{st}) \leftarrow V_1(1^\lambda)$ .

Namely, the queries do not depend on the statement proven.

2. **Completeness:** For every security parameter  $1^\lambda$ , and every  $x \in L$  such that  $|x| \leq 2^\lambda$ ,

$$\Pr[V_2(x, q_1, \dots, q_k, a_1, \dots, a_k, \text{st}) = 1] = 1 - \text{negl}(\lambda),$$

where the probability is over  $(q_1, \dots, q_k, \text{st}) \leftarrow V_1(1^\lambda)$  and over  $a_i \leftarrow P_i(x, q_i)$  for every  $i \in [k]$ .

3. **Soundness:** For every  $\lambda \in \mathbb{N}$ , every  $x \notin L$  (whose size may depend on  $\lambda$ ), and any (computationally unbounded, possibly cheating) provers  $P_1^*, \dots, P_k^*$ ,

$$\Pr[V_2(x, q_1, \dots, q_k, a_1, \dots, a_k, \text{st}) = 1] = \text{negl}(\lambda),$$

where the probability is over  $(q_1, \dots, q_k, \text{st}) \leftarrow V_1(1^\lambda)$  and over  $a_i \leftarrow P_i^*(x, q_i)$  for every  $i \in [k]$ .

4. **Efficiency:** The communication complexity is  $\text{poly}(\lambda, \log T)$ . The verifier runs in time  $|x| \cdot \text{polylog}(T(|x|)) + \text{poly}(\lambda, \log T(|x|))$ , and the prover runs in time  $\text{poly}(\lambda, T(|x|))$  (assuming he has non-deterministic advice for  $x \in L$ ).

**Theorem 1.** *Babai, Fortnow, and Lund [1991]* For any  $T : \mathbb{N} \rightarrow \mathbb{N}$  and any language  $L \in \text{NTIME}(T)$ , there exists a 2-prover interactive proof  $(V, P_1, P_2)$  for  $L$  satisfying Definition 5.

The holy grail of the area of computation delegation, is to achieve the guarantees of Theorem 1 with a single prover. Unfortunately, as we mentioned, this dream is too good to be true, since the  $\text{IP} = \text{PSPACE}$  theorem says that a single prover can only prove the correctness of bounded space computations. Moreover, known interactive proofs for PSPACE require many rounds, and the class of languages that can be proved via a 1-round interactive proof is widely believed to be quite limited.

In [Section 3.1](#), we present a method first proposed by Biehl et. al. [Biehl, Meyer, and Wetzel \[1999\]](#), that converts any MIP scheme into a single prover delegation scheme, using the aid of cryptography, and in particular using a computational private information retrieval (PIR) scheme. In this article, for the sake of simplicity, we present this method using a fully homomorphic encryption (FHE) scheme, which is a stronger assumption than a PIR scheme. We chose to present this method using an FHE scheme (as opposed to a PIR scheme) only because we find the terminology to be simpler. We emphasize that all the results presented from now on hold with a PIR scheme as well.

**2.3 Fully Homomorphic Encryption (FHE).** We start by defining a *public-key encryption* scheme. Such a scheme consists of three probabilistic polynomial-time algorithms  $(\text{Gen}, \text{Enc}, \text{Dec})$ , and is defined over some message space  $\mathbb{M}$ . The key generation algorithm  $\text{Gen}$ , when given as input a security parameter  $1^\lambda$ , outputs a pair  $(\text{pk}, \text{sk})$  of public and secret keys. The encryption algorithm,  $\text{Enc}$ , on input a public key  $\text{pk}$ , and a message  $m \in \mathbb{M}$ , outputs a ciphertext  $\hat{m}$ , and the decryption algorithm,  $\text{Dec}$ , when given the ciphertext  $\hat{m}$  and the secret key  $\text{sk}$ , outputs the original message  $m$  (with overwhelming probability).

**Definition 6.** *A public key encryption over a message space  $\mathbb{M}$  consists of three PPT algorithms  $(\text{Gen}, \text{Enc}, \text{Dec})$  such that for every  $m \in \mathbb{M}$ ,*

$$\Pr[\text{Dec}(\hat{m}, \text{sk}) = m] = 1 - \text{negl}(\lambda),$$

where the probability is over  $(\text{pk}, \text{sk}) \leftarrow \text{Gen}(1^\lambda)$ , and over  $\hat{m} \leftarrow \text{Enc}(m, \text{pk})$ .

**Definition 7.** *Goldwasser and Micali [1984] A public-key encryption scheme  $(\text{Gen}, \text{Enc}, \text{Dec})$  is (semantically) secure if for every PPT algorithm  $\mathcal{G}$ , for every  $\lambda \in \mathbb{N}$  and for every two messages  $m_1, m_2 \in \mathbb{M}$  such that  $|m_1| = |m_2|$ ,*

$$|\Pr[\mathcal{G}(\text{pk}, \hat{m}_1) = 1] - \Pr[\mathcal{G}(\text{pk}, \hat{m}_2) = 1]| = \text{negl}(\lambda)$$

where the probabilities are over  $(\text{pk}, \text{sk}) \leftarrow \text{Gen}(1^\lambda)$ , over  $\hat{m}_1 \leftarrow \text{Enc}(m_1, \text{pk})$ , and over  $\hat{m}_2 \leftarrow \text{Enc}(m_2, \text{pk})$ .

**Definition 8.** *A tuple of PPT algorithms  $(\text{Gen}, \text{Enc}, \text{Dec}, \text{Eval})$  is a fully-homomorphic encryption scheme over the message space  $\{0, 1\}^*$  if  $(\text{Gen}, \text{Enc}, \text{Dec})$  is a public-key encryption scheme over the message space  $\{0, 1\}^*$ , and in addition the following condition holds:*

**Homomorphic Evaluation:**  $\text{Eval}$  takes as input a public key  $\text{pk}$ , a circuit  $C : \{0, 1\}^k \rightarrow \{0, 1\}^\ell$ , where  $k, \ell \leq \text{poly}(\lambda)$ , and a ciphertext  $\hat{m}$  that is an encryption of a message  $m \in$

$\{0, 1\}^k$  with respect to  $\text{pk}$ , and outputs a string  $\psi$  such that for every  $C : \{0, 1\}^k \rightarrow \{0, 1\}^\ell$ , where  $k, \ell \leq \text{poly}(\lambda)$ , and every  $m \in \{0, 1\}^k$ ,

$$\Pr[\text{Dec}(\psi, \text{sk}) = C(m)] = 1 - \text{negl}(\lambda),$$

where the probability is over  $(\text{pk}, \text{sk}) \leftarrow \text{Gen}(1^\lambda)$ , over  $\hat{m} \leftarrow \text{Enc}(m, \text{pk})$ , and over  $\psi = \text{Eval}(\text{pk}, C, \hat{m})$ .

Moreover, the length of  $\psi$  is polynomial in  $\lambda$  and  $\ell$  (and is independent of the size of  $C$ ).

Starting from the breakthrough work of [Gentry \[2009\]](#), and of [Brakerski and Vaikuntanathan \[2011\]](#), such homomorphic encryption schemes were constructed based on the standard Learning with Error Assumption [Regev \[2003\]](#). The message space in these constructions is  $\mathfrak{M} = \{0, 1\}$ , though one can use these schemes to encrypt any message in  $\{0, 1\}^*$  by encrypting the message in a bit-by-bit manner.

### 3 From MIP to Non-Interactive Delegation

**Notation.** Throughout this section we denote by  $k = k(\lambda)$  the number of provers in the MIP scheme. For a vector  $a = (a_1, \dots, a_k)$  and a subset  $S \subseteq [k]$ , we denote by  $a_S$  the sequence of elements of  $a$  that are indexed by indices in  $S$ , that is,  $a_S = (a_i)_{i \in S}$ .

**3.1 The Biehl, Meyer, and Wetzel [1999] Heuristic.** Biehl et. al. [Biehl, Meyer, and Wetzel \[ibid.\]](#) suggested a heuristic for converting any MIP into a 1-round delegation scheme, by using a computational private information retrieval (PIR) scheme. As mentioned above, we present this heuristic using a fully homomorphic encryption (FHE) scheme (see [Definition 8](#)).

The Biehl et. al. heuristic is natural and elegant. Loosely speaking, the idea is the following: The verifier of the delegation scheme computes all the queries for the MIP provers, and sends all these queries to the (single) prover, each encrypted using an FHE scheme, where each query is encrypted with its own (freshly generated) key. The prover then computes for each of the MIP provers its response homomorphically, underneath the layer of the FHE encryption.

In what follows we give a formal description of the Biehl et. al. heuristic.

**The Biehl, Meyer, and Wetzel [ibid.] Heuristic.** Fix any language  $L$ , an MIP scheme  $(V, P_1, \dots, P_k)$  for  $L$ , and an FHE scheme  $(\text{Gen}, \text{Enc}, \text{Dec}, \text{Eval})$ . Consider the following 1-round delegation scheme  $(P^{\text{del}}, V^{\text{del}})$ , where  $V^{\text{del}} = (V_1^{\text{del}}, V_2^{\text{del}})$ :

- The PPT algorithm  $V_1^{\text{del}}$  takes as input the security parameter  $1^\lambda$ , and does the following:
  1. Compute  $(q_1, \dots, q_k, \text{st}) \leftarrow V_1(1^\lambda)$ .
  2. Run  $\text{Gen}(1^\lambda)$  independently  $k$  times to generate  $\{(\text{pk}_i, \text{sk}_i)\}_{i \in [k]}$ .
  3. For every  $i \in [k]$  compute  $\hat{q}_i \leftarrow \text{Enc}(q_i, \text{pk}_i)$ .
  4. Set  $\text{pp}^{\text{del}} = (\hat{q}_1, \dots, \hat{q}_k)$  and set  $\text{st}^{\text{del}} = (\text{sk}_1, \dots, \text{sk}_k, q_1, \dots, q_k, \text{st})$ .
- The prover  $P^{\text{del}}(x, \text{pp}^{\text{del}})$  does the following:
  1. Parse  $\text{pp}^{\text{del}} = (\hat{q}_1, \dots, \hat{q}_k)$ .
  2. For every  $i \in [k]$  compute  $\hat{a}_i \leftarrow \text{Eval}(P_i(x, \cdot), \hat{q}_i)$ .
  3. Send  $(\hat{a}_1, \dots, \hat{a}_k)$  to the verifier.
- Upon receiving  $x$  and  $(\hat{a}_1, \dots, \hat{a}_k)$ , the verifier  $V_2^{\text{del}}(x, \hat{a}_1, \dots, \hat{a}_k, \text{st}^{\text{del}})$  does the following:
  1. Parse  $\text{st}^{\text{del}} = (\text{sk}_1, \dots, \text{sk}_k, q_1, \dots, q_k, \text{st})$ .
  2. For each  $i \in [k]$  compute  $a_i \leftarrow \text{Dec}(\hat{a}_i, \text{sk}_i)$ .
  3. Accept if and only if  $V_2(x, q_1, \dots, q_k, a_1, \dots, a_k, \text{st}) = 1$ .

This is a beautiful and natural heuristic. it is easy to see that it satisfies the efficiency and completeness properties of a delegation scheme. The main question is:

*Is this Heuristic Sound?*

The intuition for why this heuristic was believed to be sound is the following: When a cheating prover answers each of the queries, the other queries are encrypted using different (independently generated) keys, and hence are indistinguishable from encryptions of 0. Therefore, each answer should be indistinguishable from the answer the cheating prover would have provided in the case where the other queries were all 0, and clearly having encryptions of 0 cannot help a prover cheat, since he can generate these encryptions on his own.

Surprisingly, despite this intuition, Dwork et. al. [Dwork, Langberg, Naor, Nissim, and Reingold \[2004\]](#) showed that this heuristic, in general, can be insecure. The reason is that the soundness of the MIP is ensured only against cheating provers that answer each query *locally*, only as a function of the corresponding query. In this delegation scheme a cheating prover is not restricted to use local strategies. Rather the security of the FHE scheme ensures that each answer (provided by a cheating prover) does not “signal” information

about the other queries, since if it did then we could use this prover to break the security of the FHE scheme.

However, there are strategies that are neither signaling nor local. Dwork et. al. [Dwork, Langberg, Naor, Nissim, and Reingold *ibid.*] refer to such strategies as “spooky interactions”. Such strategies are known in the quantum literature as *no-signaling* strategies (defined formally in Section 3.2, below). The intuition above suggests that these no-signaling strategies are useless. However, in the quantum literature it is well known that this is not the case.

Very recently, Dodis, Halevi, R. D. Rothblum, and Wichs [2016] showed that indeed the Biehl, Meyer, and Wetzel [1999] heuristic is insecure! Specifically, they construct an MIP scheme and a FHE scheme, for which when applying the Biehl, Meyer, and Wetzel *ibid.* heuristic to these MIP and FHE schemes, the resulting delegation scheme is not sound. To this end, they construct an MIP scheme whose soundness can be broken via a no-signaling strategy, and this no-signaling strategy can be implemented under the layer of the FHE.

**3.2 MIPs with No-Signaling Provers.** The works of Kalai, Raz, and R. D. Rothblum [2013, 2014] attempt to prove the soundness of the Biehl, Meyer, and Wetzel [1999] heuristic, by considering a variant of the MIP model, where the cheating provers are more powerful.

In the standard MIP model, each prover answers his own query locally, without knowing the queries that were sent to the other provers. The no-signaling model allows each answer to depend on all the queries, as long as for any subset  $S \subset [k]$ , and any queries  $q_S$  for the provers in  $S$ , the distribution of the answers  $a_S$ , conditioned on the queries  $q_S$ , is independent of all the other queries.

Intuitively, this means that the answers  $a_S$  do not give the provers in  $S$  information about the queries of the provers outside  $S$ , except for information that they already have by seeing the queries  $q_S$ .

Formally, denote by  $D$  the alphabet of the queries and denote by  $\Sigma$  the alphabet of the answers. For every  $q = (q_1, \dots, q_k) \in D^k$ , let  $\mathcal{Q}_q$  be a distribution over  $\Sigma^k$ . We think of  $\mathcal{Q}_q$  as the (joint) distribution of the answers for queries  $q$ .

**Definition 9.** *We say that the family of distributions  $\{\mathcal{Q}_q\}_{q \in D^k}$  is no-signaling if for every subset  $S \subset [k]$  and every two sequences of queries  $q, q' \in D^k$ , such that  $q_S = q'_S$ , the following two random variables are identically distributed:*

- $a_S$ , where  $a \leftarrow \mathcal{Q}_q$
- $a'_S$  where  $a' \leftarrow \mathcal{Q}_{q'}$



If the two distributions are computationally (resp. statistically) indistinguishable (see [Definition 2](#)), rather than identical, we say that the family of distributions  $\{\mathcal{Q}_q\}_{q \in D^k}$  is computationally (resp. statistically) no-signaling.

**Definition 10.** An MIP  $(V, P_1, \dots, P_k)$  for a language  $L$  is said to be sound against no-signaling strategies (or provers) if the following (more general) soundness property is satisfied:

**Ni-Signaling Soundness:** For every  $\lambda \in \mathbb{N}$ , every  $x \notin L$ , and any no-signaling family of distributions  $\{\mathcal{Q}_q\}_{q \in D^k}$ ,

$$\Pr[V_2(x, q_1, \dots, q_k, a_1, \dots, a_k, \text{st}) = 1] = \text{negl}(\lambda)$$

where the probability is over  $(q_1, \dots, q_k, \text{st}) \leftarrow V_1(1^\lambda)$  and over  $(a_1, \dots, a_k) \leftarrow \mathcal{Q}_{(q_1, \dots, q_k)}$ .

If this property is satisfied for any computationally (resp. statistically) no-signaling family of distributions  $\{\mathcal{Q}_q\}_{q \in D^k}$ , we say that the MIP has soundness against computationally (resp. statistically) no-signaling strategies.

No-signaling strategies were first studied in physics in the context of Bell inequalities by [Khalfin and Tsirelson \[1985\]](#) and [Rastall \[1985\]](#), and they gained much attention after they were reintroduced by [Popescu and Rohrlich \[1994\]](#). MIPs that are sound against no-signaling provers were extensively studied in the literature (see for example [Toner \[2009\]](#), [Barrett, Linden, Massar, Pironio, Popescu, and D. Roberts \[2005\]](#), [Avis, Imai, and Ito \[2006\]](#), [Kempe, Kobayashi, Matsumoto, Toner, and Vidick \[2008\]](#), [Ito, Kobayashi, and Matsumoto \[2009\]](#), [Holenstein \[2009\]](#), and [Ito \[2010\]](#)). We denote the class of MIP's that are sound against no-signaling provers by  $\text{MIP}^{\text{NS}}$ .

The study of MIPs that are sound against no-signaling provers was originally motivated by the study of MIPs with provers that share entangled quantum states. No-signaling provers are allowed to use arbitrary strategies, as long as their strategies cannot be used for communication between any two disjoint sets of provers. By the physical principle that information cannot travel faster than light, a consequence of Einstein's special relativity theory, it follows that if the provers are placed far enough apart, then the only strategies that can be realized by these provers, even if they share entangled quantum states, are no-signaling strategies.

Moreover, the principle that information cannot travel faster than light is a central principle in physics, and is likely to remain valid in any future ultimate theory of nature, since its violation means that information could be sent from future to past. Therefore, soundness against no-signaling strategies is likely to ensure soundness against provers that obey a future ultimate theory of physics, and not only the current physical theories that we have, that are known to be incomplete.

The study of MIPs that are sound against no-signaling provers is very appealing also because no-signaling strategies have a simple mathematical characterization.

Ito et al. [Ito, Kobayashi, and Matsumoto 2009] proved that the set of languages in  $\text{MIP}^{\text{NS}}$  contains PSPACE and is contained in EXP. We emphasize that they use the traditional MIP definition, which allows the honest provers to be computationally unbounded, and indeed in their  $\text{MIP}^{\text{NS}}$  for PSPACE the provers run in super-polynomial time. Moreover, they assume the verifier runs in time at most  $\text{poly}(|x|)$  (which is the traditional requirement).<sup>8</sup> We note that if they used our efficiency requirement where the verifier is allowed to run in time  $|x| \cdot \text{polylog}(T) + \text{poly}(\lambda, \log T)$ , and the communication complexity is at most  $\text{poly}(\lambda, \log T)$ , they would get that each  $\text{MIP}^{\text{NS}}$  is contained in  $\text{DTIME}(T)$ .

For the case of *two* provers, Ito [2010] showed that the corresponding complexity class is contained in (and therefore equal to) PSPACE. This is in contrast to the class MIP (with soundness against local strategies), which is known to be equal to NEXP.

The connection between MIPs with no-signaling soundness and computation delegation was first observed in Kalai, Raz, and R. D. Rothblum [2013]. Loosely speaking, they prove that the Biehl, Meyer, and Wetzel [1999] heuristic is sound when applied to any MIP that is secure against *statistically* no-signaling strategies, denoted by  $\text{MIP}^{\text{sNS}}$ .<sup>9</sup> In Kalai, Raz, and R. D. Rothblum [2013, 2014] they also characterize the exact power of MIPs that are secure against statistically no-signaling provers, and prove that  $\text{MIP}^{\text{sNS}} = \text{EXP}$ . More specifically, they prove the following theorem.

**Theorem 2.** *Kalai, Raz, and R. D. Rothblum [2013, 2014] For any  $T : \mathbb{N} \rightarrow \mathbb{N}$ , and any language in  $L \in \text{DTIME}(T)$ , there exists an MIP with statistical no-signaling soundness (as in defined in Definitions 5 and 10).*

In particular, these works prove the following theorem.

**Theorem 3.** *Kalai, Raz, and R. D. Rothblum [2013, 2014] For any  $T : \mathbb{N} \rightarrow \mathbb{N}$  and any  $L \in \text{DTIME}(T)$  there exists a 1-round delegation scheme for  $L$  (as defined in Definition 3), assuming the existence of an FHE scheme that is secure against quasi-polynomial time adversaries.*

To achieve non-interactive delegation (as opposed to 1-round delegation) we need to use an MIP scheme that is sound against *adaptive* no-signaling strategies, as defined in Brakerski, Holmgren, and Kalai [2017].

<sup>8</sup>They show that one can find the best no-signaling strategy for the provers by solving an exponential (in  $|x|$ ) size linear program.

<sup>9</sup>Their result relies on the stronger assumption that the underlying FHE is not only secure against PPT adversaries, but also again quasi-polynomial time adversaries.

**Definition 11.** An MIP  $(V, P_1, \dots, P_k)$  for a language  $L$  is said to be adaptively sound against no-signaling strategies (or provers) if the following adaptive soundness property is satisfied:

**Adaptive Soundness:** For every  $\lambda \in \mathbb{N}$  and any no-signaling family of distributions  $\{\mathcal{Q}_q\}_{q \in D^k}$ ,

$$\Pr[V_2(x, q_1, \dots, q_k, a_1, \dots, a_k, \text{st}) = 1] = \text{negl}(\lambda)$$

where the probability is over  $(q_1, \dots, q_k, \text{st}) \leftarrow V_1(1^\lambda)$  and over  $(x, a_1, \dots, a_k) \leftarrow \mathcal{Q}_{(q_1, \dots, q_k)}$ , where  $x$  should be thought of as corresponding to an additional (dummy) query  $q_0$ , and thus should signal no information about the other queries  $q_1, \dots, q_k$ .

If this property is satisfied for any computationally (resp. statistically) no-signaling family of distributions  $\{\mathcal{Q}_q\}_{q \in D^k}$ , we say that the MIP has adaptive soundness against computationally (resp. statistically) no-signaling strategies.

We denote the class of MIP scheme that have adaptive soundness against computational no-signaling strategies by  $\text{MIP}^{\text{adaptive-cns}}$ . Brakerski et. al. [Brakerski, Holmgren, and Kalai \[2017\]](#) proved that  $\text{MIP}^{\text{adaptive-cns}} = \text{EXP}$ . More specifically, they prove the following theorem, which is a strengthening of [Theorem 2](#).

**Theorem 4.** For any  $T : \mathbb{N} \rightarrow \mathbb{N}$ , and any language in  $L \in \text{DTIME}(T)$ , there exists an MIP with adaptive computational no-signaling soundness (as in defined in [Definitions 5 and 11](#)).

In addition, they proved that applying the [Biehl, Meyer, and Wetzal \[1999\]](#) heuristic to any MIP that is adaptively sound against computational no-signaling strategies, results with a *non-interactive delegation scheme* that is sound assuming the standard (PPT) security of the underlying FHE scheme.

**Theorem 5.** [Brakerski, Holmgren, and Kalai \[2017\]](#) For every  $T : \mathbb{N} \rightarrow \mathbb{N}$  and for every  $L \in \text{DTIME}(T)$  there exists a non-interactive delegation scheme for  $L$  (as defined in [Definition 4](#)), assuming the existence of an FHE scheme.

Due to lack of space we do not provide any intuition behind the proof of [Theorem 4](#), and instead provide a proof sketch of [Theorem 5](#) (assuming [Theorem 4](#)).

**Proof Sketch of Theorem 5.** Fix an FHE scheme  $(\text{Gen}, \text{Enc}, \text{Dec}, \text{Eval})$ , a time bound  $T = T(\lambda)$ , and a language  $L \in \text{DTIME}(T)$ . Let

$$\text{MIP}^{\text{adaptive-cns}} = (V, P_1, \dots, P_k)$$

be an MIP for  $L$  with adaptive soundness against computationally no-signaling strategies. The existence of such a proof system for  $L$  follows from [Theorem 4](#).

The non-interactive delegation scheme, denoted by  $(V^{\text{del}}, P^{\text{del}})$  is the one obtained by applying the [Biehl, Meyer, and Wetzel \[1999\]](#) heuristic to  $\text{MIP}^{\text{adaptive-cNS}}$  and FHE.

Suppose for contradiction that there exists a cheating prover  $P^*$  such that for infinitely many  $\lambda \in \mathbb{N}$ ,

$$\Pr[V_2^{\text{del}}(x, \text{pf}, \text{st}) = 1 \wedge (x \notin L)] \geq \frac{1}{\text{poly}(\lambda)}$$

where the probability is over  $(\text{pp}, \text{st}) \leftarrow V_1^{\text{del}}(1^\lambda)$  and over  $(x, \text{pf}) \leftarrow P^*(\text{pp})$ .

We use  $P^*$  to construct an adaptive computational no-signaling strategy that contradicts the adaptive soundness condition of  $\text{MIP}^{\text{adaptive-cNS}}$ .

To this end, for every possible set of queries  $q = (q_1, \dots, q_k)$ , consider the distribution of answers  $\mathcal{Q}_q$  defined as follows:

1. For every  $i \in [k]$  sample  $(\text{pk}_i, \text{sk}_i) \leftarrow \text{Gen}(1^\lambda)$ .
2. For every  $i \in [k]$  sample  $\hat{q}_i \leftarrow \text{Enc}(q_i, \text{pk}_i)$ .
3. Let  $\text{pp} = (\hat{q}_1, \dots, \hat{q}_k)$ .
4. Compute  $(x, \text{pf}) \leftarrow P^*(\text{pp})$ .
5. Parse  $\text{pf} = (\hat{a}_1, \dots, \hat{a}_k)$ .
6. For every  $i \in [k]$  decrypt  $a_i \leftarrow \text{Dec}(\hat{a}_i, \text{sk}_i)$ .
7. Output  $(x, a_1, \dots, a_k)$ .

To reach a contradiction it remains to argue that the strategy  $\{\mathcal{Q}_q\}$  is computationally no-signaling. This follows from the security of the underlying FHE scheme. We omit the proof due to lack of space.

## References

- Prabhanjan Ananth, Yu-Chi Chen, Kai-Min Chung, Huijia Lin, and Wei-Kai Lin (2016). “[Delegating RAM Computations with Adaptive Soundness and Privacy](#)”. In: *Theory of Cryptography - 14th International Conference, TCC 2016-B, Beijing, China, October 31 - November 3, 2016, Proceedings, Part II*, pp. 3–30 (cit. on p. [3364](#)).
- Sanjeev Arora, Carsten Lund, Rajeev Motwani, Madhu Sudan, and Mario Szegedy (1998). “Proof Verification and the Hardness of Approximation Problems”. *J. ACM* 45.3, pp. 501–555 (cit. on p. [3360](#)).
- Sanjeev Arora and Shmuel Safra (1992). “[Probabilistic Checking of Proofs; A New Characterization of NP](#)”. In: *33rd Annual Symposium on Foundations of Computer Science, Pittsburgh, Pennsylvania, USA, 24-27 October 1992*, pp. 2–13 (cit. on p. [3360](#)).

- David Avis, Hiroshi Imai, and Tsuyoshi Ito (2006). “On the relationship between convex bodies related to correlation experiments with dichotomic observables”. *Journal of Physics A: Mathematical and General*, 39(36) 39.36, p. 11283 (cit. on p. 3372).
- László Babai, Lance Fortnow, Leonid A. Levin, and Mario Szegedy (1991). “Checking Computations in Polylogarithmic Time”. In: *Proceedings of the 23rd Annual ACM Symposium on Theory of Computing*. ACM, pp. 21–31 (cit. on p. 3360).
- László Babai, Lance Fortnow, and Carsten Lund (1991). “Non-Deterministic Exponential Time has Two-Prover Interactive Protocols”. *Computational Complexity* 1, pp. 3–40 (cit. on pp. 3359, 3367).
- Boaz Barak (2001). “How to Go Beyond the Black-Box Simulation Barrier”. In: *FOCS*, pp. 106–115 (cit. on p. 3362).
- Jonathan Barrett, Noah Linden, Serge Massar, Stefano Pironio, Sandu Popescu, and David Roberts (2005). “Nonlocal correlations as an information-theoretic resource”. *Physical Review A*, 71(022101) 71.2, p. 022101 (cit. on p. 3372).
- Mihir Bellare and Phillip Rogaway (1993). “Random Oracles are Practical: A Paradigm for Designing Efficient Protocols”. In: *ACM Conference on Computer and Communications Security*. Ed. by Dorothy E. Denning, Raymond Pyle, Ravi Ganesan, Ravi S. Sandhu, and Victoria Ashby. ACM, pp. 62–73 (cit. on p. 3362).
- Michael Ben-Or, Shafi Goldwasser, Joe Kilian, and Avi Wigderson (1988). “Multi-Prover Interactive Proofs: How to Remove Intractability Assumptions”. In: *Proceedings of the 20th Annual ACM Symposium on Theory of Computing*, pp. 113–131 (cit. on pp. 3358, 3359).
- Eli Ben-Sasson, Alessandro Chiesa, Christina Garman, Matthew Green, Ian Miers, Eran Tromer, and Madars Virza (2014). “[Zerocash: Decentralized Anonymous Payments from Bitcoin](#)”. In: *2014 IEEE Symposium on Security and Privacy, SP 2014, Berkeley, CA, USA, May 18-21, 2014*, pp. 459–474 (cit. on p. 3356).
- Ingrid Biehl, Bernd Meyer, and Susanne Wetzel (1999). “[Ensuring the Integrity of Agent-Based Computations by Short Proofs](#)”. In: *Proceedings of the Second International Workshop on Mobile Agents*. MA ’98. London, UK, UK: Springer-Verlag, pp. 183–194 (cit. on pp. 3363, 3368, 3369, 3371, 3373–3375).
- Nir Bitansky, Ran Canetti, Alessandro Chiesa, Shafi Goldwasser, Huijia Lin, Aviad Rubinfeld, and Eran Tromer (2014). “[The Hunting of the SNARK](#)”. *IACR Cryptology ePrint Archive* 2014, p. 580 (cit. on p. 3362).
- Nir Bitansky, Ran Canetti, Alessandro Chiesa, and Eran Tromer (2013). “[Recursive composition and bootstrapping for SNARKS and proof-carrying data](#)”. In: *Symposium on Theory of Computing Conference, STOC’13, Palo Alto, CA, USA, June 1-4, 2013*. Ed. by Dan Boneh, Tim Roughgarden, and Joan Feigenbaum. ACM, pp. 111–120 (cit. on p. 3362).

- Nir Bitansky, Alessandro Chiesa, Yuval Ishai, Rafail Ostrovsky, and Omer Paneth (2013). “[Succinct Non-interactive Arguments via Linear Interactive Proofs](#)”. In: *TCC*, pp. 315–333 (cit. on p. 3362).
- Nir Bitansky, Sanjam Garg, Huijia Lin, Rafael Pass, and Sidharth Telang (2015). “Succinct Randomized Encodings and their Applications”. *IACR Cryptology ePrint Archive* 2015, p. 356 (cit. on p. 3364).
- Andrew J. Blumberg, Justin Thaler, Victor Vu, and Michael Walfish (2014). “[Verifiable computation using multiple provers](#)”. *IACR Cryptology ePrint Archive* 2014, p. 846 (cit. on p. 3358).
- Zvika Brakerski, Justin Holmgren, and Yael Tauman Kalai (2017). “[Non-interactive delegation and batch NP verification from standard computational assumptions](#)”. In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pp. 474–482 (cit. on pp. 3364, 3373, 3374).
- Zvika Brakerski and Vinod Vaikuntanathan (2011). “[Efficient Fully Homomorphic Encryption from \(Standard\) LWE](#)”. In: *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pp. 97–106 (cit. on pp. 3363, 3369).
- Gilles Brassard, David Chaum, and Claude Crépeau (1988). “Minimum Disclosure Proofs of Knowledge”. *J. Comput. Syst. Sci.* 37.2, pp. 156–189 (cit. on p. 3361).
- Ran Canetti, Oded Goldreich, and Shai Halevi (2004). “The random oracle methodology, revisited”. *J. ACM* 51.4, pp. 557–594 (cit. on p. 3362).
- Ran Canetti and Justin Holmgren (2016). “Fully Succinct Garbled RAM”. In: *ITCS. ACM*, pp. 169–178 (cit. on p. 3364).
- Ran Canetti, Justin Holmgren, Abhishek Jain, and Vinod Vaikuntanathan (2015). “Succinct Garbling and Indistinguishability Obfuscation for RAM Programs”. In: *STOC. ACM*, pp. 429–437 (cit. on p. 3364).
- Yu-Chi Chen, Sherman S. M. Chow, Kai-Min Chung, Russell W. F. Lai, Wei-Kai Lin, and Hong-Sheng Zhou (2016). “Cryptography for Parallel RAM from Indistinguishability Obfuscation”. In: *ITCS. ACM*, pp. 179–190 (cit. on p. 3364).
- Graham Cormode, Michael Mitzenmacher, and Justin Thaler (2012). “[Practical verified computation with streaming interactive proofs](#)”. In: *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*. Ed. by Shafi Goldwasser. ACM, pp. 90–112 (cit. on p. 3358).
- Ivan Damgård (1992). “Towards Practical Public Key Systems Secure Against Chosen Ciphertext Attacks”. In: *Proceedings of CRYPTO-91*, pp. 445–456 (cit. on p. 3362).

- Ivan Damgård, Sebastian Faust, and Carmit Hazay (2012). “Secure Two-Party Computation with Low Communication”. In: *Theory of Cryptography - 9th Theory of Cryptography Conference, TCC 2012, Taormina, Sicily, Italy, March 19-21, 2012. Proceedings*, pp. 54–74 (cit. on p. 3362).
- Yevgeniy Dodis, Shai Halevi, Ron D. Rothblum, and Daniel Wichs (2016). “Spooky Encryption and Its Applications”. In: *Advances in Cryptology - CRYPTO 2016 - 36th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 14-18, 2016, Proceedings, Part III*, pp. 93–122 (cit. on pp. 3363, 3371).
- Cynthia Dwork, Michael Langberg, Moni Naor, Kobbi Nissim, and Omer Reingold (2004). “Succinct Proofs for NP and Spooky Interactions”. Unpublished manuscript, available at [http://www.cs.bgu.ac.il/~kobbi/papers/spooky\\_sub\\_crypto.pdf](http://www.cs.bgu.ac.il/~kobbi/papers/spooky_sub_crypto.pdf) (cit. on pp. 3363, 3370, 3371).
- Uriel Feige, Shafi Goldwasser, László Lovász, Shmuel Safra, and Mario Szegedy (1991). “Approximating Clique is Almost NP-Complete (Preliminary Version)”. In: *FOCS*. IEEE Computer Society, pp. 2–12 (cit. on p. 3360).
- Lance Fortnow, John Rempel, and Michael Sipser (1994). “On the Power of Multi-Prover Interactive Protocols”. *Theor. Comput. Sci.* 134.2, pp. 545–557 (cit. on p. 3360).
- Rosario Gennaro, Craig Gentry, Bryan Parno, and Mariana Raykova (2013). “Quadratic Span Programs and Succinct NIZKs without PCPs”. In: *Advances in Cryptology - EUROCRYPT 2013, 32nd Annual International Conference on the Theory and Applications of Cryptographic Techniques, Athens, Greece, May 26-30, 2013. Proceedings*, pp. 626–645 (cit. on p. 3362).
- Craig Gentry (2009). “Fully homomorphic encryption using ideal lattices”. In: *Proceedings of the 41st Annual ACM Symposium on Theory of Computing, STOC 2009, Bethesda, MD, USA, May 31 - June 2, 2009*, pp. 169–178 (cit. on pp. 3363, 3369).
- Oded Goldreich, Silvio Micali, and Avi Wigderson (1987). “How to play ANY mental game”. In: *STOC '87: Proceedings of the nineteenth annual ACM symposium on Theory of computing*, pp. 218–229 (cit. on p. 3359).
- Shafi Goldwasser and Yael Tauman Kalai (2003). “On the (In)security of the Fiat-Shamir Paradigm”. In: *FOCS*, pp. 102– (cit. on p. 3362).
- Shafi Goldwasser, Yael Tauman Kalai, and Guy N. Rothblum (2008). “Delegating computation: interactive proofs for muggles”. In: *STOC*, pp. 113–122 (cit. on pp. 3357, 3358).
- Shafi Goldwasser and Silvio Micali (1984). “Probabilistic Encryption”. *Journal of Computer and System Sciences* 28.2, pp. 270–299 (cit. on p. 3368).
- Shafi Goldwasser, Silvio Micali, and Ronald L. Rivest (1988). “A Digital Signature Scheme Secure Against Adaptive Chosen-Message Attacks”. *SIAM J. Comput.* 17.2, pp. 281–308 (cit. on pp. 3356, 3359).

- Jens Groth (2010). “Short Pairing-Based Non-interactive Zero-Knowledge Arguments”. In: *ASIACRYPT*. Vol. 6477. Lecture Notes in Computer Science. Springer, pp. 321–340 (cit. on p. 3362).
- Thomas Holenstein (2009). “Parallel Repetition: Simplification and the No-Signaling Case”. *Theory of Computing* 5.1, pp. 141–172 (cit. on p. 3372).
- Tsuyoshi Ito (2010). “Polynomial-Space Approximation of No-Signaling Provers”. In: *ICALP (1)*, pp. 140–151 (cit. on pp. 3372, 3373).
- Tsuyoshi Ito, Hirotada Kobayashi, and Keiji Matsumoto (2009). “Oracularization and Two-Prover One-Round Interactive Proofs against Nonlocal Strategies”. In: *IEEE Conference on Computational Complexity*, pp. 217–228 (cit. on pp. 3372, 3373).
- Yael Tauman Kalai and Omer Paneth (2015). “Delegating RAM Computations”. *IACR Cryptology ePrint Archive* 2015, p. 957 (cit. on p. 3364).
- Yael Tauman Kalai, Ran Raz, and Ron D. Rothblum (2013). “[Delegation for bounded space](#)”. In: *Symposium on Theory of Computing Conference, STOC’13, Palo Alto, CA, USA, June 1-4, 2013*. Ed. by Dan Boneh, Tim Roughgarden, and Joan Feigenbaum. ACM, pp. 565–574 (cit. on pp. 3362, 3363, 3371, 3373).
- (2014). “How to delegate computations: the power of no-signaling proofs”. In: *STOC*. ACM, pp. 485–494 (cit. on pp. 3362, 3363, 3371, 3373).
- Julia Kempe, Hirotada Kobayashi, Keiji Matsumoto, Ben Toner, and Thomas Vidick (2008). “Entangled Games are Hard to Approximate”. In: *FOCS*, pp. 447–456 (cit. on p. 3372).
- Leonid A. Khalfin and Boris S. Tsirelson (1985). “Quantum and quasi-classical analogs of Bell inequalities”. In: *In Symposium on the Foundations of Modern Physics*, pp. 441–460 (cit. on p. 3372).
- Joe Kilian (1992). “A Note on Efficient Zero-Knowledge Proofs and Arguments (Extended Abstract)”. In: *STOC*, pp. 723–732 (cit. on p. 3362).
- Venkata Koppula, Allison Bishop Lewko, and Brent Waters (2015). “Indistinguishability Obfuscation for Turing Machines with Unbounded Memory”. In: *STOC*. ACM, pp. 419–428 (cit. on p. 3364).
- Helger Lipmaa (2012). “Progression-Free Sets and Sublinear Pairing-Based Non-Interactive Zero-Knowledge Arguments”. In: *TCC*, pp. 169–189 (cit. on p. 3362).
- Carsten Lund, Lance Fortnow, Howard J. Karloff, and Noam Nisan (1992). “Algebraic Methods for Interactive Proof Systems”. *J. ACM* 39.4, pp. 859–868 (cit. on p. 3357).
- Silvio Micali (1994). “[CS Proofs \(Extended Abstracts\)](#)”. In: *35th Annual Symposium on Foundations of Computer Science, Santa Fe, New Mexico, USA, 20-22 November 1994*. Full version in [Micali \[2000\]](#). IEEE Computer Society, pp. 436–453 (cit. on pp. 3361, 3362).
- (2000). “Computationally Sound Proofs.” *SIAM J. Comput.* 30.4, pp. 1253–1298 (cit. on p. 3379).



- Omer Paneth and Guy N. Rothblum (2017). “On Zero-Testable Homomorphic Encryption and Publicly Verifiable Non-interactive Arguments”. In: *Theory of Cryptography - 15th International Conference, TCC 2017, Baltimore, MD, USA, November 12-15, 2017, Proceedings, Part II*, pp. 283–315 (cit. on p. 3364).
- Sandu Popescu and Daniel Rohrlich (1994). “Quantum nonlocality as an axiom”. *Foundations of Physics* 24.3, pp. 379–385 (cit. on p. 3372).
- Peter Rastall (1985). “Locality, Bell’s theorem, and quantum mechanics”. *Foundations of Physics* 15.9, pp. 963–972 (cit. on p. 3372).
- O. Regev (2003). “New lattice based cryptographic constructions”. In: *Proceedings of the 35th Annual ACM Symposium on Theory of Computing*, pp. 407–416 (cit. on p. 3369).
- Omer Reingold, Guy N. Rothblum, and Ron D. Rothblum (2016). “Constant-round interactive proofs for delegating computation”. In: *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pp. 49–62 (cit. on p. 3358).
- Adi Shamir (1992). “IP = PSPACE”. *Journal of the ACM* 39.4, pp. 869–877 (cit. on p. 3357).
- Justin Thaler (2013). “Time-Optimal Interactive Proofs for Circuit Evaluation”. In: *Advances in Cryptology - CRYPTO 2013 - 33rd Annual Cryptology Conference, Santa Barbara, CA, USA, August 18-22, 2013. Proceedings, Part II*, pp. 71–89 (cit. on p. 3358).
- Justin Thaler, Mike Roberts, Michael Mitzenmacher, and Hanspeter Pfister (2012). “Verifiable Computation with Massively Parallel Interactive Proofs”. In: *4th USENIX Workshop on Hot Topics in Cloud Computing, HotCloud’12, Boston, MA, USA, June 12-13, 2012* (cit. on p. 3358).
- Ben Toner (2009). “Monogamy of non-local quantum correlations”. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Science* 465.2101, pp. 59–69 (cit. on p. 3372).
- Victor Vu, Srinath T. V. Setty, Andrew J. Blumberg, and Michael Walfish (2013). “A Hybrid Architecture for Interactive Verifiable Computation”. In: *2013 IEEE Symposium on Security and Privacy, SP 2013, Berkeley, CA, USA, May 19-22, 2013*, pp. 223–237 (cit. on p. 3358).
- Riad S. Wahby, Max Howald, Siddharth J. Garg, Abhi Shelat, and Michael Walfish (2016). “Verifiable ASICs”. In: *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pp. 759–778 (cit. on p. 3358).
- Riad S. Wahby, Ye Ji, Andrew J. Blumberg, Abhi Shelat, Justin Thaler, Michael Walfish, and Thomas Wies (2017). “Full accounting for verifiable outsourcing”. *IACR Cryptology ePrint Archive* 2017, p. 242 (cit. on p. 3358).

Yupeng Zhang, Daniel Genkin, Jonathan Katz, Dimitrios Papadopoulos, and Charalampos Papamanthou (2017). “[vSQL: Verifying Arbitrary SQL Queries over Dynamic Outsourced Databases](#)”. In: *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pp. 863–880 (cit. on p. [3358](#)).

Received 2017-12-19.

Yael Tauman Kalai  
Microsoft Research, MIT  
[yael@microsoft.com](mailto:yael@microsoft.com)



# GRADIENTS AND FLOWS: CONTINUOUS OPTIMIZATION APPROACHES TO THE MAXIMUM FLOW PROBLEM

ALEKSANDER MAŁDZY

## Abstract

We use the lens of the maximum flow problem, one of the most fundamental problems in algorithmic graph theory, to describe a new framework for design of graph algorithms. At a high level, this framework casts the graph problem at hand as a convex optimization task and then applies to it an appropriate method from the continuous optimization toolkit. We survey how this new approach led to the first in decades progress on the maximum flow problem and then briefly sketch the challenges that still remain.

## 1 Introduction

The maximum flow problem is one of the most fundamental and extensively studied graph problems in combinatorial optimization [Schrijver \[2003\]](#), [Ahuja, Magnanti, and Orlin \[1993\]](#), and [Schrijver \[2002\]](#). It has a wide range of applications (see [Ahuja, Magnanti, Orlin, and Reddy \[1995\]](#)), is often used as subroutine in other algorithms (see, e.g., [Arora, Hazan, and Kale \[2012\]](#) and [Sherman \[2009\]](#)), and a number of other important problems – e.g., the minimum  $s$ - $t$  cut problem and the bipartite matching problem [Cormen, Leiserson, Rivest, and C. Stein \[2009\]](#) – can be reduced to it. Furthermore, this problem was often a testbed for development of fundamental algorithmic tools and concepts. Most prominently, the max-flow min-cut theorem [Elias, Feinstein, and Shannon \[1956\]](#) and [Ford and Fulkerson \[1956\]](#) constitutes a prototypical primal-dual relation.

Several decades of work resulted in a number of developments on fast algorithms for the maximum flow problem (see [Goldberg and S. Rao \[1998\]](#) for an overview) and many of this problem’s generalizations and special cases. The algorithms underlying these developments tended to be combinatorial in spirit. That is, they operated on various combinatorial notions associated with a graph, such as paths, cuts, trees, and partitions, and then used sophisticated data structures to make these operations efficient. Employing this

kind of approaches is fairly natural in this context – after all, graphs are combinatorial objects – and it was very successful too. The resulting techniques were shaping much of our understanding of not only graph algorithms but also algorithms at large.

Still, despite all this effort and successes, the basic problem of computing a maximum  $s$ - $t$  flow in general graphs resisted progress for a long time. The best known combinatorial algorithm for that problem runs in time  $O(m \min\{m^{\frac{1}{2}}, n^{\frac{2}{3}}\} \log(n^2/m) \log U)$  and it was developed already 20 years ago by [Goldberg and S. Rao \[1998\]](#). In fact, this running time bound, in turn, matches the  $O(m \min\{m^{\frac{1}{2}}, n^{\frac{2}{3}}\})$  bound that [Even and Tarjan \[1975\]](#) – and, independently, [Karzanov \[1973\]](#) – established for unit-capacity graphs almost 40 years ago.<sup>1</sup> It is thus evident that such purely combinatorial techniques have certain fundamental limitations. Consequently, there is a need for development of a different, broader perspective on graph algorithms.

In this survey, we describe such a new perspective. In a sense, this new view can be seen as a more general form of the continuous approaches to understanding graphs that were developed in the context of spectral graph theory [Chung \[1997\]](#). At a high level, it relies on casting the graph problem at hand as an optimization task that is continuous and then applying to it an appropriate tool from continuous optimization, a field that aims to design efficient algorithms for finding (approximate) minimum of a given (continuous) function over a continuous domain.

Over the last decade this new approach enabled us to make first in decades progress on a number of fundamental graph problems. Most prominently, it provided us with algorithms for the  $\varepsilon$ -approximate undirected maximum flow problem [Christiano, J. Kelner, Mądry, D. Spielman, and Teng \[2011\]](#), [Lee, S. Rao, and Srivastava \[2013\]](#), [Sherman \[2013\]](#), [J. Kelner, Lee, Orecchia, and Sidford \[2014\]](#), and [Sherman \[2017a\]](#) and the exact, directed maximum flow problem [Mądry \[2013\]](#), [Lee and Sidford \[2014\]](#), and [Mądry \[2016\]](#) that finally improve over the classic bounds due to [Even and Tarjan \[1975\]](#) and [Karzanov \[1973\]](#) as well as [Goldberg and S. Rao \[1998\]](#).

The goal of this exposition is to present a unified view on these developments. In particular, we aim to directly connect the maximum flow algorithms that have been proposed in this context to the underlying methods and notions from the field of continuous optimization. It turns out that this “tale of one problem” enables us to survey a large part of continuous optimization’s landscape. Specifically, along the way, we discuss almost all of the most fundamental tools and concepts of that field, such as different variants of the gradient descent method, and the Newton’s method. This shows that the maximum flow

---

<sup>1</sup>Here,  $m$  denotes the number of edges,  $n$  – the number of vertices, and  $U$  is the largest (integer) edge capacity.

problem might end up becoming a fertile testbed for development of new continuous optimization methods, and thus play a role that is similar to the one it has played already in the context of combinatorial methods.

**1.1 Organization of the Paper.** We begin the technical part of the paper in [Section 2](#) by introducing the basic notions that will be needed in our discussion. Then, in [Section 3](#), we provide a brief overview of basic continuous optimization tools. In [Section 4](#), we show how these tools can be used to obtain fast  $\varepsilon$ -approximate algorithms for the maximum flow problem in undirected graphs. Next, in [Section 5](#), we describe continuous optimization-based approaches to computing an exact maximum flow in directed graphs. We conclude in [Section 6](#) with a discussion of some of the key challenges that the future work in this area might be able to address.

## 2 Preliminaries

We will be viewing graphs as having both lower and upper capacities. Specifically, we will denote by  $G = (V, E, u)$  a directed graph with a vertex set  $V$ , an edge set  $E$ , and two (non-negative) integer capacities  $u_e^-$  and  $u_e^+$ , for each edge  $e \in E$ . (The role of these capacities is described below.) Usually,  $m$  will denote the number  $|E|$  of edges of the graph in question,  $n = |V|$  the number of its vertices, and  $U$  the largest edge capacity. We view each edge  $e$  of  $G$  as having an orientation  $(u, v)$ , where  $u$  is its *tail* and  $v$  is its *head*.

**Maximum Flow Problem.** Given a graph  $G$ , we view a flow in  $G$  as a vector  $f \in \mathbb{R}^m$  that assigns a value  $f_e$  to each edge  $e$  of  $G$ . When  $f_e$  is non-negative (resp. negative) we interpret it as having a flow of  $|f_e|$  flowing in (resp. opposite to) the direction of the edge  $e$  orientation.

We say that a flow  $f$  is *valid* for some demands  $\sigma \in \mathbb{R}^n$  iff it satisfies *flow conservation constraints* with respect to that demands. That is, we have that

$$(1) \quad \sum_{e \in E^+(v)} f_e - \sum_{e \in E^-(v)} f_e = \sigma_v, \quad \text{for each vertex } v \in V.$$

Here,  $E^+(v)$  (resp.  $E^-(v)$ ) is the set of edges of  $G$  that are oriented towards (resp. out of) vertex  $v$ . Intuitively, these constraints enforce that the net balance of the total in-flow into vertex  $v$  and the total out-flow out of that vertex is equal to  $\sigma_v$ , for every  $v \in V$ . (Observe that this implies, in particular, that  $\sum_v \sigma_v = 0$ .)

Now, we say that a flow  $f$  is *feasible* in  $G$  iff  $f$  obeys the *capacity constraints*:

$$(2) \quad -u_e^- \leq f_e \leq u_e^+, \quad \text{for each arc } e \in E.$$

In other words, we want each arc  $e$  to have a flow that is at most  $u_e^+$  if it flows in the direction of  $e$ 's orientation (i.e.,  $f_e \geq 0$ ), and at most  $u_e^-$ , if it flows in the opposite direction (i.e.,  $f_e < 0$ ). Note that orienting the edges accordingly and setting all  $u_e^-$ s be equal to zero recovers the standard notion of flows in directed graphs. Similarly, setting  $u_e^- = u_e^+$  for each edge  $e$  corresponds to the setting of undirected flows.

One type of flows that will be of special interest to us are  $s$ - $t$  flows, where  $s$  (the *source*) and  $t$  (the *sink*) are two distinguish vertices of  $G$ . Formally, an  $s$ - $t$  flow is a  $\sigma$ -flow whose demand vector  $\sigma$  is equal to  $F \cdot \chi_{s,t}$ , where  $F \geq 0$  is called the *value* of  $f$  and  $\chi_{s,t}$  is a demand vector that has  $-1$  (resp.  $1$ ) at the coordinate corresponding to  $s$  (resp.  $t$ ) and zeros everywhere else.

Now, the *maximum flow problem* corresponds to a task in which we are given a graph  $G = (V, E, u)$  with integer capacities as well as a source vertex  $s$  and a sink vertex  $t$  and want to find a *feasible* (in the sense of (2))  $s$ - $t$  flow of maximum value. We will denote this maximum value as  $F^*$ .

**Vector Norms.** We will find it useful to work with various  $\ell_p$ -norms of vectors. To this end, for any  $p > 0$ , we define the  $\ell_p$ -norm  $\|h\|_p$  of a vector  $h$  as  $\|h\|_p := (\sum_i |h_i|^p)^{\frac{1}{p}}$ . In particular, the  $\ell_\infty$ -norm is given by  $\|h\|_\infty := \max_i |h_i|$ . Finally, for a given positive definite matrix  $A$ , we define  $\|h\|_A := \sqrt{h^T A h}$ .

### 3 A Primer on Continuous Optimization

The framework that will be at the center of our considerations is continuous optimization or, more precisely, its part called *convex optimization*. Therefore, in this section, we provide a brief overview of this framework. (For a much more comprehensive treatment of this subject, the reader is referred to [Nemirovskii, Yudin, and Dawson \[1983\]](#), [Nesterov \[2004\]](#), [Nocedal and S. Wright \[2000\]](#), [Boyd and Vandenberghe \[2004\]](#), and [Bubeck \[2015\]](#).) Later, we will discuss how this methodology can be applied to flow problems.

**3.1 (Constrained) Minimization Problem.** At a high level, one can view continuous optimization as a set of tools designed to solve a single, general task: (*constrained*) *minimization problem*. In this problem, we are given a continuous *objective function*  $g : \mathbb{R}^k \rightarrow \mathbb{R}$  and want to solve the following optimization problem.

$$(3) \quad \min_{x \in \mathcal{K}} g(x),$$

where  $\mathcal{K} \subseteq \mathbb{R}^k$  is the *feasible set*. In its full generality, the problem (3) is intractable (or even impossible to solve). Therefore, in the context of convex optimization – which is

the context we will focus on here – we assume that both  $\mathcal{K}$  and  $g$  are convex and that a minimum we intend to find indeed exists. Additionally, whenever we have that  $\mathcal{K} = \mathbb{R}^k$ , we will call (3) *unconstrained*.

The most popular way of solving problem (3) is to apply an iterative approach to it. Specifically, we start with some initial feasible solution  $x^0 \in \mathcal{K}$  and then, repeatedly, given a current solution  $x^{t-1} \in \mathcal{K}$ , we provide a procedure (*update rule*) that produces a new, improved solution  $x^t$ . We require that

$$\lim_{t \rightarrow \infty} x^t = x^*,$$

for some optimal solution  $x^* \in \mathcal{K}$  to problem (3), and are interested in the rate of this convergence. Specifically, for a given  $\varepsilon > 0$ , we would like to understand the number  $T_\varepsilon$  of update steps needed to have that

$$(4) \quad \min_{t=0, \dots, T_\varepsilon} g(x^t) - g(x^*) \leq \varepsilon.$$

That is, the number of steps needed to guarantee that we find an  $\varepsilon$ -approximate minimizer of  $g$  in  $\mathcal{K}$ .

Convex optimization has developed a number of different update rules. Each of these rules gives rise to a different algorithm and thus a different type of bound on  $T_\varepsilon$ . Most of these methods – including each one we will discuss – require making additional assumptions on the function  $g$  that go beyond assuming that it is convex. The general principle is that the stronger conditions on  $g$  we assume, the better convergence bounds we can obtain.

In what follows, we describe the most basic examples of such algorithms: the *subgradient descent* and the *gradient descent* methods. These methods fall into the broad category of so-called *first-order methods*, i.e., algorithms whose update rules rely only on the local first-order information about the objective function  $g$ . Later on, in [Section 5](#), we will also discuss more advanced algorithms.

**3.2 Subgradient Descent Method.** Recall that a *subgradient* of a function  $g$  at a point  $x$  is any vector  $s \in \mathbb{R}^k$  such that

$$(5) \quad g(x) - g(y) \leq s^T(x - y),$$

for every  $y$ . So, in other words,  $s$  defines a linear function  $g(x) + s^T(y - x)$  that lower-bounds the function  $g$  everywhere. We denote by  $\partial g(x)$  the set of all subgradients of  $g$  at the point  $x$ .

Now, the key observation is that if  $s \in \partial g(x)$  for some (non-optimal) solution  $x \in \mathcal{K}$  and  $x^*$  is a minimizer of  $g$  in  $\mathcal{K}$ , then, by (5), it must be the case that

$$(6) \quad 0 < g(x) - g(x^*) \leq s^T(x - x^*) = (-s)^T(x^* - x),$$



i.e., the direction in which  $x^*$  lies with respect to  $x$  is positively correlated, i.e., has a positive inner product, with the direction that is *opposite* to the subgradient direction.

The above observation motivates the following natural update rule

$$(7) \quad x^t \leftarrow x^{t-1} - s,$$

where  $s \in \partial g(x^{t-1})$ .

This update rule has, however, two important issues. First of all, even if moving in the direction of  $-s$  might indeed bring us closer to a minimum  $x^*$ , it might make the point  $x^t$  lay outside of the feasible set  $\mathcal{K}$ . We thus need to have a way to map such a point outside of  $\mathcal{K}$  back to  $\mathcal{K}$  (and do it in a way that does not cancel our progress). To this end, we employ the operation of a *projection*  $\Pi_{\mathcal{K}}$  on the set  $\mathcal{K}$  defined as

$$(8) \quad \Pi_{\mathcal{K}}(x) = \operatorname{argmin}_{y \in \mathcal{K}} \|y - x\|_2,$$

and project our new solution back on the feasible set  $\mathcal{K}$  in each step. A key property of such projection that we use here is its contractivity. In particular, we have that, for any point  $y$ ,

$$\|\Pi_{\mathcal{K}}(x^*) - \Pi_{\mathcal{K}}(y)\|_2 = \|\Pi_{\mathcal{K}}(x^*) - \Pi_{\mathcal{K}}(y)\|_2 \leq \|\Pi_{\mathcal{K}}(x^*) - y\|_2,$$

where the first equality follows from the fact that  $\Pi_{\mathcal{K}}(x^*) = x^*$ , as  $x^* \in \mathcal{K}$ . As a result, projecting a point on  $\mathcal{K}$  can only bring it closer to a given minimum  $x^*$ .

The second shortcoming of the update rule (7) is related to the fact that it is not clear if fully moving in the direction of  $-s$  is not too drastic. After all, the correlation expressed by (6) does not tell us much about how *far* from  $x$  the minimum  $x^*$  lies. It only informs the direction we should take. Consequently, moving by too much could lead to vast “over-shooting” of the minimum  $x^*$  and thus lack of convergence.

To cope with this problem we need to make (a fairly minimal) assumption about the objective function  $g$ . Namely, we require that  $g$  is *L-Lipschitz*, i.e., that

$$(9) \quad |g(x) - g(y)| \leq L\|x - y\|_2,$$

for every  $x$  and  $y$  and some fixed parameter  $L$ , and then we modulate the size of our step appropriately.

Specifically, our final form of subgradient descent method update becomes

$$(10) \quad x^t \leftarrow \Pi_{\mathcal{K}}(x^{t-1} - \eta s),$$

where  $s \in \partial g(x^{t-1})$  and  $\eta > 0$  is the scalar *step size*.

With this update rule in place, one can establish the following convergence bounds for the resulting algorithm.

**Theorem 3.1** (see, e.g., Theorem 3.2 in [Bubeck \[2015\]](#)). *If the objective function  $g$  is  $L$ -Lipschitz then, for any  $\varepsilon > 0$ , the update rule (10) with  $\eta = \frac{\varepsilon}{L^2}$  delivers an  $\varepsilon$ -approximate solution after at most*

$$T_\varepsilon \leq L^2 R^2 \varepsilon^{-2}$$

*iterations, where  $R = \|x^0 - x^*\|$  is the distance of the initial solution  $x^0$  to a minimum  $x^*$ .*

The above bound is fairly remarkable as it requires very minimal assumptions on the objective function  $g$ . In particular, we do not even need  $g$  to be differentiable. Still, as we will see shortly, once stronger assumptions on  $g$  can be made, we can obtain significantly improved bounds.

**3.3 Gradient Descent Method.** Arguably, the most well-known algorithm in continuous optimization is gradient descent method. In our setting, this algorithm can be viewed as a variant of the subgradient descent method considered above in the case when the objective function  $g$  is differentiable everywhere. In this case, the *gradient*  $\nabla g(x)$  of  $g$  exists at every point  $x$  and, consequently, we have that  $\partial g(x) = \{\nabla g(x)\}$  everywhere. That is, the gradients are the (unique) subgradients of  $g$ .

The update rule (10) thus becomes

$$(11) \quad x^t \leftarrow \Pi_{\mathcal{K}}(x^{t-1} - \eta \nabla g(x^{t-1})),$$

where, again,  $\eta > 0$  is the step size.

Clearly, setting  $\eta$  as in [Theorem 3.1](#) immediately recovers the corresponding bounds. (Note that the Lipschitz constant  $L$  (cf. [Equation \(9\)](#)) corresponds to the bound on the norm of the gradient.)

However, one can get an even better bound provided  $g$  is not only Lipschitz but also has Lipschitz gradients. That is, the additional assumption to make on  $g$  is to require that it is  $\beta$ -smooth, which is defined as

$$(12) \quad \|\nabla g(x) - \nabla g(y)\|_2 \leq \beta \|x - y\|_2,$$

for every  $x$  and  $y$ .

To understand how  $\beta$ -smoothness enables us to get a better control over the progress made by the update step (11), let us assume for the sake of exposition that  $g$  is infinitely differentiable. (However, every conclusion that follows holds also without this assumption.) Applying Taylor expansion to  $g$  around a given solution  $x$ , we obtain that

$$(13) \quad g(x + \Delta) = \underbrace{g(x) + \nabla g(x)^T \Delta}_{\varphi_x(\Delta)} + \underbrace{\frac{1}{2} \Delta^T \nabla^2 g(x) \Delta + \dots}_{\varrho_x(\Delta)},$$

where  $\Delta$  is the step that we intend to take and  $\nabla^2 g(x)$  is the *Hessian* of  $g$  at  $x$ .

One should view this expansion as comprising two terms. A linear (in  $\Delta$ ) term  $\varphi_x(\Delta) = g(x) + \nabla g(x)^T \Delta$  that corresponds to a (local) linear approximation of our objective function  $g$  at the point  $x$ , and  $\varrho_x(\Delta)$  being the “tail error” of this approximation.

Now, the key point is that the convexity and  $\beta$ -smoothness of  $g$  enables us to have a fairly tight control of the tail error term  $\varrho_x(\Delta)$ . Specifically, we have that

$$(14) \quad 0 \leq \varrho_x(\Delta) \leq \frac{\beta}{2} \|\Delta\|_2^2,$$

for all  $x$  and  $\Delta$ . That is, the objective function  $g$  can be not only lowerbounded by the linear function  $\varphi_x(\Delta)$ , as before, but it also can be upperbounded by the same linear function after adding a quadratic term  $\frac{\beta}{2} \|\Delta\|_2^2$  to it.

Consequently, instead of trying to choose the step  $\Delta$  so as to directly minimize  $g$ , one can aim to minimize this upperbounding function. More precisely, we can choose  $\Delta$  so as

$$(15) \quad \Delta_x^* = \operatorname{argmin}_{\Delta} \left( \varphi_x(\Delta) + \frac{\beta}{2} \|\Delta\|_2^2 \right) = -\frac{1}{\beta} \operatorname{argmax}_{\Delta} \left( \nabla g(x)^T \Delta - \frac{1}{2} \|\Delta\|_2^2 \right).$$

An elementary calculation shows that  $\Delta_x^* = -\frac{1}{\beta} \nabla g(x)$ . This, in turn, corresponds to the update step (11) with the setting of  $\eta = \frac{1}{\beta}$ .

Indeed, with such a setting of  $\eta$  one obtains the following, improved convergence bound.

**Theorem 3.2** (see, e.g., Theorem 3.3 in [Bubeck \[2015\]](#)). *If the objective function  $g$  is  $\beta$ -smooth then, for any  $\varepsilon > 0$ , the update rule (11) with  $\eta = \frac{1}{\beta}$  delivers an  $\varepsilon$ -approximate solution to problem (3) after at most*

$$T_\varepsilon \leq O(\beta R^2 \varepsilon^{-1})$$

*iterations, where  $R = \|x^0 - x^*\|_2$  is the distance of the initial solution  $x^0$  to a minimum  $x^*$ .*

We remark that the update rule (11) and the resulting iteration bound provided above is not optimal. [Nesterov \[1983, 2005\]](#) put forth a much more involved update rule: so-called *accelerated scheme*, that enables one to obtain a significantly improved convergence.

**Theorem 3.3** (see, e.g., Theorem 3.19 in [Bubeck \[2015\]](#)). *If the objective function  $g$  is  $\beta$ -smooth then, for any  $\varepsilon > 0$ , one can compute an  $\varepsilon$ -approximate solution to problem (3) after at most*

$$T_\varepsilon \leq O\left(\sqrt{\beta} R \varepsilon^{-\frac{1}{2}}\right)$$

*iterations, where  $R = \|x^0 - x^*\|_2$  is the distance of the initial solution  $x^0$  to a minimum  $x^*$ .*

**3.4 Gradient Descent Method and Strong Convexity.** The bound provided by the gradient descent method (cf. [Theorem 3.2](#)) is a clear improvement over the convergence bound delivered by the subgradient descent method (cf. [Theorem 3.1](#)). Still, this bound is not fully satisfying as it depends polynomially on  $\varepsilon^{-1}$ . As a result, obtaining a solution that is very close to optimal, say, has  $\varepsilon = \frac{1}{n^c}$ , for some constant  $c > 1$ , becomes very expensive computationally.

It turns out, however, that gradient descent, with the exact same update rule (11), can converge *significantly* faster as long as the objective function  $g$  has an additional property: it is *strongly* convex. Formally, we say that  $g$  is  $\alpha$ -strongly convex, for some  $\alpha > 0$ , if we have that

$$(16) \quad g(x + \Delta) \geq g(x) + \nabla g(x)^T \Delta + \frac{\alpha}{2} \|\Delta\|_2^2,$$

for every  $x$  and  $\Delta$ . (Note that standard convexity corresponds to taking  $\alpha = 0$  above.)

Clearly,  $\alpha$ -strong convexity immediately implies the following straightening of the tail error approximation (14)

$$(17) \quad \frac{\alpha}{2} \|\Delta\|_2^2 \leq \varrho_x(\Delta) \leq \frac{\beta}{2} \|\Delta\|_2^2,$$

for all  $x$  and  $\Delta$ . That is, now, we can both lower- and upperbound the objective  $g$  at point  $x$  by quadratic functions. This much tighter control of the error tail  $\varrho_x(\Delta)$  leads to the following convergence bound.

**Theorem 3.4** (see, e.g., Theorem 3.10 in [Bubeck \[ibid.\]](#)). *If the objective function  $g$  is  $\alpha$ -strongly convex and  $\beta$ -smooth then, for any  $\varepsilon > 0$ , the update rule (11) with  $\eta = \frac{1}{\beta}$  delivers an  $\varepsilon$ -approximate solution to problem (3) after at most*

$$T_\varepsilon \leq O\left(\frac{\beta}{\alpha} \log \frac{R}{\varepsilon}\right)$$

iterations, where  $R = \|x^0 - x^*\|_2$  is the distance of the initial solution  $x^0$  to a minimum  $x^*$ .

Observe that the convergence bound in the above theorem is only *logarithmic* in  $\varepsilon^{-1}$  (and  $R$ ). So, this dependence is small enough that we can afford getting solutions that are close to optimal. Consequently, the key factor influencing the convergence of gradient descent in this case is the ratio  $\frac{\beta}{\alpha}$ . This ratio can be seen as expressing the worst-case ill-conditioning of the level sets of  $g$  and is thus often referred to as the *condition number* of  $g$ .

Finally, as in the previous section, the above iteration bound is not optimal. In particular, the dependence on the condition number that the above theorem presents can be improved

using an accelerated scheme due to [Nesterov \[1983, 2005\]](#), which corresponds to a certain more sophisticated update rule.

**Theorem 3.5** (see, e.g., Theorem 3.18 in [Bubeck \[2015\]](#)). *If the objective function  $g$  is  $\alpha$ -strongly convex and  $\beta$ -smooth then, for any  $\varepsilon > 0$ , one can compute an  $\varepsilon$ -approximate solution to problem (3) after at most*

$$T_\varepsilon \leq O\left(\sqrt{\frac{\beta}{\alpha}} \log \frac{R}{\varepsilon}\right)$$

*iterations, where  $R = \|x^0 - x^*\|_2$  is the distance of the initial solution  $x^0$  to a minimum  $x^*$ .*

## 4 First-Order–Based Approaches to Undirected Maximum Flow

The previous section laid down the key components of the continuous optimization framework we will need: the basic first-order convex optimization methods. We are thus ready to demonstrate how these methods can be applied to the maximum flow problem.

For now, our focus will be on solving a very basic variant of the problem: the one that corresponds to the graph being *undirected* and all capacities being *unit*, and in which we are interested in obtaining only an *approximately* optimal solution. Then, in [Section 5](#), we will extend our approach to make it deliver exact solution to the general problem.

One should note that the classic, combinatorial approaches to the maximum flow problem are not known to be able to offer improved performance for this special variant of the problem. Specifically, the best combinatorial algorithm for this setting is still the classic  $O(m \min\{\sqrt{m}, n^{\frac{2}{3}}\})$ -time algorithm for the (exact) unit-capacity maximum flow problem due to [Even and Tarjan \[1975\]](#) and [Karzanov \[1973\]](#).

**4.1 Maximum Flow as an Optimization Problem.** Our point of start is casting the (undirected) maximum flow problem as a continuous optimization task. To this end, we need first to encode our graph and the flows in it in the language of linear algebra, i.e., as vectors and matrices. Conveniently, our definition of a flow already views it as a vector in  $m$ -dimensional space. As a result, for a given demand vector  $\sigma \in \mathbb{R}^n$ , we can compactly express the flow conservation constraints (1) as

$$(18) \quad B^T f = \sigma,$$

where  $B$  is an  $m \times n$  *edge-vertex incidence matrix* defined as

$$(19) \quad B_{e,v} := \begin{cases} 1 & \text{if } v \text{ is } e\text{'s head} \\ -1 & \text{if } v \text{ is } e\text{'s tail} \\ 0 & \text{otherwise.} \end{cases}$$

Now, one should observe that by employing a simple binary search (and incurring an  $O(\log nU) = O(\log n)$  factor running time overhead) we can reduce the task of ( $\varepsilon$ -approximate) solving of the maximum flow problem to solving an ( $\varepsilon$ -approximate) *flow feasibility problem*. In the latter problem, we are given a candidate value  $F \geq 0$  as well as the desired accuracy  $\varepsilon > 0$ , and our goal is to either return a flow of value  $F/(1 + \varepsilon)$  that is feasible or to conclude that  $F > F^*$ .

The flow feasibility problem can be in turn cast as the following optimization task

$$(20) \quad \begin{aligned} \min_x \quad & \|x\|_\infty \\ \text{s.t.} \quad & x \in \mathcal{F}_{s,t}, \end{aligned}$$

where  $s$  is the source vertex,  $t$  is the sink vertex and

$$(21) \quad \mathcal{F}_{s,t} = \{x \mid B^T x = F\chi_{s,t}\},$$

i.e.,  $\mathcal{F}_{s,t}$  is the affine subspace of all the vectors  $x$  that represent an  $s$ - $t$  flow of value  $F$ .

Note that both the set  $\mathcal{F}_{s,t}$  and the objective  $\|\cdot\|_\infty$  are convex. Furthermore, it is not hard to see that obtaining an  $\varepsilon$ -approximate solution to the minimization problem (20) (and scaling this solution down by  $(1 + \varepsilon)$  when needed) gives us a solution the desired  $\varepsilon$ -approximate flow feasibility problem. So, from now on, we can focus on the former task.

**4.2 Projections and Electrical Flows.** The most straightforward way to solve the problem (20) is to apply to it the subgradient descent method (cf. Section 3.2). However, to apply the corresponding update rule (10), we need to describe how to compute the projection  $\Pi_{\mathcal{F}_{s,t}}$  on the feasible set  $\mathcal{F}_{s,t}$  (cf. (8)).

Since  $\mathcal{F}_{s,t}$  is an affine subspace (cf. (21)), a simple calculations shows that, for any vector  $x$ ,

$$(22) \quad \Pi_{\mathcal{F}_{s,t}}(x) = x - B(BB^T)^\dagger (B^T x - F\chi_{s,t}).$$

The above expression turns out to have a very natural interpretation. It corresponds to canceling out any deviation of the demand vector of the flow represented by  $x$  from the desired demand vector  $F\chi_{s,t}$  by routing it in the graph using electrical flows, i.e., flows

that minimize the energy (wrt to unit edge resistances). In particular, the matrix  $BB^T$  is the Laplacian matrix of our graph and multiplying its pseudoinverse  $(BB^T)^\dagger$  by a vector corresponds to solving a Laplacian system. Importantly, there is now a long line of work [D. A. Spielman and Teng \[2004\]](#), [Koutis, Miller, and Peng \[2010, 2011\]](#), [J. A. Kelner, Orecchia, Sidford, and Zhu \[2013\]](#), [Cohen, Kyng, Miller, Pachocki, Peng, A. B. Rao, and Xu \[2014\]](#), [Kyng, Lee, Peng, Sachdeva, and D. A. Spielman \[2016\]](#), and [Kyng and Sachdeva \[2016\]](#) that builds on an earlier work of [Vaidya \[n.d.\]](#) and [D. A. Spielman and Teng \[2003\]](#) that enables us to solve such a system, and thus also compute the projection  $\Pi_{\mathcal{F}_{s,t}}(x)$ , in nearly-linear time, i.e., in time  $\tilde{O}((m))$ .<sup>2</sup>

Once we know how to implement each iteration of the update rule (11) efficiently, we can use [Theorem 3.1](#) to bound the running time of the resulting algorithm. To this end, observe that our objective function  $\|\cdot\|_\infty$  has a Lipschitz constant (cf. (9)) of  $L = 1$ , and with some care one can show that  $R = \sqrt{m}$ , provided we start with  $x^0 = B(BB^T)^\dagger F\chi_{s,t}$ . Given that each iteration can be implemented in  $\tilde{O}(m)$  time, this yields an overall running time bound of

$$(23) \quad \tilde{O}(m) \cdot \frac{LR^2}{\varepsilon^2} = \tilde{O}(m^2\varepsilon^{-2}),$$

to get an  $\varepsilon$ -approximate solution to our problem.

**4.3 Smoothing Technique.** The running time bound (23) is hardly satisfying given that classic algorithms [Even and Tarjan \[1975\]](#) and [Karzanov \[1973\]](#) run in  $O(m \min\{\sqrt{m}, n^{\frac{2}{3}}\})$  time and deliver an *exact* solution to the problem. In fact, even the most basic Ford-Fulkerson algorithm [Ford and Fulkerson \[1956\]](#) and [Elias, Feinstein, and Shannon \[1956\]](#) offers a running time of  $O(mn)$ , which is also superior to the bound (23).

Still, this should not be viewed as the evidence that continuous optimization is an inadequate toolkit in this context. To the contrary! After all, the algorithm we just obtained came out of a very straightforward attempt. This attempt almost completely ignored the structure of the problem. It turns out that there is a principled methodology that one can apply here to address the shortcomings of this first attempt.

More precisely, a problematic issue with formulation (20) is that the objective function is convex and Lipschitz but not differentiable, let alone smooth. At first, this might seem to be an insurmountable obstacle. After all, (20) captures exactly the problem we want to solve! However, this turns out to not be entirely correct. Even though the objective itself captures our problem and is non-differentiable, it is still possible to *approximate* it with a different “proxy” objective function that is differentiable and, in fact, smooth. (Note that

---

<sup>2</sup>The notation  $\tilde{O}(t(n))$  suppresses factors that are polylogarithmic in  $t(n)$ . Also, we ignore here the fact that the solutions computed by the solver are not exact, as this shortcoming can be alleviated in a standard way.

we are looking here for approximate solution anyway, so this additional approximation is not detrimental.) This approach of substituting the original objective function with such a proxy function is known as *smoothing* Nesterov [2005]. In the context of our specific objective function, the corresponding proxy function is the *softmax* function defined as

$$(24) \quad \text{smax}_\delta(x) := \delta \ln \left( \frac{\sum_{i=1}^m e^{\frac{x_i}{\delta}} + e^{\frac{-x_i}{\delta}}}{2m} \right),$$

for any  $\delta > 0$ .

Elementary calculation shows that  $\text{smax}_\delta$  is  $\frac{1}{\delta}$ -smooth and we have that

$$\|x\|_\infty - \delta \ln 2m \leq \text{smax}_\delta(x) \leq \|x\|_\infty,$$

for any vector  $x$ . So, the parameter  $\delta$  trades off the quality of approximation of the  $l_\infty$  norm and the smoothness of the resulting function. In particular, it is not hard to see that setting  $\delta = \frac{\varepsilon}{2 \ln 2m}$  suffices for our purposes.

Now, by applying gradient descent update (11) to such *smoothened* version of the problem (20), by Theorem 3.2, we get an  $\varepsilon$ -approximation to the maximum flow problem in time

$$(25) \quad \widetilde{O}(m) \cdot O\left(\frac{\beta R^2}{\varepsilon}\right) \leq \widetilde{O}(m) \cdot O\left(\frac{R^2}{\delta \varepsilon}\right) = \widetilde{O}(m^2 \varepsilon^{-2}).$$

This new running time bound unfortunately matches the bound (23) we obtained using subgradient descent method. Still, the important differences is that now we are able to work directly with gradients of our objective function. As a result, the range of possible tools we can apply to the problem is much broader.

In particular, Christiano, J. Kelner, Mądry, D. Spielman, and Teng [2011] showed that one can use a certain variant of gradient descent method: the so-called *multiplicative weights update method* Plotkin, Shmoys, and Tardos [1995], N. E. Young [1995], N. Young [2001], and Arora, Hazan, and Kale [2012] to obtain an improved running time of  $\widetilde{O}\left(m^{\frac{3}{2}} \varepsilon^{-\frac{5}{2}}\right)$ . This running time, in a sense, already matches the classic  $O(m \min\{\sqrt{m}, n^{\frac{2}{3}}\})$  running time bound of Even and Tarjan [1975] and Karzanov [1973] whenever the graph is sparse, i.e.,  $m = O(n)$ .

**4.4 Importance of Choosing the “Right” Geometry.** As we have seen above, fairly standard gradient descent-based approaches are able to largely recover the best known classic running time bounds. At least when we only aim for an  $\varepsilon$ -approximate solution to the undirected variant of the maximum flow problem. Naturally, the fact that we merely recover these bounds might not be fully satisfying. So, it is important to understand what are the key obstacles preventing us from obtaining an actual improvement here.



The choice that turns out to play a key role in this context is the choice of the geometry we use when projecting back onto the feasible space in the update rule (11) of the gradient descent method. Specifically, we defined our projection  $\Pi_{\mathcal{F}_{s,t}}$  to be an  $\ell_2$ -projection, i.e., to always project a given point  $x$  to a feasible point  $y$  that is the closest one with respect to the  $\ell_2$  distance  $\|x - y\|_2$  (see (8)). This choice was convenient since it enables us to compute this projection fast via Laplacian system solvers (see (22)). However, the  $\ell_2$ -based geometry this projection works with ends up being ill-suited to the maximum flow problem.

The root of the problem here is that the maximum flow problem corresponds to an  $\ell_\infty$ -based geometry. In particular, as made explicit in formulation (20), its objective is to find a minimum  $\ell_\infty$  norm point in an affine space  $(\mathcal{F}_{s,t})$ . It is not hard to see, however, that the  $\ell_2$ - and  $\ell_\infty$ -based notions of distance can sometimes vary significantly. So, points that are “close” with respect to  $\ell_\infty$ -based distance might seem far when computing the projection with respect to  $\ell_2$ -based distance. More precisely, if we are working in  $m$ -dimensional space, which is the case for us, we have that

$$(26) \quad \|x\|_\infty \leq \|x\|_2 \leq \sqrt{m} \|x\|_\infty,$$

and both these inequalities are tight, e.g., when  $x$  has just a single non-zero coordinate or is an all-ones vector, respectively.

This  $\sqrt{m}$  discrepancy manifests itself directly when establishing the upper bound  $R$  on the initial distance to the optimum solution (see Theorem 3.2). In a sense, the fact that our bound on  $R$  was only  $O(\sqrt{m})$  is tied closely to the worst-case discrepancy captured by (26).

This realization prompted Christiano, J. Kelner, Małdzy, D. Spielman, and Teng [2011] to change the geometry used in the projection. Specifically, instead of working with the distance induced by the  $\ell_2$  norm, they work with an  $\ell_2$  norm that is coordinate-wise reweighted in an *adaptive* manner. That is, in each step  $t$ , one projects with respect to the distance induced by the norm  $\|\cdot\|_{D_t}$ , where each  $D_t$  is a positive diagonal matrix and, additionally,  $D_{t+1} \geq D_t$  for each  $t$ . The choice of these matrices  $D_t$  is such that the corresponding distance measure approximates the  $\ell_\infty$ -based distances well (at least, in the directions relevant for the current solution) and thus avoids the  $O(\sqrt{m})$  worst-case discrepancy discussed above. Also, since  $D_t$  is a diagonal matrix,  $\|\cdot\|_{D_t}$  is still at its core an  $\ell_2$  norm. So, this enables us to use the Laplacian solver-based approach to make the projections step remain fast.

These ideas give rise to the following result that finally improves over the classic running time bounds.

**Theorem 4.1** (Christiano, J. Kelner, Mądry, D. Spielman, and Teng [ibid.]). *For any  $\varepsilon > 0$ , one can compute an  $\varepsilon$ -approximate maximum flow in undirected graph in time  $\tilde{O}\left(m^{\frac{4}{3}}\varepsilon^{-3}\right)$ .*

Christiano, J. Kelner, Mądry, D. Spielman, and Teng [ibid.] also demonstrate how to use sparsification techniques D. A. Spielman and Teng [2008] and D. A. Spielman and Srivastava [2008] to get an  $\tilde{O}\left(mn^{\frac{1}{3}}\varepsilon^{-\frac{11}{3}}\right)$ -time algorithm. This algorithm is more favorable for dense graph setting, at the expense of slightly higher dependence on  $\frac{1}{\varepsilon}$ . Finally, it is worth noting that Lee, S. Rao, and Srivastava [2013] subsequently demonstrated that one can obtain a similar result using the accelerated gradient descent method (see Theorem 3.3).

**4.5 Interlude: Gradient Descent Method for General Norms.** As discussed in the previous section, the key to obtaining faster algorithms for the maximum flow problem is to understand and exploit the interplay between the geometries of the problem and the one used by the gradient descent method. Once we manage to align these two geometries better, we can obtain an improved running time.

This gives rise to a question: how much flexibility does the gradient descent framework have in terms of the geometry it can work in? So far, all our considerations revolved around the  $\ell_2$  geometry (and its coordinate-wise reweightings). However, it turns out that gradient descent method can be applied to *any* geometry that is induced by a norm.

In fact, our treatment of gradient descent method presented in Section 3.3 can be translated into this broader, general norm setting almost syntactically. (Although there are certain important differences.) The point of start is extending the notion of  $\beta$ -smoothness (12). We say that an objective function is  $\beta$ -smooth (with respect to a general norm  $\|\cdot\|$ ) iff

$$(27) \quad \|\nabla g(x) - \nabla g(y)\|^* \leq \beta \|x - y\|,$$

for every  $x$  and  $y$ , where  $\|\cdot\|^*$  denotes the *dual norm* of  $\|\cdot\|$ , defined as

$$(28) \quad \|y\|^* = \max_{x \neq 0} \frac{y^T x}{\|x\|}.$$

Observe that if  $\|\cdot\|$  is the  $\ell_2$  norm then  $\|\cdot\|^*$  is also the  $\ell_2$  norm. (This corresponds to the fact that  $\ell_2$  norm is self-dual.) Thus, the definition (12) is a special case of the above definition. In general, the primal norm  $\|\cdot\|$  and its dual norm  $\|\cdot\|^*$  are different. In particular,  $\|\cdot\|_p^* = \|\cdot\|_q$ , where  $\frac{1}{p} + \frac{1}{q} = 1$ .

Now, similarly as in Section 3.3, the fact that the objective function  $g$  is  $\beta$ -smooth (and convex) enables us to derive the following analogue of the bound (14) on the behavior of

the tail error term  $\varrho_x(\Delta)$  of the linear Taylor approximation  $\varphi_x(\Delta)$  (see (13)).

$$(29) \quad 0 \leq \varrho_x(\Delta) \leq \frac{\beta}{2} \|\Delta\|^2,$$

for all  $x$  and  $\Delta$ .

This, in turn, enables us to derive the optimal update  $\Delta_x^*$  (see (15)) to be

$$(30) \quad \begin{aligned} \Delta_x^* &= \operatorname{argmin}_{\Delta} \left( \varphi_x(\Delta) + \frac{\beta}{2} \|\Delta\|^2 \right) \\ &= -\frac{1}{\beta} \operatorname{argmax}_{\Delta} \left( \nabla g(x)^T \Delta - \frac{1}{2} \|\Delta\|^2 \right) = -\frac{1}{\beta} (\nabla g(x))^{\#}, \end{aligned}$$

where  $\cdot^{\#}$  operator is defined as

$$(31) \quad y^{\#} = \operatorname{argmax}_{\Delta} \left( y^T \Delta - \frac{1}{2} \|\Delta\|^2 \right).$$

Observe that, again, if  $\|\cdot\|$  is the  $\ell_2$  norm then  $(y)^{\#} = y$  and (15) becomes a special case of (30). In general, however,  $(y)^{\#} \neq y$  and, for example, when  $\|\cdot\|$  is the  $\ell_{\infty}$  norm, we have that

$$(y)_i^{\#} = \operatorname{sign}(y_i) \cdot |y|_1,$$

for each coordinate  $i$ .

The final step is to make our projection  $\Pi_{\mathcal{K}}$  correspond to the distance induced by our general norm  $\|\cdot\|$ . Namely, we take (cf. (8))

$$(32) \quad \Pi_{\mathcal{K}}(x) = \operatorname{argmin}_{y \in \mathcal{K}} \|y - x\|,$$

for any  $x$ . As a result, our overall update rule becomes

$$(33) \quad x^t \leftarrow \Pi_{\mathcal{K}} \left( x^{t-1} - \eta (\nabla g(x^{t-1}))^{\#} \right),$$

where, again,  $\eta > 0$  is the step size.

Once we put all these elements together, we obtain a direct analogue of [Theorem 3.2](#).

**Theorem 4.2.** *If the objective function  $g$  is  $\beta$ -smooth with respect to norm  $\|\cdot\|$  then, for any  $\varepsilon > 0$ , the update rule (33) with  $\eta = \frac{1}{\beta}$  delivers an  $\varepsilon$ -approximate solution after at most*

$$T_{\varepsilon} \leq \beta R^2 \varepsilon^{-1}$$

*iterations, where  $R = \|x^0 - x^*\|$  is the distance of the initial solution  $x^0$  to a minimum  $x^*$ .*

It is important to remember that even though the above theorem seems to be an almost verbatim repetition of [Theorem 3.2](#), the fact that all the notions correspond to a general norm is crucial and, as we will see shortly, it significantly enhances the power of this framework.

Finally, we note that once we extend the notion of  $\alpha$ -strong convexity (16) in an analogous manner, i.e., define the objective function to be  $\alpha$ -strong convex (with respect to a general norm  $\|\cdot\|$ ) iff

$$(34) \quad g(x + \Delta) \geq g(x) + \nabla g(x)^T \Delta + \frac{\alpha}{2} \|\Delta\|^2,$$

for every  $x$  and  $\Delta$ , we can obtain a corresponding tighter control on the behavior of the tail error term  $\varrho_x(\Delta)$  of our (local) linear approximation  $\varphi_x(D)$  of the objective function  $g$  at  $x$  (cf (13)). Specifically, as a direct analogue of the inequalities (17), we have that

$$(35) \quad \frac{\alpha}{2} \|\Delta\|^2 \leq \varrho_x(\Delta) \leq \frac{\beta}{2} \|\Delta\|^2,$$

and thus the following extension of [Theorem 3.4](#) can be established.

**Theorem 4.3.** *If the objective function  $g$  is  $\alpha$ -strongly convex and  $\beta$ -smooth with respect to a norm  $\|\cdot\|$  then, for any  $\varepsilon > 0$ , the update rule (33) with  $\eta = \frac{1}{\beta}$  delivers an  $\varepsilon$ -approximate solution after at most*

$$T_\varepsilon \leq \frac{\beta}{\alpha} \log \frac{R}{\varepsilon}$$

iterations, where  $R = \|x^0 - x^*\|$  is the distance of the initial solution  $x^0$  to a minimum  $x^*$ .

#### 4.6 Solving the $\varepsilon$ -approximate Undirected Maximum Flow in $\tilde{O}(m\varepsilon^{-1} \log U)$ Time.

As first noted by [Sherman \[2013\]](#) and [J. Kelner, Lee, Orecchia, and Sidford \[2014\]](#) independently, the framework described in the previous section is particularly well-suited to tackle the  $\varepsilon$ -approximate undirected maximum flow problem.<sup>3</sup> Indeed, let us consider applying an  $\ell_\infty$  norm variant of the gradient descent method to a version of the problem (20) that was smoothened as described in [Section 4.3](#). One can readily notice that such smoothened objective has a smoothness of

$$\beta = \frac{1}{\delta} = \frac{2 \ln 2m}{\varepsilon}$$

<sup>3</sup>Strictly speaking, the variant of the framework presented here corresponds to the one employed in the work of [J. Kelner, Lee, Orecchia, and Sidford \[2014\]](#). [Sherman \[2013\]](#) relied on its slightly different, dual variant.

also with respect to the  $\ell_\infty$  norm. Additionally, one can bound the distance to the optimum  $R$  in that norm to be only  $O(1)$ . As a result, [Theorem 4.2](#) yields an iteration bound of only  $O(\frac{\ln m}{\varepsilon^2})!$

Unfortunately, this alone does not yet provide us with a nearly-linear time algorithm. The issue is that now in each iteration of the algorithm we need to compute a projection  $\Pi_{\mathcal{F}_{s,t}}$  with respect to the  $\ell_\infty$  norm – instead of the (reweighted)  $\ell_2$  norm. This is problematic as, in general, computing this projection is as hard as solving the *exact* maximum flow problem. Specifically, if we consider a projection  $\Pi_{\mathcal{F}_{s,t}}(0)$  of the all-zero vector, we note that

$$\Pi_{\mathcal{F}_{s,t}}(0) = \operatorname{argmin}_{y \in \mathcal{F}_{s,t}} \|y\|_\infty,$$

which is exactly the *optimal* solution to our problem (20).

To circumvent this conundrum, [Sherman \[2013\]](#) and [J. Kelner, Lee, Orecchia, and Sidford \[2014\]](#) make a crucial observation: we do not need to compute the projection  $\Pi_{\mathcal{F}_{s,t}}(x)$  exactly. It suffices to compute some  $\gamma$ -approximation of it. Even if the value of  $\gamma$  is fairly large, its only impact is on the running time, i.e., the running time will depend polynomially on  $\gamma$ , but *not* on the quality of the solution one obtains in the end. With this insight in mind, they build on the work of [Małdzy \[2010\]](#) that gave an  $\tilde{O}(m^{1+o(1)})$ -time  $n^{o(1)}$ -approximation algorithm for the *value* of the undirected maximum flow, to compute such  $\gamma = n^{o(1)}$ -approximate projection in  $\tilde{O}(m^{1+o(1)})$  time. This gives rise to the following result.

**Theorem 4.4** ([Sherman \[2013\]](#) and [J. Kelner, Lee, Orecchia, and Sidford \[2014\]](#)). *For any  $\varepsilon > 0$ , one can compute an  $\varepsilon$ -approximate maximum flow in undirected graph in time  $O(m^{1+o(1)}\varepsilon^{-2} \log U)$ .*

Note that the above running time already refers to the general, not necessary unit-capacity, version of the undirected maximum flow problem. (The necessary adjustment boils down to applying an appropriate coordinate-wise scaling that corresponds to edge capacities. It also introduces the  $O(\log U)$  term in the running time due to the need to perform a binary search in the reduction of the maximum flow problem to the flow feasibility problem – see [Section 4.1](#).)

In follow up work, [Peng \[2016\]](#) provided an improved variant of the algorithm that runs in  $\tilde{O}(m\varepsilon^{-2})$  time. Finally, [Sherman \[2017a\]](#) further improved the dependence on  $\frac{1}{\varepsilon}$ , giving rise to the following theorem.

**Theorem 4.5** ([Sherman \[ibid.\]](#)). *For any  $\varepsilon > 0$ , one can compute an  $\varepsilon$ -approximate maximum flow in undirected graph in time  $\tilde{O}(m\varepsilon^{-1} \log U)$ .*

It is worth noting that in the unit capacity setting (i.e., when  $U = 1$ ), if we wanted to obtain an *exact* solution to the undirected maximum flow problem, it suffices to set  $\varepsilon = \frac{1}{\sqrt{n}}$

and then “fix” the resulting approximate solution by rounding the flow and computing at most  $\sqrt{n}$  augmenting paths, which would take  $O(m\sqrt{n})$  time. So, the above result also matches (up to polylogarithmic factors) – and, in fact, for dense graphs improves – the classic  $O(m \min\{\sqrt{m}, n^{\frac{2}{3}}\})$ -time algorithms of [Even and Tarjan \[1975\]](#) and [Karzanov \[1973\]](#) also in the regime of exact answers.

## 5 Computing Maximum Flows with Second-Order Methods

In the previous section, we demonstrated how continuous optimization–based approaches such as gradient descent method can make substantial progress on the problem of computing  $\varepsilon$ -approximately maximum flows in undirected graphs (cf. [Theorem 4.5](#)). Can though some of this progress be leveraged to get improved algorithms also for the general variant of the maximum flow problem, i.e., the task of computing *exact* maximum flows in *directed* graphs?

At first glance, the key challenge in adapting the techniques from previous section to this new setting might be that these techniques were defined only in the context of undirected graphs. So, it is unclear how to extend them to the directed graph context. It turns out, however, that this alone is not a problem. Specifically, one can show that the maximum flow problem in directed graphs can be *efficiently* reduced to the maximum flow problem in *undirected* graphs. At least, in the regime of sparse graph. (See Theorem 3.6.1 in [Mařdry \[2011\]](#) for details.)

Crucially, however, this reduction only holds if we are able to solve the undirected maximum flow problem (almost) exactly. The key shortcoming of the methods from [Section 4](#) is thus that, due to their running time bounds being polynomial in  $\varepsilon^{-1}$ , they do not offer sufficiently good accuracy here. This deficiency is, in a sense, inherent to the first-order framework that these methods rely on. Specifically, this shortcoming is tied to the fact that our objective function  $g$  in problem (20) is *not* strongly convex (and smoothening does not affect this aspect). As a result, we are unable to benefit from the corresponding, much improved convergence bound given by [Theorem 4.3](#).

This realization motivates us to consider a more powerful continuous optimization methodology: the so-called *second-order methods*, i.e., approaches that, in contrast to the first-order approaches that rely solely on the information conveyed by the gradients of the objective function, also take advantage of probing the *Hessians* of that function.

**5.1 Barrier–Based Maximum Flow Problem Formulation.** Our point of start here is casting the maximum flow problem as a special type of a constrained minimization: linear program (LP). That is, instead of formulating the maximum flow problem as a problem of  $\ell_\infty$  norm minimization over an affine subspace – as was the case in (20), we phrase it

as a task of minimizing a simple, *linear* objective function over a more complex, convex feasible set.

Specifically, the maximum flow problem can be cast as the following LP.

$$\begin{aligned}
 (36) \quad & \min_x c^T x \\
 & s.t. B^T x = 0 \\
 & x_e \leq u_e^+ \forall e \\
 & x_e \geq u_e^- \forall e.
 \end{aligned}$$

Here,  $B$  is an edge-vertex incidence matrix (cf. (19)) of the input graph after we added to it an edge  $\widehat{e}$  that connects the sink  $t$  to the source  $s$  and has its lower capacity  $u_{\widehat{e}}^-$  be 0 and its upper capacity  $u_{\widehat{e}}^+$  be unbounded. Also,  $c$  is a vector in which  $c_{\widehat{e}} = -1$  and  $c_e = 0$  for all other edges  $e$ . (This added edge  $\widehat{e}$  ensures that the circulation  $x$  that we find in the augmented graph corresponds to a valid  $s$ - $t$  flow in the original graph, and the penalty we apply to that edge ensures that in the optimal solution the corresponding  $s$ - $t$  flow is indeed a maximum one.)

As the LP (36) is an example of a constrained minimization problem, we are, in principle, able to apply to it the gradient descent framework we described in Section 3. However, in that case, the update rule (11) would require us to compute in each step the projection onto the feasible set of this LP, and it is difficult to implement that task efficiently. The key reason here is the presence of the inequality constraints. (Note that projection on the kernel of the matrix  $B$  would again correspond to solving a Laplacian system – cf. (22), and thus could be computed fast.)

To cope with this problem, we will employ a technique that is inspired by one of the most popular family of approaches to solving linear programs: the *interior point methods* (see Boyd and Vandenberghe [2004], S. J. Wright [1997], and Ye [1997]). This technique, instead of maintaining the inequality constraints explicitly, enforces them implicitly by introducing an appropriate term to the objective. Specifically, instead of solving the LP (36), we aim to solve the following constrained minimization problem.

$$\begin{aligned}
 (37) \quad & \min_x c^T x + \psi_\mu(x) \\
 & s.t. B^T x = 0,
 \end{aligned}$$

where

$$(38) \quad \psi_\mu(x) = -\mu \sum_e (\ln(u_e^+ - x_e) + \ln(x_e - u_e^-)),$$

is the *barrier function* and  $\mu > 0$  is a positive number.

Observe that as long as  $\mu > 0$  and our initial solution is feasible, the barrier function will ensure that any solution we find using an iterative minimization scheme, such as gradient descent method, remains feasible. Also, it is not hard to see that the smaller  $\mu > 0$  is the closer the optimal solution to the problem (37) is to the desired optimal solution to the LP (36). In particular, one can show that to get an  $\varepsilon$ -approximate solution to the LP (36) it suffices to set  $\mu \leq \frac{\varepsilon}{2m}$  and find the optimal solution to the corresponding problem (37).

Finally, observe that the Hessian of the objective function of (37) at some point  $x$  is equal to the Hessian  $\nabla^2 \psi_\mu(x)$  of the barrier function. The latter turns out to be a diagonal matrix with

$$(\nabla^2 \psi_\mu(x))_{e,e} = \mu \left( \frac{1}{(u_e^+ - x_e)^2} + \frac{1}{(x_e - u_e^-)^2} \right),$$

for each edge  $e$ . Thus, the Hessian of our objective function is positive definite (provided  $\mu > 0$ ).

Now, one should note that the fact that this Hessian is positive definite implies that the objective function is *strongly* convex. So, if we attempted to solve problem (37) using gradient descent method we can take advantage of the improved iteration bound described by Theorem 3.4. Importantly, this bound depends only logarithmically on  $\varepsilon^{-1}$ , which is what we need in order to be able to compute the (almost) exact solutions.

Unfortunately, even though this objective function is indeed  $\alpha$ -strongly convex and  $\beta$ -smooth, its condition number  $\frac{\beta}{\alpha}$  might be very large. The bound delivered by Theorem 4.3 would thus still be prohibitively large, even if it would have the “right” dependence on  $\varepsilon^{-1}$ . It turns out that in order to get a more efficient algorithm we need to resort to a more powerful technique: the Newton’s method.

**5.2 Interlude: Newton’s Method.** Recall that the key quantity that impacts the iteration bound given by Theorem 4.3 is the condition number  $\frac{\beta}{\alpha}$ . This number captures the degree of control we have on the behavior of the tail error  $\varrho_x(\Delta)$  – see (35), in terms of the norm  $\|\cdot\|$  we work with. However, in principle, we have a complete freedom in choosing this norm. So, one might wonder: what is the “best” norm to choose in order to make this condition number be as small as possible?

Observe that, for a given point  $x$  and sufficiently small  $\Delta$ , Taylor expansion (13) gives us that

$$\varrho_x(\Delta) \approx \frac{1}{2} \Delta^T \nabla^2 g(x) \Delta = \frac{1}{2} \|\Delta\|_{\nabla^2 g(x)}^2,$$

where the norm  $\|\cdot\|_{\nabla^2 g(x)}$  is called the *local norm* of  $g$  at  $x$ . (Note that, crucially, this norm might be different at each point  $x$ .)

So, as long as the objective function  $g$  is strongly convex for *some*  $\alpha$  with respect to the  $\ell_2$  norm (and thus  $\nabla^2 g(x) \succ 0$ ), the local norm is well-defined at each point  $x$  and



locally, i.e., in a sufficiently close neighborhood of that point  $x$ , the condition number of  $g$  with respect to  $\|\cdot\|_{\nabla^2 g(x)}$  is close to best possible, i.e., close to 1!

This suggests employing an iterative method in which we repeatedly perform a gradient descent update (33) with respect to such local norm (at the current point). This method is known as the *Newton's method*. Clearly, its most attractive feature is that whenever the “sufficiently close neighborhood” of  $x$  contains the minimum  $x^*$  of the objective function  $g$ , [Theorem 4.3](#) ensures very fast convergence *regardless of the “natural”, i.e.,  $\ell_2$ -based, condition number of  $g$* . (In fact, one can show that in this case the convergence can be even faster than the one promised by that theorem.)

Unfortunately, this method has also two important shortcomings. First of all, it has no meaningful convergence guarantees if the minimum  $x^*$  of the objective function  $g$  is not sufficiently close to the current point  $x$ . In fact, even the notion of “sufficiently close” is in general not well defined. One thus usually is able to analyze Newton's method only for special class of functions such as self-concordant functions [Nesterov and Nemirovskii \[1994\]](#). (The barrier function (38) is self-concordant by design.)

Also, the other shortcoming is that each update step of Newton's method can be computationally expensive. Specifically, recall that by (33) and (31), we have that this method's update rule becomes

$$(39) \quad x^t \leftarrow \Pi_{\mathcal{K}} \left( x^{t-1} - \eta (\nabla g(x^{t-1}))^\# \right) = \Pi_{\mathcal{K}} \left( x^{t-1} - \eta (\nabla^2 g(x^{t-1}))^{-1} \nabla g(x^{t-1}) \right),$$

where we used the fact that  $y^\# = A^{-1}y$  when we work with respect to the norm  $\|\cdot\|_A$ . So, implementing each step of the Newton's method requires solving a linear system in the local Hessian of the objective function.

**5.3 Faster Algorithms for the Maximum Flow Problem.** The shortcomings of the Newton's method that we identified above severely limit its usefulness as a general optimization procedure. Still, this method turns out to be very powerful when carefully applied. In particular, it is a key element of the interior point method-based approaches to LP solving [Boyd and Vandenberghe \[2004\]](#), [Ye \[1997\]](#), and [S. J. Wright \[1997\]](#).

More concretely, the so-called path-following variants of the interior point methods solve the LP (36) by solving a *sequence* of barrier problems (37) (instead of solving directly the one that corresponds to the desired, sufficiently small value of  $\mu$ ). Specifically, they start by obtaining a (near) optimal solution to the problem (37) for a *large* value of  $\mu$ . (One can show that after appropriate preprocessing of the problem, such a solution is readily available.) Then, these algorithms repeatedly use the Newton's method to compute a (near) optimal solution to the barrier problem (37) for a slightly smaller value of  $\mu$  *while using the previously obtained solution as a warm start*. (This warm starting is crucial as it ensures that the Newton's method is always in its rapid convergence stage.)

Now, one of the most central theoretical challenges in continuous optimization is establishing bounds on the number of iterations that the above path-following procedure requires. That is, bounding the number of such barrier subproblems that need to be solved in order to obtain an  $\varepsilon$ -approximate solution to the LP (36). In 1988, Renegar [1988] established a general iteration bound of  $O(\sqrt{m} \log \frac{1}{\varepsilon})$ . Also, Daitch and D. A. Spielman [2008] observed that when one solves flow LPs such as (36), the Newton’s method update step (39) for the corresponding barrier problem (37) can be implemented in nearly-linear time using a Laplacian solver. This enabled them to obtain an  $\tilde{O}\left(m^{\frac{3}{2}} \log U\right)$ -time algorithm for the maximum flow problem as well as a host of its generalizations.

However, this running time is hardly satisfying. In particular, it is still inferior to the running time of  $O(m \min\{m^{\frac{1}{2}}, n^{\frac{2}{3}}\} \log(n^2/m) \log U)$  due to Goldberg and S. Rao [1998]. Unfortunately, obtaining an improvement here hinges on going beyond the Renegar’s  $O(\sqrt{m} \log \frac{1}{\varepsilon})$  iteration bound, which is one of the longstanding open problems in the field.

To address this challenge, Mądry [2013] developed a more fine-grained understanding of the convergence behavior of such interior point methods. This understanding showed that, similarly as it was the case in the context of undirected maximum flow problem (see Section 4.4), the slower convergence is directly tied to an underlying “geometry mismatch”. In particular, the  $O(\sqrt{m})$  term in Renegar’s bound arises due to the compounded worst-case discrepancies between the  $\ell_\infty$  and  $\ell_4$  as well as  $\ell_4$  and  $\ell_2$  norms. Mądry then builds on the idea of adaptive coordinate-wise reweighting of the  $\ell_2$  norm put forth in the work of Christiano, J. Kelner, Mądry, D. Spielman, and Teng [2011] (see Section 4.4) to alleviate this worst case discrepancies.

Specifically, after translating to our setting, the approach of Mądry [2013] can be viewed as considering a coordinate-wise reweighted version of the barrier function  $\psi_\mu$  defined as

$$(40) \quad \psi_\mu(x) = -\mu \sum_e v_e (\ln(u_e^+ - x_e) + \ln(x_e - u_e^-)),$$

where each  $v_e \geq 1$  is an individual weight tied to edge  $e$ ’s constraints. Then, Mądry develops an adaptive scheme to update the weights  $v_e$  that is inspired by (but more involved than) the reweighting scheme due to Christiano, J. Kelner, Mądry, D. Spielman, and Teng [2011]. This scheme together with a careful analysis enables one to obtain the following result.

**Theorem 5.1** (Mądry [2013, 2016]). *The maximum flow problem can be solved in  $\tilde{O}\left(m^{\frac{10}{7}} U^{\frac{1}{7}}\right)$  time.*

Observe that the resulting algorithm constitutes the first improvement over the classic  $O(m \min\{m^{\frac{1}{2}}, n^{\frac{2}{3}}\})$ -time results of Even and Tarjan [1975] and Karzanov [1973] for

unit capacity graphs as well as the general capacity result of [Goldberg and S. Rao \[1998\]](#), provided the capacities are not too large and the graph is sufficiently sparse. In fact, the improved variant of interior point method developed in [Mądry \[2013\]](#) can be applied to solving *any* linear program. Unfortunately, the underlying reweighting scheme involves permanent perturbation of the cost vector of that solved LP. This method is thus of limited use if one cannot control the impact of these perturbations on the final solution we find, as one can in the context of the maximum flow problem. (It is, however, possible that introducing these perturbations is not truly necessary and an alternative reweighting scheme could avoid it.)

Later on, [Lee and Sidford \[2014\]](#) provided a different barrier function reweighting scheme. Their result builds on the work of [Vaidya \[1989\]](#) and delivers an interior point method that converges in only  $\tilde{O}(\sqrt{n} \log \frac{1}{\varepsilon})$  iterations, which yields the following theorem.

**Theorem 5.2** ([Lee and Sidford \[2014\]](#)). *The maximum flow problem can be solved in  $\tilde{O}(m\sqrt{n} \log U)$  time.*

This result thus improves over the classic work of [Even and Tarjan \[1975\]](#), [Karzanov \[1973\]](#) and [Goldberg and S. Rao \[1998\]](#) for dense graph case. Importantly, this new interior point method can be readily applied to *any* LP. (In fact, it was developed directly in the general LP setting.) In particular, the iteration bound it gives matches – and in some cases even outperforms – the bound stemming from the seminal self-concordant barrier of [Nesterov and Nemirovskii \[1994\]](#) and, in contrast to the latter, the method of Lee and Sidford is efficiently computable. Due to this generality, [Lee and Sidford \[2014\]](#) is able to provide, in particular, improved running times for a number of generalizations of the maximum flow problem too.

## 6 Open Problems

Last decade has brought us a significant progress on algorithms for the maximum flow problem (see Theorems 4.5, 5.1 and 5.2) as well as for the related graph problems such as (weighted) bipartite matching and general shortest path problems [Cohen, Mądry, Sankowski, and Vladu \[2017\]](#). However, there is still a number of key open problems that remain unsolved and, hopefully, further work on them will lead to a better understanding of both the graph algorithms as well as the continuous optimization toolkit we employ in this context. We briefly discuss some of these questions below.

Arguably, the most progress so far has been made in the context of the  $\varepsilon$ -approximate algorithms for the undirected maximum flow problem. In particular, [Theorem 4.5](#) delivers a solution that could be viewed, essentially, as the best possible in this regime. Still, it is interesting to understand how flexible the underlying framework is. In particular, if it is

possible to apply it with a similar success to other, related flow problems. Indeed, [Sherman \[2017b\]](#) showed that this framework can deliver an almost linear time  $\varepsilon$ -approximate algorithm for a variant of the problem (20) in which we minimize the  $\ell_1$  norm of the vector instead of its  $\ell_\infty$  norm. This corresponds to solving a special case of the *minimum-cost* flow problem in which the capacities are *unbounded* and costs are all unit. It is thus natural to attempt to use this approach to tackle the general undirected minimum-cost flow problem as well.

**Challenge 6.1.** *Solve the general minimum-cost flow problem in undirected graphs in  $O(m^{1+o(1)}\varepsilon^{-2}\log(C+U))$  time, where  $C$  is the largest (integer) cost and  $U$ , as usual, is the largest (integer) capacity.*

Intuitively, solving this problem corresponds to solving the problem (20) in the case when the objective is a *linear combination* of  $\ell_\infty$  and  $\ell_1$  norms, with each one of them being reweighted coordinate-wise according to the capacities and costs, respectively. Since these two norms are dual to each other, it is unclear how to “reconcile” them to get the desired running time improvement. The current state of the art for this problem is the  $\tilde{O}(m\sqrt{n}\log(U+C))$ -time *exact* algorithm due to [Lee and Sidford \[2014\]](#) and the  $\tilde{O}(m^{\frac{10}{7}}\log C)$ -time *exact* algorithm of [Cohen, Mądry, Sankowski, and Vladu \[2017\]](#) for the *unit capacity* (but general costs) variant of the problem.

In the context of the maximum flow problem, the most central direction is to extend further the progress made on the exact algorithms for the general, directed graph setting. (See Theorems 5.1 and 5.2.) In particular, given the fact that [Theorem 5.1](#) builds on the adaptive coordinate-wise  $\ell_2$  norm reweighting idea of [Christiano, J. Kelner, Mądry, D. Spielman, and Teng \[2011\]](#), one could wonder if it is possible to match the type of running time improvement that was achieved in the latter work.

**Challenge 6.2.** *Obtain an  $\tilde{O}(m^{\frac{4}{3}}\log U)$  time algorithm for the (exact) maximum flow problem.*

Note that the above challenge is interesting also in the setting of unit capacity maximum flow, i.e., when  $U = 1$ . As it would still constitute an improvement over the current best  $\tilde{O}(m^{\frac{10}{7}}U^{\frac{1}{7}})$ -time algorithm due to [Mądry \[2013, 2016\]](#).

Finally, even though our treatment so far focused on applying the continuous optimization framework to graph algorithms, this connection can be also used in the opposite way. That is, we can view graph algorithmic problems as a useful “testbed” for understanding the full power and limitations of the current continuous optimization tools. This understanding can, in turn, be translated into progress on the challenges in core continuous optimization.

In fact, the advances on the maximum flow problem made so far already yielded important progress on such core questions. On one hand, the  $\varepsilon$ -approximation result described in [Theorem 4.5](#) required going beyond the standard gradient descent–based framework to overcome its fundamental limitation: the inability to obtain the acceleration (as in [Theorem 3.3](#)) for the  $\ell_\infty$  norm based variant of gradient descent. On the other hand, the progress on the exact maximum flow problem captured by [Theorems 5.1](#) and [5.2](#) brought us key advances on the convergence bounds for general interior point methods. In the light of this, the natural next goal here would be to use the adaptive  $\ell_2$  norm reweighing ideas to get the following improvement.

**Challenge 6.3.** *Develop an interior point method that computes an  $\varepsilon$ -approximate solution to any linear program using  $\tilde{O}\left(m^{\frac{4}{3}} \log \frac{1}{\varepsilon}\right)$  iterations, with each iteration requiring only  $O(1)$  linear system solves<sup>4</sup>.*

One should note that this challenge subsumes [Challenge 6.2](#). This is so as the maximum flow problem can be cast as a linear program in which the linear systems to be solved in each iteration of the interior point method are Laplacian (and thus can be solved efficiently) – see [Section 5.1](#). We also view tackling the above challenge as the current most promising way to approach [Challenge 6.2](#).

## References

- Ravindra K. Ahuja, Thomas L. Magnanti, and James B. Orlin (1993). *Network flows: theory, algorithms, and applications*. Prentice-Hall (cit. on p. [3379](#)).
- Ravindra K. Ahuja, Thomas L. Magnanti, James B. Orlin, and M. R. Reddy (1995). *Applications of Network Optimization*. Vol. 7. Handbooks in Operations Research and Management Science. North-Holland, pp. 1–75 (cit. on p. [3379](#)).
- Sanjeev Arora, Elad Hazan, and Satyen Kale (2012). “The Multiplicative Weights Update Method: a Meta-Algorithm and Applications”. *Theory of Computing* 8.1, pp. 121–164 (cit. on pp. [3379](#), [3391](#)).
- Stephen Boyd and Lieven Vandenbergh (2004). *Convex Optimization*. Cambridge University Press (cit. on pp. [3382](#), [3398](#), [3400](#)).
- Sébastien Bubeck (2015). “Convex Optimization: Algorithms and Complexity”. *Found. Trends Mach. Learn.* 8.3-4, pp. 231–357 (cit. on pp. [3382](#), [3385–3388](#)).

---

<sup>4</sup>To be precise, the linear systems to solve should be in a matrix of a form  $A^T D A$ , where  $A$  is the constraint matrix of the linear program and  $D$  is some positive diagonal matrix.

- Paul Christiano, Jonathan Kelner, Aleksander Mądry, Daniel Spielman, and Shang-Hua Teng (2011). “Electrical Flows, Laplacian Systems, and Faster Approximation of Maximum Flow in Undirected Graphs”. In: *STOC’11: Proceedings of the 43rd Annual ACM Symposium on Theory of Computing*, pp. 273–281 (cit. on pp. [3380](#), [3391–3393](#), [3401](#), [3403](#)).
- Fan R. K. Chung (1997). *Spectral Graph Theory*. American Mathematical Society (cit. on p. [3380](#)).
- Michael B. Cohen, Rasmus Kyng, Gary L. Miller, Jakub W. Pachocki, Richard Peng, Anup B. Rao, and Shen Chen Xu (2014). “Solving SDD Linear Systems in Nearly  $m \log^{1/2} n$  Time”. In: *STOC’14: Proceedings of the 46th Annual ACM Symposium on Theory of Computing*, pp. 343–352 (cit. on p. [3390](#)).
- Michael B. Cohen, Aleksander Mądry, Piotr Sankowski, and Adrian Vladu (2017). “Negative-Weight Shortest Paths and Unit Capacity Minimum Cost Flow in  $\tilde{O}(m^{10/7} \log W)$  Time”. In: *SODA’17: Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms* (cit. on pp. [3402](#), [3403](#)).
- T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein (2009). *Introduction to Algorithms*. 3rd. The MIT Press (cit. on p. [3379](#)).
- Samuel I. Daitch and Daniel A. Spielman (2008). “Faster approximate lossy generalized flow via interior point algorithms”. In: *STOC’08: Proceedings of the 40th Annual ACM Symposium on Theory of Computing*, pp. 451–460 (cit. on p. [3401](#)).
- Peter Elias, Amiel Feinstein, and Claude E. Shannon (1956). “A note on the maximum flow through a network”. *IRE Transactions on Information Theory* 2 (cit. on pp. [3379](#), [3390](#)).
- Shimon Even and R. Endre Tarjan (1975). “Network Flow and Testing Graph Connectivity”. *SIAM Journal on Computing* 4.4, pp. 507–518 (cit. on pp. [3380](#), [3388](#), [3390](#), [3391](#), [3397](#), [3401](#), [3402](#)).
- Lester R Ford and Delbert R Fulkerson (1956). “Maximal Flow Through a Network”. *Canadian Journal of Mathematics* 8, pp. 399–404 (cit. on pp. [3379](#), [3390](#)).
- Andrew V. Goldberg and Satish Rao (1998). “Beyond the flow decomposition barrier”. *Journal of the ACM* 45.5, pp. 783–797 (cit. on pp. [3379](#), [3380](#), [3401](#), [3402](#)).
- Alexander V. Karzanov (1973). “O nakhozhdanii maksimal’nogo potoka v setyakh spetsial’nogo vida i nekotorykh prilozheniyakh”. *Matematicheskie Voprosy Upravleniya Proizvodstvom* 5. (in Russian; title translation: On finding maximum flows in networks with special structure and some applications), pp. 81–94 (cit. on pp. [3380](#), [3388](#), [3390](#), [3391](#), [3397](#), [3401](#), [3402](#)).
- Jonathan A. Kelner, Lorenzo Orecchia, Aaron Sidford, and Zeyuan Allen Zhu (2013). “A Simple, Combinatorial Algorithm for Solving SDD Systems in Nearly-Linear Time”. In: *STOC’13: Proceedings of the 45th Annual ACM Symposium on the Theory of Computing*, pp. 911–920 (cit. on p. [3390](#)).

- Jonathan Kelner, Yin Tat Lee, Lorenzo Orecchia, and Aaron Sidford (2014). “An Almost-Linear-Time Algorithm for Approximate Max Flow in Undirected Graphs, and its Multicommodity Generalizations”. In: *SODA’14: Proceedings of the 25th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 217–226 (cit. on pp. [3380](#), [3395](#), [3396](#)).
- Ioannis Koutis, Gary L. Miller, and Richard Peng (2010). “Approaching optimality for solving SDD systems”. In: *FOCS’10: Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, pp. 235–244 (cit. on p. [3390](#)).
- (2011). “A Nearly  $m \log n$ -Time Solver for SDD Linear Systems”. In: *FOCS’11: Proceedings of the 52nd Annual IEEE Symposium on Foundations of Computer Science*, pp. 590–598 (cit. on p. [3390](#)).
- Rasmus Kyng, Yin Tat Lee, Richard Peng, Sushant Sachdeva, and Daniel A. Spielman (2016). “Sparsified Cholesky and Multigrid Solvers for Connection Laplacians”. In: *STOC’16: Proceedings of the 48th Annual ACM Symposium on Theory of Computing* (cit. on p. [3390](#)).
- Rasmus Kyng and Sushant Sachdeva (2016). “Approximate Gaussian Elimination for Laplacians: Fast, Sparse, and Simple”. In: *FOCS’16: Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science* (cit. on p. [3390](#)).
- Yin Tat Lee, Satish Rao, and Nikhil Srivastava (2013). “A New Approach to Computing Maximum Flows using Electrical Flows”. In: *STOC’13: Proceedings of the 45th Annual ACM Symposium on the Theory of Computing*, pp. 755–764 (cit. on pp. [3380](#), [3393](#)).
- Yin Tat Lee and Aaron Sidford (2014). “Path Finding Methods for Linear Programming: Solving Linear Programs in  $\tilde{O}(\sqrt{\text{rank}})$  Iterations and Faster Algorithms for Maximum Flows”. In: *FOCS’14: Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, pp. 424–433 (cit. on pp. [3380](#), [3402](#), [3403](#)).
- Aleksander Mađry (2010). “Fast Approximation Algorithms for Cut-based Problems in Undirected Graphs”. In: *FOCS’10: Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, pp. 245–254 (cit. on p. [3396](#)).
- (2011). “From Graphs to Matrices, and Back: New Techniques for Graph Algorithms”. PhD thesis. Massachusetts Institute of Technology (cit. on p. [3397](#)).
  - (2013). “Navigating Central Path with Electrical Flows: from Flows to Matchings, and Back”. In: *FOCS’13: Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, pp. 253–262 (cit. on pp. [3380](#), [3401–3403](#)).
  - (2016). “Computing Maximum Flow with Augmenting Electrical Flow”. In: *FOCS’16: Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, pp. 593–602 (cit. on pp. [3380](#), [3401](#), [3403](#)).
- Arkadii Nemirovskii, David Borisovich Yudin, and Edgar Ronald Dawson (1983). *Problem complexity and method efficiency in optimization*. Wiley (cit. on p. [3382](#)).



- Yurii Nesterov (1983). “A method of solving a convex programming problem with convergence rate  $O(1/k^2)$ ”. *Soviet Mathematics Doklady* 27.2, pp. 372–376 (cit. on pp. 3386, 3388).
- (2004). *Introductory lectures on convex optimization*. Vol. 87. Applied Optimization. A basic course. Kluwer Academic Publishers, Boston, MA, pp. xviii+236. MR: 2142598. IA: [springer\\_10.1007-978-1-4419-8853-9](#) (cit. on p. 3382).
  - (2005). “Smooth minimization of non-smooth functions”. *Mathematical Programming* 103.1, pp. 127–152 (cit. on pp. 3386, 3388, 3391).
- Yurii Nesterov and Arkadi Nemirovskii (1994). *Interior-Point Polynomial Algorithms in Convex Programming*. Society for Industrial and Applied Mathematics (cit. on pp. 3400, 3402).
- Jorge Nocedal and Stephen Wright (2000). *Numerical Optimization*. Springer New York (cit. on p. 3382).
- Richard Peng (2016). “Approximate Undirected Maximum Flows in  $O(m \text{polylog}(n))$  Time”. In: *SODA’16: Proceedings of the 27th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1862–1867 (cit. on p. 3396).
- Serge A. Plotkin, David B. Shmoys, and Eva Tardos (1995). “Fast Approximation Algorithms for Fractional Packing and Covering Problems”. *Mathematics of Operations Research* 20, pp. 257–301 (cit. on p. 3391).
- James Renegar (1988). “A polynomial-time algorithm, based on Newton’s method, for linear programming”. *Mathematical Programming* 40, pp. 59–93 (cit. on p. 3401).
- Alexander Schrijver (2002). “On the history of the transportation and maximum flow problems”. *Mathematical Programming* 91, pp. 437–445 (cit. on p. 3379).
- (2003). *Combinatorial Optimization: Polyhedra and Efficiency*. Springer (cit. on p. 3379).
- Jonah Sherman (2009). “Breaking the Multicommodity Flow Barrier for  $O(\sqrt{\log n})$ -Approximations to Sparsest Cuts”. In: *FOCS’09: Proceedings of the 50th Annual IEEE Symposium on Foundations of Computer Science*, pp. 363–372 (cit. on p. 3379).
- (2013). “Nearly Maximum Flows in Nearly Linear Time”. In: *FOCS’13: Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, pp. 263–269 (cit. on pp. 3380, 3395, 3396).
  - (2017a). “Area-convexity,  $\ell_\infty$  regularization, and undirected multicommodity flow”. In: *STOC’17: Proceedings of the 49th Annual ACM Symposium on Theory of Computing*, pp. 452–460 (cit. on pp. 3380, 3396).
  - (2017b). “Generalized Preconditioning and Undirected Minimum-Cost Flow”. In: *SODA’17: Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 772–780 (cit. on p. 3403).



- Daniel A. Spielman and Nikhil Srivastava (2008). “Graph sparsification by effective resistances”. In: *STOC’08: Proceedings of the 40th Annual ACM Symposium on the Theory of Computing*, pp. 563–568 (cit. on p. [3393](#)).
- Daniel A. Spielman and Shang-Hua Teng (2003). “Solving Sparse, Symmetric, Diagonally-Dominant Linear Systems in Time  $O(m^{1.31})$ ”. In: *FOCS’03: Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, pp. 416–427 (cit. on p. [3390](#)).
- (2004). “Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems”. In: *STOC’04: Proceedings of the 36th Annual ACM Symposium on the Theory of Computing*, pp. 81–90 (cit. on p. [3390](#)).
  - (2008). “Spectral Sparsification of Graphs”. *CoRR* abs/0808.4134 (cit. on p. [3393](#)).
- Pravin M. Vaidya (n.d.). “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, Mineapolis. (cit. on p. [3390](#)).
- (1989). “Speeding-up linear programming using fast matrix multiplication”. In: *FOCS’89: Proceedings of the 30th Annual IEEE Symposium on Foundations of Computer Science*, pp. 332–337 (cit. on p. [3402](#)).
- Stephen J. Wright (1997). *Primal-Dual Interior-Point Methods*. Society for Industrial and Applied Mathematics (cit. on pp. [3398](#), [3400](#)).
- Yinyu Ye (1997). *Interior Point Algorithms: Theory and Analysis*. John Wiley & Sons (cit. on pp. [3398](#), [3400](#)).
- N. Young (2001). “Sequential and Parallel Algorithms for Mixed Packing and Covering”. In: *FOCS’01: Proceedings of the 42nd Annual IEEE Symposium on Foundations of Computer Science* (cit. on p. [3391](#)).
- Neal E. Young (1995). “Randomized rounding without solving the linear program”. In: *SODA’95: Proceedings of the 6th Annual ACM-SIAM Symposium on Discrete Algorithms*. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, pp. 170–178 (cit. on p. [3391](#)).

Received 2018-03-05.

ALEKSANDER MADRY

MIT

[madry@mit.edu](mailto:madry@mit.edu)

# HIGH DIMENSIONAL ESTIMATION VIA SUM-OF-SQUARES PROOFS

PRASAD RAGHAVENDRA, TSELIL SCHRAMM AND DAVID STEURER

## Abstract

Estimation is the computational task of recovering a *hidden parameter*  $x$  associated with a distribution  $\mathfrak{D}_x$ , given a *measurement*  $y$  sampled from the distribution. High dimensional estimation problems can be formulated as system of polynomial equalities and inequalities, and thus give rise to natural probability distributions over polynomial systems.

Sum of squares proofs not only provide a powerful framework to reason about polynomial systems, but they are constructive in that there exist efficient algorithms to search for sum-of-squares proofs. The efficiency of these algorithms degrade exponentially in the degree of the sum-of-squares proofs.

Understanding and characterizing the power of sum-of-squares proofs for estimation problems has been a subject of intense study in recent years. On one hand, there is a growing body of work utilizing sum-of-squares proofs for recovering solutions to polynomial systems whenever the system is feasible. On the other hand, a broad technique referred to as *pseudocalibration* has been developed towards showing lower bounds on degree of sum-of-squares proofs. Finally, the existence of sum-of-squares refutations of a polynomial system has been shown to be intimately connected to the spectrum of associated low-degree matrix valued functions. This article will survey all of these developments in the context of estimation problems.

## Contents

<b>1</b>	<b>Introduction</b>	<b>3408</b>
<b>2</b>	<b>Algorithms for high-dimensional estimation</b>	<b>3415</b>
<b>3</b>	<b>Lower bounds</b>	<b>3424</b>

## 4 Connection to Spectral Algorithms

3431

## 1 Introduction

An estimation problem is specified by a family of distributions  $\{\mathcal{D}_x\}$  over  $\mathbb{R}^N$  parametrized by  $x \in \mathbb{R}^n$ . The input consists of a sample  $y \in \mathbb{R}^N$  drawn from  $\mathcal{D}_x$  for some  $x \in \mathbb{R}^n$ , and the goal is to recover the value of the parameter  $x$ . Here  $x$  is referred to as the *hidden variable* or the *parameter*, while the sample  $y$  is the *measurement* or the *instance*. Often, it is information theoretically impossible to recover hidden variables  $x$  in that their value is not completely determined by the measurements. Further, even if the hidden variable  $x$  is completely determined by the measurements, in many high-dimensional settings it is computationally intractable to recover  $x$ . For these reasons, one often seeks to recover  $x$  approximately by minimizing the expected loss for an appropriate loss function. For example, if  $\theta(y)$  denotes the estimate for  $x$  given the measurement  $y$ , a natural goal would be to minimize the expected mean-square loss given by  $\mathbb{E}_{y \sim \mathcal{D}_x} [\|\theta(y) - x\|^2]$ .

Such a minimization problem can often be equivalently stated as the problem of finding a solution to a system of polynomial inequalities and equalities. By classical NP-completeness results, general polynomial systems in many variables are computationally intractable in the worst case. In the context of estimation problems, the estimation problem gives rise to a probability distribution over polynomial systems, and the goal is to reason about a typical system drawn from the distribution. If the underlying distributions are sufficiently well-behaved, polynomial systems yield an avenue to design algorithms for high-dimensional estimation problems.

The central tool that we will bring to bear on polynomial systems is that of sum-of-squares proofs. Sum-of-squares proofs yield a complete proof system to reason about polynomial systems [Krivine \[1964\]](#) and [Stengle \[1974\]](#). More importantly, sum-of-squares proofs are constructive: the problem of finding a sum-of-squares proof can be formulated as a semidefinite program, and thus algorithms for convex optimization can be used to find a sum-of-squares proof when one exists. The computational complexity of the algorithm grows exponentially with the degree of the polynomials involved in the sum-of-squares proof. Thus, low-degree sum-of-squares proofs can be found efficiently.

Applying low-degree sum-of-squares proofs in the context of estimation problems lays open a rich family of questions. For natural distributions of polynomial systems, if a system drawn from the distribution is feasible, can one harness the sum-of-squares proofs towards actually solving the polynomial system? (surprisingly, the answer is yes!) If the system is typically infeasible, what is the smallest degree of a sum-of-squares refutation? Are there structural characterizations of the degree of sum-of-squares refutations in terms of the properties of the distribution? Is there a connection between the existence of

low-degree sum-of-squares proofs and the spectra of random matrices associated with the distribution? In the past few years, significant strides have been made on all these fronts, exposing the contours of a rich theory that lies hidden. This survey will be devoted to expounding some of the major developments in this context.

**1.1 Estimation problems.** We will start by describing a few estimation problems that will be recurring examples in our survey.

**Example 1.1 ( $k$ -CLIQUE).** Fix a positive integer  $k \leq n$ . In the  $k$ -CLIQUE problem, a clique of size  $k$  is planted within a random graph drawn from the Erdős–Rényi distribution denoted  $\mathbb{G}(n, 1/2)$ . The goal is to recover the  $k$  clique. Formally, the structured family  $\{\mathfrak{g}\}$  is parametrized by subsets  $S \subset \binom{[n]}{k}$ . For a subset  $S \in \binom{[n]}{k}$ , the distribution  $\mathfrak{g}_S$  over  $\{0, 1\}^{\binom{[n]}{2}}$  is specified by the following sampling procedure:

- Sample a graph  $G' = ([n], E(G'))$  from the Erdős–Rényi distribution  $\mathbb{G}(n, 1/2)$  and set  $G = ([n], E(G') \cup E(K_S))$  where  $K_S$  denotes the clique on the vertices in  $S$ . Let  $y \in \{1, -1\}^{\binom{[n]}{2}}$  denote the natural  $\{1, -1\}$ -encoding of the graph  $G$ , namely,  $y_{ij} = \frac{1}{2}(1 - 2\mathbf{1}[(i, j) \in E(G)])$  for all  $i, j \in \binom{[n]}{2}$ . Set  $x := \mathbf{1}_S \in \{0, 1\}^n$ .

We will refer to the variables  $y_{ij}$  as *instance variables* as they specify the input to the problem. The variables  $x_i$  will be referred to as the *hidden variables*.

It is easy to see that for all  $k \gg 2 \log n$ , the clique  $S$  can be exactly recovered with high probability given the graph  $G$ . However, there is no known polynomial time algorithm for the problem with the best algorithm being a brute force search running in time  $n^{O(\log n)}$ . We will now see how to encode the problem as a polynomial system by encoding the constraints one at a time, i.e.,

(1-1)

$x_i$  are Boolean

$$\{x_i(1 - x_i) = 0\}_{i \in [n]}$$

(1-2)

if  $(i, j) \notin E(G)$  then  $\{i, j\}$  are not both in clique  $\{(1 - y_{ij})x_i x_j = 0\}_{\forall i, j \in \binom{[n]}{2}}$

(1-3)

at least  $k$  vertices in clique

$$\sum_{i \in [n]} x_i - k \geq 0$$

Note that the instance variables  $y_{ij}$  are given, and the hidden variables  $\{x_i\}$  are the unknowns in the polynomial system. It is easy to check that the only feasible solutions  $x \in \mathbb{R}^n$  for this system of polynomial equations are Boolean vectors  $x \in \{0, 1\}^n$  which are supported on cliques of size at least  $k$  in  $G$ .

For every estimation problem that we will encounter in this survey, one can associate two related computational problems termed refutation and distinguishing.

Estimation can be thought of as searching for a hidden structure within the input instance  $y$ . The goal of refutation is to certify that there is no hidden structure, when there is none. More precisely, a *null* distribution is a probability distribution over instances  $y$  for which there is no hidden structure  $x$ . For example, in the  $k$ -CLIQUE problem, the corresponding null distribution is just the Erdős–Rényi random graph  $\mathbb{G}(n, 1/2)$  (without a planted clique in it). With high probability, a graph  $y \sim \mathbb{G}(n, 1/2)$  has no clique with significantly more than  $2 \log n$  vertices. Therefore, for a fixed  $k \gg 2 \log n$ , given a graph  $y \sim \mathbb{G}(n, 1/2)$ , the goal of a refutation algorithm is to certify that  $y$  has no clique of size  $k$ . Equivalently, the goal of a refutation algorithm is to certify the infeasibility of the associated polynomial system.

The most rudimentary computational task associated with estimation and refutation is that of distinguishing. The setup of the distinguishing problem is as follows. Fix a prior distribution  $\pi$  on the hidden variables  $x \in \mathbb{R}^n$ , which in turn induces a distribution  $\mathfrak{D}_*$  on  $\mathbb{R}^N$ , obtained by first sampling  $x \sim \pi$  and then sampling  $y \sim \mathfrak{I}_x$ . The input consists of a sample  $y$  which is with equal probability drawn from the structured distribution  $\mathfrak{D}_*$  or the null distribution  $\mathfrak{D}_\emptyset$ . The computational task is to identify which distribution the sample  $y$  is drawn from, with a probability of success  $\frac{1}{2} + \delta$  for some constant  $\delta > 0$ . For example, the structured distribution for  $k$ -CLIQUE is obtained by setting the prior distribution of  $x$  to be uniform on subsets of size  $k$ . In the distinguishing problem, the input is a graph drawn from either  $\mathfrak{D}_*$  or the null distribution  $\mathbb{G}(n, 1/2)$  and the algorithm is required to identify the distribution. For every problem included in this survey, the distinguishing task is formally no harder than estimation or refutation, i.e., the existence of algorithms for estimation or refutation immediately implies a distinguishing algorithm.

**Example 1.2.** (TENSOR PCA) The family of structured distributions  $\{\mu_x\}$  is parametrized by unit vectors  $x \in \mathbb{R}^n$ . A sample from  $\mu_x$  consists of a symmetric 4-tensor  $y = x^{\otimes 4} + \zeta$  where  $\zeta \in \mathbb{R}^{n \times n \times n \times n}$  is a symmetric 4-tensor whose entries are i.i.d Gaussian random variables sampled from  $N(0, \sigma^2)$ . The goal is to recover a vector  $x'$  that is close as possible to  $x$ .

A canonical strategy to recover  $x$  given  $y = x^{\otimes 4} + \zeta$  is to maximize the degree-4 polynomial associated with the symmetric 4 tensor  $y$ . Specifically, if we set

$$x' = \operatorname{argmax}_{\|x'\| \leq 1} \langle y, x'^{\otimes 4} \rangle$$

then one can show that  $\|x - x'\|_2 \leq O(n^{1/2} \cdot \sigma)$  with high probability over  $\zeta$ . If  $y \sim \mathfrak{I}_x$  then  $\langle y, x^{\otimes 4} \rangle = 1$ . Furthermore, when  $\sigma \ll n^{-1/2}$  it can be shown that  $x \in \mathbb{R}^n$  is close to the unique maximizer of the function  $\phi(z) = \langle y, z^{\otimes 4} \rangle$ . So the problem of recovering

$x$  can be encoded as following polynomial system:

$$(1-4) \quad \|x\|^2 \leq 1, \quad \sum_{i,j,k,\ell \in [n]^4} y_{ijkl} x_i x_j x_k x_\ell \geq \tau.$$

where  $\tau := 1$ .

In the distinguishing and refutation versions of this problem, we will take the *null* distribution  $\mathfrak{D}_\emptyset$  to be the distribution over 4-tensors with independent Gaussian entries sampled from  $N(0, \sigma^2)$  (matching the distribution of the noise  $\zeta$  from  $\mathfrak{D}_*$ ). For a 4-tensor  $y$ , the maximum of  $y(x) = \langle x^{\otimes 4}, y \rangle$  over the unit ball is referred to as the *injective tensor norm* of the tensor  $y$ , and is denoted by  $\|y\|_{\text{inj}}$ . If  $y \sim \mathfrak{D}_\emptyset$  then  $\|y\|_{\text{inj}} \leq O(n^{1/2} \cdot \sigma)$  with high probability over choice of  $y$ . Thus when  $\sigma \ll n^{-1/2}$ , the refutation version of the TENSOR PCA problem reduces to certifying an upper bound on  $\|y\|_{\text{inj}}$ . If we could compute  $\|y\|_{\text{inj}}$  exactly, then we can certify that  $y \sim \mathfrak{D}_\emptyset$  for  $\sigma$  as large as  $\sigma = O(n^{-1/2})$ . The injective tensor norm is known to be computationally intractable in the worst case [Gurvits \[2003\]](#), [Gharibian \[2010\]](#), and [Barak, Brandão, Harrow, Kelner, Steurer, and Zhou \[2012\]](#).

**Example 1.3.** (Matrix & Tensor Completion) In matrix completion, the hidden parameter is a rank- $r$  matrix  $X \in \mathbb{R}^{n \times n}$ . For a parameter  $X$ , the measurement consists of a partial matrix revealing a subset of entries of  $X$ , namely  $X_\Omega$  for a subset  $\Omega \subset [n] \times [n]$  with  $|\Omega| = m$ . The probability distribution  $\mu_X$  over measurements is obtained by picking the set  $\Omega$  to be a uniformly random subset of  $m$  entries. To formulate a polynomial system for recovering a rank- $r$  matrix consistent with the measurement  $X_\Omega$ , we will use a  $n \times r$  matrix of variables  $B$ , and write the following system of constraints on it:

$$(BB^T)_\Omega = X_\Omega \quad (BB^T \text{ is consistent with measurement})$$

Tensor completion is the analogous problem with  $X$  being a higher-order tensor namely,  $X = \sum_{i=1}^r a_i^{\otimes k}$  for some fixed  $k \in \mathbb{N}$ . The corresponding polynomial system is again over a  $n \times r$  matrix of variables  $B$  with columns  $b_1, \dots, b_r$  and the following system of constraints,

$$\left( \sum_{i \in [r]} b_i^{\otimes k} \right)_\Omega = X_\Omega \quad (\sum_{i=1}^r b_i^{\otimes k} \text{ is consistent with measurement})$$

**1.2 Sum-of-squares proofs.** The sum-of-squares (SoS) proof system is a restricted class of proofs for reasoning about polynomial systems. Fix a set of polynomial inequalities  $\mathcal{Q} = \{p_i(x) \geq 0\}_{i \in [m]}$  in variables  $x_1, \dots, x_n$ . We will refer to these inequalities as the *axioms*. Starting with the axioms  $\mathcal{Q}$ , a sum-of-squares proof of  $q(x) \geq 0$  is given by

an identity of the form,

$$\left( \sum_{i \in [m']} b_i^2(x) \right) \cdot q(x) = \sum_j s_j^2(x) + \sum_{i \in [m]} a_i^2(x) \cdot p_i(x),$$

where  $\{s_j(x)\}$ ,  $\{a_i(x)\}_{i \in [m]}$ ,  $\{b_i(x)\}_{i \in [m']}$  are real polynomials. It is clear that any identity of the above form manifestly certifies that the polynomial  $q(x) \geq 0$ , whenever each  $p_i(x) \geq 0$  for real  $x$ . The degree of the sum-of-squares proof is the maximum degree of all the summands, i.e.,  $\max\{\deg(s_j^2), \deg(a_i^2 p_i)\}_{i,j}$ .

The notion extends naturally to polynomial systems that involves a set of equations  $\{r_i(x) = 0\}$  along with a set of inequalities  $\{p_i(x) \geq 0\}$ . A syntactic approach to extend the definition would be to replace each equality  $r_i(x) = 0$  by a pair of inequalities  $r_i(x) \geq 0$  and  $-r_i(x) \geq 0$ .

We will use the notation  $\mathcal{Q} \mid \frac{x}{d} \{q(x) \geq 0\}$  to denote that the assertion that, there exists a degree  $d$  sum-of-squares proof of  $q(x) \geq 0$  from the set of axioms  $\mathcal{Q}$ . The superscript  $x$  in the notation  $\mathcal{Q} \mid \frac{x}{d} \{q(x) \geq 0\}$  indicates that the sum-of-squares proof is an identity of polynomials where  $x$  is the formal variable. We will drop the subscript or superscript when it is clear from the context, and just write  $\mathcal{Q} \vdash \{q(x) \geq 0\}$ . Sum-of-squares proofs can also be used to certify the infeasibility, a.k.a., *refute* the polynomial system. In particular, a degree  $d$  sum-of-squares refutation of a polynomial system  $\{p_i(x) \geq 0\}_{i \in [m]}$  is an identity of the form,

$$(1-5) \quad -1 = \sum_{i \in [k]} s_i^2(x) + \sum_{i \in [m]} a_i^2(x) \cdot p_i(x)$$

where  $\max\{\deg(s_j^2), \deg(a_i^2 p_i)\}_{i,j}$  is at most  $d$ .

Sum-of-square proof system have been an object of study starting with the work of Hilbert and Minkowski more than a century ago (see [Reznick \[2000\]](#) for a survey). With no restriction on degree, Stengle's Positivstellensatz imply that sum-of-squares proofs form a complete proof system, i.e., if the axioms  $\mathcal{Q}$  imply  $q(x) \geq 0$ , then there is a sum-of-squares proof of this fact.

The algorithmic implications of sum-of-squares proof system were realized starting with the work of [Parrilo \[2000\]](#) and [Lasserre \[2000\]](#), who independently arrived at families of algorithms for polynomial optimization using semidefinite programming (SDP). Specifically, these works observed that semidefinite programming can be used to find a degree- $d$  sum-of-squares proof in time  $n^{O(d)}$ , if there exists one. This family of algorithms (called a hierarchy, as we have algorithms for each even integer degree  $d$ ) are referred to as the low-degree sum-of-squares SDP hierarchy.

The SoS hierarchy has since emerged as one of the most powerful tools for algorithm design. On the one hand, a vast majority of algorithms in combinatorial optimization and

approximation algorithms developed over several decades can be systematically realized as being based on the first few levels of this hierarchy. Furthermore, the low-degree SoS SDP hierarchy holds the promise of yielding improved approximations to NP-hard combinatorial optimization problems, approximations that would beat the long-standing and universal barrier posed by the notorious unique games conjecture [Trevisan \[2012\]](#) and [Barak and Steurer \[2014\]](#).

More recently, the low-degree SoS SDP hierarchy has proved to be a very useful tool in designing algorithms for high-dimensional estimation problems, wherein the inputs are drawn from a natural probability distribution. For this survey, we organize the recent work on this topic into three lines of work.

- *When the polynomial system for an estimation problem is feasible, can sum-of-squares proofs be harnessed to retrieve the solution?* The answer is YES for many estimation problems including tensor decomposition, matrix and tensor completion. Furthermore, there is a simple and unifying principle that underlies all of these applications. Specifically, the underlying principle asserts that if there is a low-degree SoS proof that all solutions to the system are close to the hidden variable  $x$ , then low-degree SoS SDP can be used to actually retrieve  $x$ . We will discuss this broad principle and many of its implications in [Section 2](#).
- *When the polynomial system is infeasible, what is the smallest degree at which it admits sum-of-squares proof?* The degree of the sum-of-squares refutation is critical for the run-time of the SoS SDP based algorithm. Recent work by Barak et al. [Barak, Hopkins, Kelner, P. Kothari, Moitra, and A. Potechin \[2016\]](#) introduces a technique referred to as “pseudocalibration” for proving lower bounds on the degree of SoS refutation, developed in the context of the work on  $k$ -CLIQUE. [Section 3](#) is devoted to the heuristic technique of pseudocalibration, and the mystery surrounding its effectiveness.
- *Can the existence of degree- $d$  of sum-of-square refutations be characterized in terms of properties of the underlying distribution?* In [Section 4](#), we will discuss a result that shows a connection between the existence of low-degree sum-of-squares refutations and the spectra of certain low-degree matrices associated with the distribution. This connection implies that under fairly mild conditions, the SoS SDP based algorithms are no more powerful than a much simpler class of algorithms referred to as *spectral algorithms*. Roughly speaking, a spectral algorithm proceeds by constructing a matrix  $M(x)$  out of the input instance  $x$ , and then using the eigenvalues of the matrix  $M(x)$  to recover the desired outcome.



**Notation.** For a positive integer  $n$ , we use  $[n]$  to denote the set  $\{1, \dots, n\}$ . We sometimes use  $\binom{[n]}{d}$  to denote the set of all subsets of  $[n]$  of size  $d$ , and  $[n]^{\leq d}$  to denote the set of all multi-subsets of cardinality at most  $d$ .

If  $x \in \mathbb{R}^n$  and  $A \subset [n]$  is a multiset, then we will use the shorthand  $x^A$  to denote the monomial  $x^A = \prod_{i \in A} x_i$ . We will also use  $x^{\leq d}$  to denote the  $N \times 1$  vector containing all monomials in  $x$  of degree at most  $d$  (including the constant monomial 1), where  $N = \sum_{i=0}^d n^i$ . Let  $\mathbb{R}[x]_{\leq d}$  denote the space of polynomials of degree at most  $d$  in variables  $x$ .

For a function  $f(n)$ , we will say  $g(n) = O(f(n))$  if  $\lim_{n \rightarrow \infty} \frac{g(n)}{f(n)} \leq C$  for some universal constant  $C$ . We say that  $f(n) \ll g(n)$  if  $\lim_{n \rightarrow \infty} \frac{f(n)}{g(n)} = 0$ .

If  $\mu$  is a distribution over the probability space  $\mathfrak{S}$ , then we use the notation  $x \sim \mu$  for  $x \in \mathfrak{S}$  sampled according to  $\mu$ . For an event  $\mathfrak{E}$ , we will use  $\mathbf{1}[\mathfrak{E}]$  as the indicator that  $\mathfrak{E}$  occurs. We use  $\mathbb{G}(n, 1/2)$  to denote the Erdős–Rényi distribution with parameter  $1/2$ , or the distribution over graphs where each edge is included independently with probability  $1/2$ .

If  $M$  is an  $n \times m$  matrix, we use  $\lambda_{\max}(M)$  to denote  $M$ 's largest eigenvector. When  $n = m$ , then  $\text{Tr}(M)$  denotes  $M$ 's trace. If  $N$  is an  $n \times m$  matrix as well, then we use  $\langle M, N \rangle = \text{Tr}(MN^\top)$  to denote the *matrix inner product*. We use  $\|M\|_F$  to denote the Frobenius norm of  $M$ ,  $\|M\|_F = \langle M, M \rangle$ . For a subset  $S \subset [n]$ , we will use  $\mathbf{1}_S$  to denote the  $\{0, 1\}$  indicator vector of  $S$  in  $\mathbb{R}^n$ . We will also use  $\mathbf{1}$  to denote the all-1's vector.

For two matrices  $A, B$  we use  $A \otimes B$  to denote both the Kronecker product of  $A$  and  $B$ , and the order-4 tensor given by taking  $A \otimes B$  and reshaping it with modes for the rows and columns of  $A$  and of  $B$ . We also use  $A^{\otimes k}$  to denote the  $k$ -th Kronecker power of  $A$ ,  $A \otimes A \otimes \dots \otimes A$ .

**Pseudoexpectations.** If there is no degree- $d$  refutation, the dual semidefinite program gives rise to a linear functional over degree  $d$  polynomials which we term a *pseudoexpectation*. Formally, a pseudoexpectation  $\tilde{\mathbb{E}} : \mathbb{R}[x]_{\leq d} \rightarrow \mathbb{R}$  is a linear functional over polynomials of degree at most  $d$  with the properties that  $\tilde{\mathbb{E}}[1] = 1$ ,  $\tilde{\mathbb{E}}[p(x)a^2(x)] \geq 0$  for all  $p \in \mathcal{P}$  and polynomials  $a$  such that  $\deg(a^2 \cdot p) \leq d$ , and  $\tilde{\mathbb{E}}[q(x)^2] \geq 0$  whenever  $\deg(q^2) \leq d$ .

*Claim 1.4.* Suppose there exists a degree  $d$  pseudoexpectation  $\tilde{\mathbb{E}} : \mathbb{R}[x]_{\leq d} \rightarrow \mathbb{R}$  for the polynomial system  $\mathcal{P} = \{p_i(x) \geq 0\}_{i \in [m]}$ , then  $\mathcal{P}$  does not admit a degree  $d$  refutation.

*Proof.* Suppose  $\mathcal{P}$  admits a degree  $d$  refutation. Applying the pseudoexpectation operator  $\tilde{\mathbb{E}}$  to the left-hand-side of Equation (1-5), we have  $-1$ . Applying  $\tilde{\mathbb{E}}$  to the right-hand-side of Equation (1-5), the first summand must be non-negative by definition of  $\tilde{\mathbb{E}}$  since it is a sum of squares, and the second summand is non-negative, since we assumed that  $\tilde{\mathbb{E}}$  satisfies the constraints of  $\mathcal{P}$ . This yields a contradiction.  $\square$

## 2 Algorithms for high-dimensional estimation

In this section, we prove a algorithmic meta-theorem for high-dimensional estimation that provides a unified perspective on the best known algorithms for a wide range of estimation problems. Through this unifying perspective we are also able to obtain algorithms with significantly than what's known to be possible with other methods.

**2.1 Algorithmic meta-theorem for estimation.** We consider the following general class of estimation problems, which will turn out to capture a plethora of interesting problems in a useful way: In this class, an estimation problem<sup>1</sup> is specified by a set  $\mathcal{P} \subseteq \mathbb{R}^n \times \mathbb{R}^m$  of pairs  $(x, y)$ , where  $x$  is called *parameter* and  $y$  is called *measurement*. Nature chooses a pair  $(x^*, y^*) \in \mathcal{P}$ , we are given the measurement  $y^*$  and our goal is to (approximately) recover the parameter  $x^*$ .

For example, we can encode compressed sensing with measurement matrix  $A \in \mathbb{R}^{m \times n}$  and sparsity bound  $k$  by the following set of pairs,

$$\mathcal{P}_{A,k} = \{(x, y) \mid y = Ax, x \in \mathbb{R}^n \text{ is } k\text{-sparse}\}.$$

Similarly, we can encode matrix completion with observed entries  $\Omega \subseteq [n] \times [n]$  and rank bound  $r$  by the set of pairs,

$$\mathcal{P}_{\Omega,r} = \{(X, X_\Omega) \mid X \in \mathbb{R}^{n \times n}, \text{rank } X \leq r\}.$$

For both examples, the measurement was a simple (linear) function of the parameter.

**Identifiability.** In general, an estimation problem  $\mathcal{P} \subseteq \mathbb{R}^n \times \mathbb{R}^m$  may be ill-posed in the sense that, even ignoring computational efficiency, it may not be possible to (approximately) recover the parameter for a measurement  $y$  because we have  $(x, y), (x', y) \in \mathcal{P}$  for two far-apart parameters  $x$  and  $x'$ .

For a pair  $(x, y) \in \mathcal{P}$ , we say that  $y$  *identifies  $x$  exactly* if  $(x', y) \notin \mathcal{P}$  for all  $x' \neq x$ . Similarly, we say that  $y$  *identifies  $x$  up to error  $\varepsilon > 0$*  if  $\|x - x'\| \leq \varepsilon$  for all  $(x', y) \in \mathcal{P}$ . We say that  $x$  is *identifiable (up to error  $\varepsilon$ )* if every  $(x, y) \in \mathcal{P}$  satisfies that  $y$  identifies  $x$  (up to error  $\varepsilon$ ).

For example, for compressed sensing  $\mathcal{P}_{A,k}$ , it is not difficult to see that every  $k$ -sparse vector is identifiable if every subset of at most  $2k$  columns of  $A$  is linearly independent. For tensor decomposition, it turns out, for example, that the observation

<sup>1</sup> In contrast to the discussion of estimation problems in [Section 1](#), for every parameter, we have a set of possible measurements as opposed to a distribution over measurements. We can model distributions over measurements in this way by considering a set of “typical measurements”. The viewpoint in terms of sets of possible measurements will correspond more closely to the kind of algorithms we consider.

$f(x) = \sum_{i=1}^r x_i^{\otimes 3}$  is enough to identify  $x \in \mathbb{R}^{n \times r}$  (up to a permutation of its columns) if the columns  $x_1, \dots, x_r \in \mathbb{R}^n$  of  $x$  are linearly independent.

**Identifiability proofs to efficient algorithms.** By itself, identifiability typically only implies that there exists an inefficient algorithm to recover a vector  $x$  close to the parameter  $x^*$  from the observation  $y^*$ . But perhaps surprisingly, the notion of identifiability in a broader sense can also help us understand if there exists an efficient algorithm for this task. Concretely, if the *proof of identifiability* is captured by the sum-of-squares proof system at low degree, then there exists an efficient algorithm to (approximately) recover  $x$  from  $y$ .

In order to formalize this phenomenon, let the set  $\mathcal{P} \subseteq \mathbb{R}^n \times \mathbb{R}^m$  be described by polynomial equations

$$\mathcal{P} = \{(x, y) \mid \exists z. p(x, y, z) = 0\},$$

where  $p = (p_1, \dots, p_t)$  is a vector-valued polynomial and  $z$  are auxiliary variables.<sup>2</sup> (In other words,  $\mathcal{P}$  is a projection of the variety given by the polynomials  $p_1, \dots, p_t$ .) The following theorem shows that there is an efficient algorithm to (approximately) recover  $x^*$  given  $y^*$  if there exists a low-degree proof of the fact that the equation  $p(x, y^*, z) = 0$  implies that  $x$  is (close to)  $x^*$ .

**Theorem 2.1** (Meta-theorem for efficient estimation). *Let  $p$  be a vector-valued polynomial and let the triples  $(x^*, y^*, z^*)$  satisfy  $p(x^*, y^*, z^*) = 0$ . Suppose  $\mathcal{Q} \mid_{\frac{x, z}{\ell}} \{\|x^* - x\|^2 \leq \varepsilon\}$ , where  $\mathcal{Q} = \{p(x, y^*, z) = 0\}$ . Then, every level- $\ell$  pseudo-distribution  $D$  consistent with the constraints  $\mathcal{Q}$  satisfies*

$$\left\| x - \tilde{\mathbb{E}}_{D(x, z)} x \right\|^2 \leq \varepsilon.$$

Furthermore, for every  $\ell \in \mathbb{N}$ , there exists a polynomial-time algorithm (with running time  $n^{O(\ell)}$ )<sup>3</sup> that given a vector-valued polynomial  $p$  and a vector  $y$  outputs a vector  $\hat{x}(y)$  with the following guarantee: if  $\mathcal{Q} \mid_{\frac{x, z}{\ell}} \{\|x^* - x\|^2 \leq \varepsilon\}$  with a proof of bit-complexity at most  $n^\ell$ , then  $\|x^* - \hat{x}(y^*)\|^2 \leq \varepsilon + 2^{-n^\ell}$ .

Despite not being explicitly stated, the above theorem is the basis for many recent advances in algorithms for estimation problems through the sum-of-squares method [Barak, Kelner, and Steurer \[2015, 2014\]](#), [Hopkins, Shi, and Steurer \[2015\]](#), [Ma, Shi, and Steurer](#)

<sup>2</sup> We allow auxiliary variables here because they might make it easier to describe the set  $\mathcal{P}$ . The algorithms we consider depend on the algebraic description of  $\mathcal{P}$  we choose and different descriptions can lead to different algorithmic guarantees. In general, it is not clear what is the best possible description. However, typically, the more auxiliary variables the better.

[2016], Barak and Moitra [2016], A. Potechin and Steurer [2017], P. K. Kothari, Steinhardt, and Steurer [2018], and Hopkins and Li [2018].

**2.2 Matrix and tensor completion.** In matrix completion, we observe a few entries of a matrix and the goal is to fill in the missing entries. This problem is studied extensively both from practical and theoretical perspectives. One of its practical application is in recommender systems, which was the basis of the famous Netflix Prize competition. Here, we may observe a few movie ratings for each user and the goal is to infer a user’s preferences for movies that the user hasn’t rated yet.

In terms of provable guarantees, the best known polynomial time algorithm for matrix completion is based on a semidefinite programming relaxation. Let  $X = \sum_{i=1}^r \sigma_i \cdot u_i v_i^\top \in \mathbb{R}^{n \times n}$  be a rank- $r$  matrix such that its left and right singular vectors  $u_1, \dots, u_r, v_1, \dots, v_r \in \mathbb{R}^n$  are  $\mu$ -incoherent<sup>4</sup>, i.e., they satisfy  $\langle u_i, e_j \rangle^2 \leq \mu/n$  and  $\langle v_i, e_j \rangle^2 \leq \mu/n$  for all  $i \in [r]$  and  $j \in [n]$ . The algorithm observes the partial matrix  $X_\Omega$  that contains a random cardinality  $m$  subset  $\Omega \subseteq [n] \times [n]$  of the entries of  $X$ . If  $m \geq \mu r n \cdot O(\log n)^2$ , then with high probability over the choice of  $\Omega$  the algorithm recovers  $X$  exactly Candès and Recht [2009], Gross [2011], Recht [2011], and Chen [2015]. This bound on  $m$  is best-possible in several ways. In particular,  $m \geq \Omega(rn)$  appears to be necessary because an  $n$ -by- $n$  rank- $r$  matrix has  $\Omega(r \cdot n)$  degrees of freedom (the entries of its singular vectors).

In this section, we will show how the above algorithm is captured by sum-of-squares and, in particular, Theorem 2.1. We remark that this fact follows directly by inspecting the analysis of the original algorithm Candès and Recht [2009], Gross [2011], Recht [2011], and Chen [2015]. The advantage of sum-of-squares here is two-fold: First, it provides a unified perspective on algorithms for matrix completion and other estimation problems. Second, the sum-of-squares approach for matrix completion extends in a natural way to tensor completion (in a way that the original approach for matrix completion does not).

**Identifiability proof for matrix completion.** For the sake of clarity, we consider a simplified setup where the matrix  $X$  is assumed to be a rank- $r$  projector so that  $X = \sum_{i=1}^r a_i a_i^\top$  for  $\mu$ -incoherent orthonormal vectors  $a_1, \dots, a_r \in \mathbb{R}^n$ . The following theorem shows that, with high probability over the choice of  $\Omega$ , the matrix  $X$  is identified by the partial matrix  $X_\Omega$ . Furthermore, the proof of this fact is captured by sum-of-squares. Together with Theorem 2.1, the following theorem implies that there exists a polynomial-time algorithm to recover  $X$  from  $X_\Omega$ .

<sup>3</sup>In order to be able to state running times in a simple way, we assume that the total bit-complexity of  $(x, y, z)$  and the vector-valued polynomial  $p$  (in the monomial basis) is bounded by a fixed polynomial in  $n$ .

<sup>4</sup> Random unit vectors satisfy this notion of  $\mu$ -incoherence for  $\mu \leq O(\log n)$ . In this sense, incoherent vectors behave similar to random vectors.

**Theorem 2.2** (implicit in [Candès and Recht \[2009\]](#), [Gross \[2011\]](#), [Recht \[2011\]](#), and [Chen \[2015\]](#)). *Let  $X = \sum_{i=1}^r a_i a_i^\top \in \mathbb{R}^{n \times n}$  be an  $r$ -dimensional projector and  $a_1, \dots, a_r \in \mathbb{R}^n$  orthonormal with incoherence  $\mu = \max_{i,j} n \cdot \langle a_i, e_j \rangle^2$ . Let  $\Omega \subseteq [n] \times [n]$  be a random symmetric subset of size  $|\Omega| = m$ . Consider the system of polynomial equations in  $n$ -by- $r$  matrix variable  $B$ ,*

$$\mathcal{Q} = \left\{ (BB^\top)_\Omega = X_\Omega, \quad B^\top B = \text{Id}_r \right\}.$$

*Suppose  $m \geq \mu r n \cdot O(\log n)^2$ . Then, with high probability over the choice of  $\Omega$ ,*

$$\mathcal{Q} \mid \frac{B}{4} \left\{ \|BB^\top - X\|_F = 0 \right\}.$$

*Proof.* The analyses of the aforementioned algorithm for matrix completion [Candès and Recht \[2009\]](#), [Gross \[2011\]](#), [Recht \[2011\]](#), and [Chen \[2015\]](#) show the following: with high probability over the choice of  $\Omega$ , there exists<sup>5</sup> a symmetric matrix  $M$  with  $M_{\bar{\Omega}} = 0$  and  $0.9(\text{Id}_n - X) \leq M - X \leq 0.9(\text{Id}_n - X)$ . As we will see, this matrix also implies that the above proof of identifiability exists.

Since  $0 \leq X$  and  $X - 0.9(\text{Id}_n - X) \leq M$ , we have

$$\langle M, X \rangle \geq \langle X, X \rangle - 0.9 \langle \text{Id}_n - X, X \rangle = \langle X, X \rangle = r.$$

Since  $M_{\bar{\Omega}} = 0$  and  $\mathcal{Q}$  contains the equation  $(BB^\top)_\Omega = X_\Omega$ , we have  $\mathcal{Q} \mid \frac{B}{4} \langle M, BB^\top \rangle = \langle M, X \rangle \geq r$ . At the same time, we have

$$\mathcal{Q} \mid \langle M, BB^\top \rangle \leq \langle X, BB^\top \rangle + 0.9 \langle \text{Id}_n - X, BB^\top \rangle = 0.1 \langle X, BB^\top \rangle + 0.9r,$$

where the first step uses  $M \preceq X + 0.9(\text{Id}_n - X)$  and the second step uses  $\mathcal{Q} \mid \langle \text{Id}_n, BB^\top \rangle = r$  because  $\langle \text{Id}_n, BB^\top \rangle = \text{Tr } B^\top B$  and  $\mathcal{Q}$  contains the equation  $B^\top B = \text{Id}_r$ . Combining the lower and upper bound on  $\langle M, BB^\top \rangle$ , we obtain

$$\mathcal{Q} \mid \langle X, BB^\top \rangle \geq r.$$

Together with the facts  $\|X\|_F^2 = r$  and  $\mathcal{Q} \mid \|BB^\top\|_F^2 = r$ , we obtain  $\mathcal{Q} \mid \|X - BB^\top\|_F^2 = 0$  as desired.  $\square$

<sup>5</sup> Current proofs of the existence of this matrix proceed by an ingenious iterative construction of this matrix (alternatingly projecting to two affine subspaces). The analysis of this iterative construction is based on matrix concentration bounds. We refer to prior literature for details of this proof [Gross \[2011\]](#), [Recht \[2011\]](#), and [Chen \[2015\]](#).

**Identifiability proof for tensor completion.** Tensor completion is the analog of matrix completion for tensors. We observe a few of the entries of an unknown low-rank tensor and the goal is to fill in the missing entries. In terms of provable guarantees, the best known polynomial-time algorithms are based on sum-of-squares, both for exact recovery [A. Potechin and Steurer \[2017\]](#) (of tensors with orthogonal low-rank decompositions) and approximate recovery [Barak and Moitra \[2016\]](#) (of tensors with general low-rank decompositions).

Unlike for matrix completion, there appears to be a big gap between the number of observed entries required by efficient and inefficient algorithms. For 3-tensors, all known efficient algorithms require  $r \cdot \tilde{O}(n^{1.5})$  observed entries (ignoring the dependence on incoherence) whereas information-theoretically  $r \cdot O(n)$  observed entries are enough. The gap for higher-order tensors becomes even larger. It is an interesting open question to close this gap or give formal evidence that the gap is inherent.

As for matrix completion, we consider the simplified setup that the unknown tensor has the form  $X = \sum_{i=1}^r a_i^{\otimes 3}$  for incoherent, orthonormal vectors  $a_1, \dots, a_r \in \mathbb{R}^n$ . The following theorem shows that with high probability,  $X$  is identifiable from  $rn^{1.5} \cdot (\mu \log n)^{O(1)}$  random entries of  $X$  and this fact has a low-degree sum-of-squares proof.

**Theorem 2.3** ([A. Potechin and Steurer \[2017\]](#)). *Let  $a_1, \dots, a_r \in \mathbb{R}^n$  orthonormal vectors with incoherence  $\mu = \max_{i,j} n \cdot \langle a_i, e_j \rangle^2$  and let  $X = \sum_{i=1}^r a_i^{\otimes 3}$  be their 3-tensor. Let  $\Omega \subseteq [n]^3$  be a random symmetric subset of size  $|\Omega| = m$ . Consider the system of polynomial equations in  $n$ -by- $r$  matrix variable  $B$  with columns  $b_1, \dots, b_r$ ,*

$$\mathcal{Q} = \left\{ \left( \sum_{i=1}^r b_i^{\otimes 3} \right)_{\Omega} = X_{\Omega}, \quad B^{\top} B = \text{Id}_r \right\}$$

*Suppose  $m \geq rn^{1.5} \cdot (\mu \log n)^{O(1)}$ . Then, with high probability over the choice of  $\Omega$ ,*

$$\mathcal{Q} \mid_{O(1)} \left\{ \left\| \sum_{i=1}^r b_i^{\otimes 3} - X \right\|_{\text{F}}^2 = 0 \right\}$$

**2.3 Overcomplete tensor decomposition.** Tensor decomposition refers to the following general class of estimation problems: Given (a noisy version of) a  $k$ -tensor of the form  $\sum_{i=1}^r a_i^{\otimes k}$ , the goal is to (approximately) recover one, most, or all of the component vectors  $a_1, \dots, a_r \in \mathbb{R}^n$ . It turns out that under mild conditions on the components  $a_1, \dots, a_r$ , the noise, and the tensor order  $k$ , this estimation task is possible information theoretically. For example, generic components  $a_1, \dots, a_r \in \mathbb{R}^n$  with  $r \leq \Omega(n^2)$  are

identified by their 3-tensor  $\sum_{i=1}^r a_i^{\otimes 3}$  [Chiantini and Ottaviani \[2012\]](#) (up to a permutation of the components). Our concern will be what conditions on the components, the noise, and the tensor order allow us to efficiently recover the components.

Besides being significant in its own right, tensor decomposition is a surprisingly versatile and useful primitive to solve other estimation problems. Concrete examples of problems that can be reduced to tensor decomposition are latent Dirichlet allocation models, mixtures of Gaussians, independent component analysis, noisy-or Bayes nets, and phylogenetic tree reconstruction [Lathauwer, Castaing, and Cardoso \[2007\]](#), [Mossel and Roch \[2005\]](#), [Anandkumar, Foster, Hsu, S. Kakade, and Liu \[2012\]](#), [Hsu and S. M. Kakade \[2013\]](#), [Bhaskara, Charikar, Moitra, and Vijayaraghavan \[2014\]](#), [Barak, Kelner, and Steurer \[2015\]](#), [Ma, Shi, and Steurer \[2016\]](#), and [Arora, Ge, Ma, and Risteski \[2016\]](#). Through these reductions, better algorithms for tensor decomposition can lead to better algorithms for a large number of other estimation problems.

Toward better understanding the capabilities of efficient algorithms for tensor decomposition, we focus in this section on the following more concrete version of the problem.

**Problem 2.4** (Tensor decomposition, one component, constant error). Given an order- $k$  tensor  $\sum_{i=1}^r a_i^{\otimes k}$  with component vectors  $a_1, \dots, a_r \in \mathbb{R}^n$ , find a vector  $u \in \mathbb{R}^n$  that is close<sup>6</sup> to one of the component vectors in the sense that  $\max_{i \in [r]} \frac{1}{\|a_i\| \cdot \|u\|} |\langle a_i, u \rangle| \geq 0.9$ .

Algorithms for [Problem 2.4](#) can often be used to solve a-priori more difficult versions of the tensor decomposition that ask to recover most or all of the components or that require the error to be arbitrarily small.

A classical spectral algorithm attributed to [Harshman \[1970\]](#) and [Leurgans, Ross, and Abel \[1993\]](#) can solve [Problem 2.4](#) for up to  $r \leq n$  generic components if the tensor order is at least 3. (Concretely, the algorithm works for 3-tensors with linearly independent components.) Essentially the same algorithm works up to  $\Omega(n^2)$  generic<sup>7</sup> components if the tensor order is at least 5. A more sophisticated algorithm [Lathauwer, Castaing, and Cardoso \[2007\]](#) solves [Problem 2.4](#) for up to  $\Omega(n^2)$  generic<sup>8</sup> components if the tensor order is at least 4. However, these algorithms and their analyses break down if the tensor order is only 3 and the number of components issue  $n^{1+\Omega(1)}$ , even if the components are random vectors.

In this and the subsequent section, we will discuss a polynomial-time algorithm based on sum-of-squares that goes beyond these limitations of previous approaches.

<sup>6</sup>This notion of closeness ignores the sign of the components. If the tensor order is odd, the sign can often be recovered as part of some postprocessing. If the tensor order is even, the sign of the components is not identified.

<sup>7</sup>Here, the vectors  $a_1^{\otimes 2}, \dots, a_r^{\otimes 2}$  are assumed to be linearly independent.

<sup>8</sup>Concretely, the vectors  $\{a_i^{\otimes 2} \otimes a_j^{\otimes 2} \mid i \neq j\} \cup \{(a_i \otimes a_j)^{\otimes 2} \mid i \neq j\}$  are assumed to be linearly independent.

**Theorem 2.5** (Ma, Shi, and Steurer [2016] building on Barak, Kelner, and Steurer [2015], Ge and Ma [2015], and Hopkins, Schramm, Shi, and Steurer [2016]). *There exists a polynomial-time algorithm to solve Problem 2.4 for tensor order 3 and  $\tilde{\Omega}(n^{1.5})$  components drawn uniformly at random from the unit sphere.*

The strategy for this algorithm consists of two steps:

1. use sum-of-squares in order to lift the given order-3 tensor to a noisy version of the order-6 tensor with the same components,
2. apply Jennrich’s classical algorithm to decompose this order-6 tensor.

While Problem 2.4 falls outside of the scope of Theorem 2.1 (Meta-theorem for efficient estimation) because the components are only identified up to permutation, the problem of lifting a 3-tensor to a 6-tensor with the same components is captured by Theorem 2.1. Concretely, we can formalize this lifting problem as the following set of parameter–measurement pairs,

$$\mathcal{P}_{3,6;r} = \left\{ (x, y) \left| x = \sum_{i=1}^r a_i^{\otimes 6}, y = \sum_{i=1}^r a_i^{\otimes 3}, a_1, \dots, a_r \in \mathbb{R}^n \right. \right\} \subseteq \mathbb{R}^{n^6} \times \mathbb{R}^{n^3}.$$

In Section 2.4, we give the kind of sum-of-squares proofs that Theorem 2.1 requires in order to obtain an efficient algorithm to solve the above estimation problem of lifting 3-tensors to 6-tensors with the same components.

The following theorem gives an analysis of Jennrich’s algorithm that we can use to implement the second step of the above strategy for Theorem 2.5.

**Theorem 2.6** (Ma, Shi, and Steurer [2016] and Schramm and Steurer [2017]). *There exists  $\varepsilon > 0$  and a randomized polynomial-time algorithm that given a 3-tensor  $T \in (\mathbb{R}^n)^{\otimes 3}$  outputs a unit vector  $u \in \mathbb{R}^n$  with the following guarantees: Let  $a_1, \dots, a_r \in \mathbb{R}^n$  be unit vectors with orthogonality defect  $\|\text{Id}_r - A^\top A\| \leq \varepsilon$ , where  $A \in \mathbb{R}^{n \times r}$  is the matrix with columns  $a_1, \dots, a_r$ . Suppose  $\|T - \sum_i a_i^{\otimes 3}\|_F^2 \leq \varepsilon \cdot r$  and  $\max\{\|T\|_{\{1,3\}\{2\}}, \|T\|_{\{1\}\{2,3\}}\} \leq 10$ . Then, with at least inverse polynomial probability,  $\max_{i \in [r]} \langle a_i, u \rangle \geq 0.9$ .*

**2.4 Tensor decomposition: lifting to higher order.** In this section, we give low-degree sum-of-squares proofs of identifiability for the different version of the estimation problem of lifting 3-tensors to 6-tensors with the same components. These sum-of-squares proofs are a key ingredient of the algorithms for overcomplete tensor decomposition discussed in Section 2.3.



We first consider the problem of lifting 3-tensors with orthonormal components. By itself, this lifting theorem cannot be used for overcomplete tensor decomposition. However it turns out that this special case best illustrates the basic strategy for lifting tensors to higher-order tensors with the same components.

**Orthonormal components.** The following lemma shows that for orthonormal components, the 3-tensor identifies the 6-tensor with the same set of components and that this fact has a low-degree sum-of-squares proof.

**Lemma 2.7.** *Let  $a_1, \dots, a_r \in \mathbb{R}^n$  be orthonormal. Let  $\mathcal{Q} = \{\sum_{i=1}^r a_i^{\otimes 3} = \sum_{i=1}^r b_i^{\otimes 3}, B^\top \cdot B = \text{Id}\}$ , where  $B$  is an  $n$ -by- $r$  matrix of variables and  $b_1, \dots, b_r$  are the columns of  $B$ . Then,*

$$\mathcal{Q} \vdash_{\frac{B}{12}} \left\{ \left\| \sum_{i=1}^r a_i^{\otimes 6} - \sum_{i=1}^r b_i^{\otimes 6} \right\|_{\text{F}}^2 = 0 \right\}.$$

*Proof.* By orthonormality,  $\|\sum_{i=1}^r a_i^{\otimes 6}\|_{\text{F}}^2 = \|\sum_{i=1}^r a_i^{\otimes 3}\|_{\text{F}}^2 = r$  and  $\mathcal{Q} \vdash_{\frac{B}{12}} \|\sum_{i=1}^r b_i^{\otimes 6}\|_{\text{F}}^2 = \|\sum_{i=1}^r b_i^{\otimes 3}\|_{\text{F}}^2 = r$ . Thus,  $\mathcal{Q} \vdash \sum_{i,j} \langle a_i, b_j \rangle^3 = r$  it suffices to show  $\mathcal{Q} \vdash \sum_{i,j} \langle a_i, b_j \rangle^6 \geq r$ .

Using  $\sum_{i=1}^r a_i a_i^\top \preceq \text{Id}$ , a sum-of-squares version of Cauchy–Schwarz, and the fact that  $\mathcal{Q}$  contains the constraints  $\|b_1\|^2 = \dots = \|b_r\|^2 = 1$ ,

$$\mathcal{Q} \vdash r = \sum_{i,j} \langle a_i, b_j \rangle^3 \leq \frac{1}{2} \sum_{i,j} \langle a_i, b_j \rangle^2 + \frac{1}{2} \sum_{i,j} \langle a_i, b_j \rangle^4 \leq \frac{1}{2} r + \frac{1}{2} \sum_{i,j} \langle a_i, b_j \rangle^4.$$

We conclude that  $\mathcal{Q} \vdash \sum_{i,j} \langle a_i, b_j \rangle^4 = r$ . Applying the same reasoning to  $\sum_{i,j} \langle a_i, b_j \rangle^4$  instead of  $\sum_{i,j} \langle a_i, b_j \rangle^3$  yields  $\mathcal{Q} \vdash \sum_{i,j} \langle a_i, b_j \rangle^6 = r$  as desired.  $\square$

**Random components.** Let  $a_1, \dots, a_r \in \mathbb{R}^n$  be uniformly random unit vectors with  $r \leq n^{O(1)}$ . Let  $B$  be an  $n$ -by- $r$  matrix of variables and let  $b_1, \dots, b_r$  be the columns of  $B$ . Consider the following system of polynomial constraints

(2-1)

$$\mathcal{B}_\varepsilon = \left\{ \|b_i\|^2 = 1 \forall i, \left\| \sum_{i=1}^r b_i^{\otimes 3} \right\|_{\text{F}}^2 \geq (1 - \varepsilon) \cdot r, \left\| \sum_{i=1}^r b_i^{\otimes 6} \right\|_{\text{F}}^2 \leq (1 + \varepsilon) \cdot r \right\}.$$

With high probability, the vectors  $a_1, \dots, a_r$  satisfy  $\mathcal{B}_\varepsilon$  for  $\varepsilon \leq \tilde{O}(r/n^{1.5})$ . Concretely, with high probability, every pair  $(i, j) \in [r]^2$  with  $i \neq j$  satisfies  $\langle a_i, a_j \rangle^2 \leq \tilde{O}(1/n)$ . Thus,  $\left\| \sum_{i=1}^r b_i^{\otimes 3} \right\|_{\text{F}}^2 = r + \sum_{i \neq j} \langle b_i, b_j \rangle^3 \geq (1 + \tilde{O}(r/n^{1.5})) \cdot r$  and  $\left\| \sum_{i=1}^r b_i^{\otimes 6} \right\|_{\text{F}}^2 \leq (1 + \tilde{O}(r/n^3)) \cdot r$ .

**Lemma 2.8** (implicit in [Ge and Ma \[2015\]](#)). *Let  $\varepsilon > 0$  and  $a_1, \dots, a_r \in \mathbb{R}^n$  be random unit vectors with  $r \leq \varepsilon \cdot \tilde{\Omega}(n^{1.5})$ . Let  $B$  be an  $n$ -by- $r$  matrix of variables,  $b_1, \dots, b_r$  the columns of  $B$ , and  $\mathcal{Q}$  the following system of polynomial constraints,*

$$\mathcal{Q} = \mathbb{B}_\varepsilon \cup \left\{ \sum_{i=1}^r a_i^{\otimes 3} = \sum_{i=1}^r b_i^{\otimes 3} \right\}.$$

*Then,*

$$\mathcal{Q} \mid_{\frac{B}{12}} \left\{ \left\| \sum_{i=1}^r a_i^{\otimes 6} - \sum_{i=1}^r b_i^{\otimes 6} \right\|_F^2 \leq O(\varepsilon) \cdot \left\| \sum_{i=1}^r a_i^{\otimes 6} + \sum_{i=1}^r b_i^{\otimes 6} \right\|_F^2 \right\}.$$

**2.5 Clustering.** We consider the following clustering problem: given a set of points  $y_1, \dots, y_n \in \mathbb{R}^d$ , the goal is to output a  $k$ -clustering matrix  $X \in \{0, 1\}^{n \times n}$  of the points such that the points in each cluster are close to each other as possible. Here, we say that a matrix  $X \in \{0, 1\}^{n \times n}$  is a  $k$ -clustering if there is a partition  $S_1, \dots, S_k$  of  $[n]$  such that  $X_{ij} = 1$  if and only if there exists  $\ell \in [k]$  with  $i, j \in S_\ell$ .

In this section, we will discuss how sum-of-squares allow us to efficiently find clusterings with provable guarantees that are significantly stronger than for previous approaches. For concreteness, we consider in the following theorem the extensively studied special case that the points are drawn from a mixture of spherical Gaussians such that the means are sufficiently separated [Dasgupta \[1999\]](#), [Arora and Kannan \[2001\]](#), [Vempala and Wang \[2004\]](#), [Achlioptas and McSherry \[2005\]](#), [Kalai, Moitra, and Valiant \[2010\]](#), [Moitra and Valiant \[2010\]](#), and [Belkin and Sinha \[2010\]](#). Another key advantage of the approach we discuss is that it continues to work even if the points are not drawn from a mixture of Gaussians and the clusters only satisfy mild bounds on their empirical moment tensors.

**Theorem 2.9** ([Hopkins and Li \[2018\]](#), [P. K. Kothari, Steinhardt, and Steurer \[2018\]](#), and [Diakonikolas, Kane, and Stewart \[2018\]](#)). *There exists an algorithm that given  $k \in \mathbb{N}$  with  $k \leq n$  and vectors  $y_1, \dots, y_n \in \mathbb{R}^d$  outputs a  $k$ -clustering matrix  $X \in \{0, 1\}^{n \times n}$  in quasi-polynomial time  $n + (dk)^{(\log k)^{O(1)}}$  with the following guarantees: Let  $y_1, \dots, y_n$  be a sample from the uniform mixture of  $k$  spherical Gaussians  $N(\mu_1, \text{Id}), \dots, N(\mu_k, \text{Id})$  with mean separation  $\min_{i \neq j} \|\mu_i - \mu_j\| \geq O(\sqrt{\log k})$  and  $n \geq (dk)^{(\log k)^{O(1)}}$ . Let  $X^* \in \{0, 1\}^{n \times n}$  be the  $k$ -clustering matrix corresponding to the Gaussian components (so that  $X_{ij}^* = 1$  if  $y_i$  and  $y_j$  were drawn from the same Gaussian component and  $X_{ij}^* = 0$  otherwise). Then with high probability,*

$$\|X - X^*\|_F^2 \leq 0.1 \cdot \|X^*\|_F^2.$$

We remark that the same techniques also give a sequence of polynomial-time algorithms that approach the logarithmic separation of the algorithm above. Concretely, for every  $\varepsilon > 0$ , there exists an algorithm that works if the mean separation is at least  $O_\varepsilon(k^\varepsilon)$ .

These algorithms for clustering points drawn from mixtures of separated spherical Gaussians constitute a significant improvement over previous algorithms that require separation at least  $O(k^{1/4})$  [Vempala and Wang \[2004\]](#).

**Sum-of-squares approach to learning mixtures of spherical Gaussians.** In order to apply [Theorem 2.1](#), we view the clustering matrix  $X$  corresponding to the Gaussian components as parameter and a “typical sample”  $y_1, \dots, y_n$  of the mixture as measurement. Here, typical means that the empirical moments in each cluster are close to the moments of a spherical Gaussian distribution. Concretely, we consider the following set of parameter-measurement pairs,

$$\mathcal{P}_{k,\varepsilon,\ell} = \left\{ (X, Y) \mid \begin{array}{l} X \text{ is } k\text{-clustering matrix with clusters } S_1, \dots, S_k \subseteq [n] \\ \forall \kappa \in [k]. \left\| \mathbb{E}_{i \in S_\kappa} (1, x_i - \mu_\kappa)^{\otimes \ell} - \mathbb{E}_{x \sim N(0, \text{Id})} (1, x)^{\otimes \ell} \right\|_F \leq \varepsilon \end{array} \right\} \\ \subseteq \{0, 1\}^{n \times n} \times \mathbb{R}^{d \times n},$$

where  $\mu_\kappa = \mathbb{E}_{i \in S_\kappa} x_i$  is the mean of cluster  $S_\kappa \subseteq [n]$ .

It is straightforward to express  $\mathcal{P}_{k,\varepsilon,\ell}$  in terms of a system of polynomial constraints  $\mathcal{Q} = \{p(X, Y, z) = 0\}$ , so that  $\mathcal{P}_{k,\varepsilon,\ell} = \{(X, Y) \mid \exists z. p(X, Y, z) = 0\}$ . [Theorem 2.9](#) follows from [Theorem 2.1](#) using the fact that under the conditions of [Theorem 2.9](#), the following sum-of-squares proof exists with high probability for  $\ell \leq (\log k)^{O(1)}$ ,

$$\mathcal{Q} \mid \frac{X, z}{\ell} \left\{ \|X - X^*\|_F^2 \leq 0.1 \cdot \|X^*\|_F^2 \right\},$$

where  $X^*$  is the ground-truth clustering matrix (corresponding to the Gaussian components).

### 3 Lower bounds

In this section, we will be concerned with showing lower bounds on the minimum degree of sum-of-squares refutations for polynomial systems, especially those arising out of estimation problems.

The turn of the millennium saw several works that rule out degree-2 sum-of-squares refutations for a variety of problems, such as MAX CUT [Feige and Schechtman \[2002\]](#),  $k$ -CLIQUE [Feige and Krauthgamer \[2000\]](#), and SPARSEST CUT [Khot and Vishnoi \[2015\]](#), among others. These works, rather than explicitly taking place in the context of sum-of-squares proofs, were motivated by the desire to show tightness for specific SDP relaxations.

Around the same time, Grigoriev proved *linear* lower bounds on the degree of sum-of-squares refutations for  $k$ -XOR,  $k$ -SAT, and knapsack Grigoriev [2001b,a] (these bounds were later independently rediscovered by Schoenebeck [2008]). Few other lower bounds against SoS were known. Most of the subsequent works (e.g. Tulsiani [2009] and Bhaskara, Charikar, Vijayaraghavan, Guruswami, and Zhou [2012]) built on the  $k$ -SAT lower bounds via reduction; in essence, techniques for proving lower bounds against higher-degree sum-of-squares refutations were ad hoc and few.

In recent years, a series of papers Meka, A. Potechin, and Wigderson [2015], Deshpande and Montanari [2015], and Hopkins, P. Kothari, A. H. Potechin, Raghavendra, and Schramm [2016] introduced higher-degree sum-of-squares lower bounds for  $k$ -CLIQUE, culminating in the work of Barak et al. Barak, Hopkins, Kelner, P. Kothari, Moitra, and A. Potechin [2016]. Barak et al. go beyond proving lower bounds for the  $k$ -CLIQUE problem specifically, introducing a beautiful and general framework for proving SoS lower bounds. Though their work settles the  $k$ -CLIQUE refutation problem in  $\mathbb{G}(n, 1/2)$ , it leaves more questions than answers. In particular, it gives rise to a compelling conjecture, which if proven, would settle the degree needed to refute a broad class of estimation problems, including DENSEST  $k$ -SUBGRAPH, community detection problems, graph coloring, and more. We devote this section to describing the technique of pseudocalibration.

Let us begin by recalling some notation. Let  $\mathcal{P} = \{p_i(x, y) \geq 0\}_{i \in [m]}$  be a polynomial system associated with an estimation problem. The polynomial system is over hidden variables  $x \in \mathbb{R}^n$ , with coefficients that are functions of the measurement/instance variables  $y \in \mathbb{R}^N$ . We will use  $\mathcal{P}_y$  to denote the polynomial system for a fixed  $y$ . Let  $\mathcal{P}$  have degree at most  $d_x$  in  $x$  and degree at most  $d_y$  in  $y$ . If  $\mathfrak{D}_\emptyset$  denotes the null distribution, then  $\mathcal{P}_y$  is infeasible w.h.p. when  $y \sim \mathfrak{D}_\emptyset$ , and we are interested in the minimum degree of sum-of-squares refutation.

**Pseudodensities.** By Claim 1.4, to rule out degree- $d$  sum-of-squares refutations for  $\mathcal{P}_y$ , it is sufficient to construct the dual object namely the pseudoexpectation functional  $\tilde{\mathbb{E}}_y$  with the properties outlined in Section 1.2. However, it turns out to be conceptually cleaner to think about constructing a related object called *pseudodensities* rather than *pseudoexpectation functionals*. Towards defining pseudodensities, we first pick a natural background measure  $\sigma$  for  $x \in \mathbb{R}^n$ . Therefore,  $\mathbb{E}_x$  will denote the expectation over the background measure  $\sigma$ . The choice of background measure itself is not too important, but for the example we will consider, it will be convenient to pick  $\sigma$  to be uniform distribution over  $\{0, 1\}^n$ .

**Definition 3.1.** A function  $\bar{\mu} : \{0, 1\}^n \rightarrow \mathbb{R}$  is a pseudodensity for a polynomial system  $\mathcal{P} = \{p_i(x) \geq 0\}_{i \in [m]}$  if  $\tilde{\mathbb{E}}_{\bar{\mu}} : \mathbb{R}[x]_{\leq d} \rightarrow \mathbb{R}$  defined as follows:

$$\tilde{\mathbb{E}}_{\bar{\mu}}[p(x)] \stackrel{\text{def}}{=} \mathbb{E}_x \bar{\mu}(x) p(x)$$

is a valid pseudoexpectation operator, namely, it satisfies the constraints outlined in [Section 1.2](#).

To show that  $\mathcal{P}_y$  does not admit a degree  $d$  SoS refutation for most  $y \sim \mathfrak{D}_{\emptyset}$ , it suffices for us to show that with high probability over  $y \sim \mathfrak{D}_{\emptyset}$ , we can construct a pseudodensity  $\bar{\mu}_y : \{0, 1\}^n \rightarrow \mathbb{R}$ . More precisely, with high probability over the choice of  $y \sim \mathfrak{D}_{\emptyset}$ , the following must hold:

(3-1)

$$(\text{scaling}) \quad \mathbb{E}_x \bar{\mu}_y(x) dx = 1$$

(3-2)

$$(\text{pos. semidef.}) \quad \mathbb{E}_x q(x)^2 \bar{\mu}_y(x) dx \geq 0 \quad \forall q \in \mathbb{R}[x]_{\leq d}$$

(3-3)

$$(\text{constraints } \mathcal{P}) \quad \mathbb{E}_x p(x) a^2(x) \cdot \bar{\mu}_y(x) dx \geq 0 \quad \forall p \in \mathcal{P}, a \in \mathbb{R}[x], \deg(a^2 \cdot p) \leq d.$$

**3.1 Pseudocalibration.** *Pseudocalibration* is a heuristic for constructing pseudodensities for non-feasible systems in such settings. It was first introduced in [Barak, Hopkins, Kelner, P. Kothari, Moitra, and A. Potechin \[2016\]](#) for the  $k$ -CLIQUE problem, but the heuristic is quite general and can be seen to yield lower bounds for other problems as well (e.g. [Grigoriev \[2001b\]](#) and [Schoenebeck \[2008\]](#)).

At a high level, pseudocalibration leverages the existence of the structured/structured distribution of estimation problem, to construct pseudodistributions. Let  $\mathfrak{g}_*$  denote the joint *structured* distribution over  $y^* \in \{\pm 1\}^N$  and  $x^*$  is sampled from  $\sigma$ , i.e.,  $\mathbb{P}_{\mathfrak{g}_*}\{(x, y)\} = \sigma(x) \cdot \mathbb{P}_{\mathfrak{g}_x}\{y\}$ .

Let us define a joint null distribution  $\mathfrak{g}_{\emptyset}$  on pairs  $(x, y)$  to be

$$\mathfrak{g}_{\emptyset} \stackrel{\text{def}}{=} \sigma \times \mathfrak{D}_{\emptyset}.$$

As we describe pseudocalibration,  $\mathfrak{g}_{\emptyset}$  will serve as the background measure for us. Let  $\mu_* : \{\pm 1\}^N \times \{0, 1\}^n \rightarrow \mathbb{R}^+$  denote the density of the joint structured distribution  $\mathfrak{g}_*$  with respect to the background measure  $\mathfrak{g}_{\emptyset}$ , namely

$$\mu^*(x, y) = \frac{\mathbb{P}_{\mathfrak{g}_*}(x, y)}{\mathbb{P}_{\mathfrak{g}_{\emptyset}}(x, y)} = \frac{\mathbb{P}_{\mathfrak{D}_*}\{y\}}{\mathbb{P}_{\mathfrak{D}_{\emptyset}}\{y\}} \cdot \frac{\mathbb{P}_{\mathfrak{g}_*}\{x|y\}}{\sigma(x)}$$

At first glance, a candidate construction of pseudodensities  $\bar{\mu}_y$  would be the partially-evaluated relative joint density  $\mu_*$  namely

$$\bar{\mu}_y = \mu_*(y, \cdot).$$

This construction already satisfies two of the three constraints namely Equation (3-2) and Equation (3-3). Note that for any polynomial  $p(x, y)$ ,

$$\mathbb{E}_x p(x) \bar{\mu}_y(x) = \frac{\mathbb{P}_{\mathfrak{D}_* \{y\}}}{\mathbb{P}_{\mathfrak{D}_\emptyset \{y\}}} \cdot \mathbb{E}_x p(x) \frac{\mathbb{P}_{\mathfrak{D}_* \{x|y\}}}{\sigma(x)} = \frac{\mathbb{P}_{\mathfrak{D}_* \{y\}}}{\mathbb{P}_{\mathfrak{D}_\emptyset \{y\}}} \cdot \mathbb{E}_{x \sim \mathfrak{D}_{|y}} p(x).$$

From the above equality, Equation (3-2) follows directly because

$$\mathbb{E}_x q(x)^2 \bar{\mu}_y(x) = \frac{\mathbb{P}_{\mathfrak{D}_* \{y\}}}{\mathbb{P}_{\mathfrak{D}_\emptyset \{y\}}} \cdot \mathbb{E}_{x \sim \mathfrak{D}_{|y}} q^2(x) \geq 0.$$

Similarly, Equation (3-3) is again an immediate consequence of the fact that  $\mathfrak{Y}$  is supported on feasible pairs for  $\mathcal{P}$ ,

$$\mathbb{E}_x p(x) a^2(x) \bar{\mu}_y(x) = \frac{\mathbb{P}_{\mathfrak{D}_* \{y\}}}{\mathbb{P}_{\mathfrak{D}_\emptyset \{y\}}} \cdot \mathbb{E}_{x \sim \mathfrak{D}_{|y}} p(x) a^2(x) \geq 0.$$

However, the scaling constraint Equation (3-1) is far from satisfied because,

$$\mathbb{E}_x \bar{\mu}_y(x) = \frac{\mathbb{P}_{\mathfrak{D}_* \{y\}}}{\mathbb{P}_{\mathfrak{D}_\emptyset \{y\}}} \cdot \mathbb{E}_{x \sim \mathfrak{D}_{|y}} 1 = \frac{\mathbb{P}_{\mathfrak{D}_* \{y\}}}{\mathbb{P}_{\mathfrak{D}_\emptyset \{y\}}}$$

is a quantity that is really large for  $y \in \text{supp}(\mathfrak{D}_*)$  and 0 otherwise. As a saving grace, the constraint Equation (3-1) is satisfied in expectation over  $y$ , i.e.,

$$\mathbb{E}_{y \sim \mathfrak{D}_\emptyset} \mathbb{E}_x \bar{\mu}_y(x) = \mathbb{E}_{y \sim \mathfrak{D}_\emptyset} \mathbb{E}_x \mu^*(x, y) = \mathbb{E}_{(x, y) \sim \mathfrak{D}_\emptyset} \mu^*(x, y) = 1,$$

since  $\mu^*$  is a density.

The relative joint density  $\mu_*(y, x)$  faces an inherent limitation in that it is only nonzero on  $\text{supp}(\mathfrak{D}_*)$ , which accounts for a negligible fraction of  $y \sim \mathfrak{D}_\emptyset$ . Intuitively, the constraints of  $\mathcal{P}$  are low-degree polynomials in  $x$  and  $y$ . Therefore, our goal is to construct a  $\bar{\mu}_y$  that has the same low-degree structure of  $\mu_*$  but has a much higher entropy a.k.a., its mass is not too all concentrated on a small fraction of instances.

The most natural candidate to achieve this is to just project the joint density  $\mu_*$  in to the space of low-degree polynomials. Formally, let  $L_2(\mathfrak{D}_\emptyset)$  denote the vector space of functions over  $\mathbb{R}^N \times \mathbb{R}^n$  equipped with the inner product  $\langle f, g \rangle_{\mathfrak{D}_\emptyset} = \mathbb{E}_{y \sim \mathfrak{D}_\emptyset} f(x, y) g(x, y)$ . For  $d_x, D_y \in \mathbb{N}$ , let  $V_{d_x, D_y} \subseteq L_2(\mathfrak{D}_\emptyset)$  denote the following vector space

$$V_{d_x, D_y} = \text{span}\{q(x, y) \in \mathbb{R}[x, y] \mid \deg_x(q) \leq d_x, \deg_y(q) \leq D_y\}$$

If  $\Pi_{d_x, D_y}$  denote the projection on to  $V_{d_x, D_y}$ , then the pseudo-calibration recipe suggests the use of the following pseudodistribution:

$$(3-4) \quad (\text{Pseudo-calibration}) \quad \bar{\mu}_y(x) = \Pi_{d_x, D_y} \circ \mu_*(x, y)$$

where  $d_x$  is the target degree for the pseudodistribution and  $D_y \in \mathbb{N}$  is to be chosen sufficiently large given  $d_x$ .

Consider a constraint  $\{p(x, y) \geq 0\} \in \mathcal{P}$  in the polynomial system. As long as  $\deg_x(p) \leq d$  and  $\deg_y(p) \leq D_y$ , the pseudodensity  $\bar{\mu}_y$  satisfies the constraint in expectation over  $y$ . This is immediate from the following calculation,

$$\begin{aligned} \mathbb{E} \bar{\mu}_y(x) p(x, y) &= \mathbb{E}_{(x, y) \sim \mathfrak{g}_\emptyset} (\Pi_{d_x, D_y} \circ \mu_*(x, y)) p(x, y) \\ &= \mathbb{E}_{(x, y) \sim \mathfrak{g}_\emptyset} \mu_*(x, y) p(x, y) \\ &= \mathbb{E}_{(x, y) \sim \mathfrak{g}_*} p(x, y) \geq 0. \end{aligned}$$

We require that the constraints are satisfied for each  $y \sim \mathfrak{D}_\emptyset$ , rather than in expectation. Under mild conditions on the joint distribution  $\mathfrak{g}_*$ , the pseudocalibrated construction satisfies the constraints approximately with high probability over  $y \in \mathfrak{D}_\emptyset$ . Specifically, the following theorem holds.

**Theorem 3.2.** *Suppose  $\{p(x, y) \geq 0\} \in \mathcal{P}$  is always satisfied for  $(x, y) \sim \mathfrak{g}_*$  and let  $B := \max_{(x, y) \in \mathfrak{g}_\emptyset} p(x, y)$  and let  $d_y := \deg_y(p)$ . If  $\bar{\mu}_y$  is the pseudocalibrated pseudodensity as defined in Equation (3-4) then*

$$\mathbb{P}_{y \in \mathfrak{D}_\emptyset} [\mathbb{E}_x p(x, y) \bar{\mu}_y(x) \leq -\varepsilon] \leq \frac{B^2}{\varepsilon^2} \|\Pi_{d, D_y + 2d_y} \circ \mu_* - \Pi_{d, D_y} \circ \mu_*\|_{2, \mathfrak{g}_\emptyset}^2$$

where  $\Pi_{d, D}$  is the projection on to span of polynomials of degree at most  $D$  in  $y$  and degree  $d$  in  $x$ .

The theorem suggests that if the projection of the structured density  $\mu_*$  decays with increasing degree then the pseudocalibrated density  $\bar{\mu}_y$  satisfies the same constraints as those satisfied by  $\mu_*$ , with high probability. This decay in the Fourier spectrum of the structured density is a common feature in all known applications of pseudocalibration. We defer the proof of the Theorem 3.2 to the full version.

**Verifying non-negativity of squares.** In light of Theorem 3.2, the chief obstacle in establishing  $\bar{\mu}(y, \cdot)$  as a valid pseudodensity is in proving that it satisfies the constraint  $\mathbb{E}_{\mathfrak{g}_\emptyset} p(y, x)^2 \bar{\mu}(y, x) dx \geq 0$ , for every polynomial  $p$  of degree at most  $\frac{d}{2}$  in  $x$ . As we will

see in [Claim 3.3](#), this condition is equivalent to establishing the positive-semidefiniteness (PSDness) of the matrix

$$M_d(y) \stackrel{\text{def}}{=} \mathbb{E}_x \left( x^{\leq d/2} \right) \left( x^{\leq d/2} \right)^\top \cdot \bar{\mu}(y, x) dx,$$

where  $x^{\leq d/2}$  is the  $O(n^{d/2}) \times 1$  vector whose entries contain all monomials in  $x$  of degree at most  $d/2$ .

*Claim 3.3.*  $\mathbb{E}_{\mathfrak{g}_\emptyset} q(y, x)^2 \bar{\mu}(y, x) \geq 0$  for all polynomials  $q(y, x)$  of degree at most  $d/2$  in  $x$  if and only if the matrix  $M(y) \stackrel{\text{def}}{=} \mathbb{E}[(x^{\leq d/2})(x^{\leq d/2})^\top](y)$  is positive semidefinite.

*Proof.* The first direction is given by expressing  $q(y, x)$  with its vector of coefficients of monomials of  $x$ ,  $\hat{q}(y)$ , so that  $\langle \hat{q}(y), x^{\leq d/2} \rangle = q(y, x)$ . Then

$$\mathbb{E}_{\mathfrak{g}_\emptyset} q(y, x)^2 \bar{\mu}(y, x) = \mathbb{E}_{\mathfrak{g}_\emptyset} [\hat{q}(y)^\top (x^{\leq d/2})(x^{\leq d/2})^\top \hat{q}(y)] = \hat{q}(y)^\top M(y) \hat{q}(y) \geq 0,$$

by the positive-semidefiniteness of  $M(y)$ .

We now prove the contrapositive: if  $M(y)$  is not positive-semidefinite, then there is some negative eigenvector  $v(y)$  so that  $v(y)^\top M(y) v(y) < 0$ . Taking  $q(y, x) = \langle v(y), x^{\leq d/2} \rangle$ , we have our conclusion.  $\square$

Each entry of  $M_d(y)$  is a degree- $D$  polynomial in  $y \sim \mathfrak{D}_\emptyset$ . Since the entries of  $M_d(y)$  are not independent, and because  $M_d(y)$  cannot be decomposed easily into a sum of independent random matrices, standard black-box matrix concentration arguments such as matrix Chernoff bounds and Wigner-type laws do not go far towards characterizing the spectrum of  $M_d(y)$ . This ends up being a delicate and involved process, and the current proofs are very tailored to the specific choice of  $\mathfrak{D}_\emptyset$ , and in some cases they are quite technical.

**3.1.1 Pseudocalibration: a partial answer, and many questions.** While [Theorem 3.2](#) establishes some desirable properties for  $\mu$ , we are left with many unanswered questions. Ideally, we would be able to identify simple, general sufficient conditions on the structured distribution  $\mathfrak{D}_*$  and on  $d$  the degree in  $x$  and  $D$  the degree in  $y$ , for which the answer to the above questions is affirmative. The following conjecture stipulates one such choice of conditions:

**Conjecture 3.4.** *Suppose that  $\mathcal{P}$  contains no polynomial of degree more than  $k$  in  $y$ . Let  $D = O(kd \log n)$  and  $D = \Omega(kd)$ . Then the  $D$ -pseudocalibrated function  $\bar{\mu}(y, \cdot)$  is with high probability a valid degree- $d$  pseudodistribution which satisfies  $\mathcal{P}$  if and only if there is no polynomial  $q(y)$  of degree  $D$  in  $y$  such that*

$$\mathbb{E}_{y \sim \mathfrak{D}_\emptyset} [q(y)] < n^{O(d)} \cdot \mathbb{E}_{y \sim \mathfrak{D}_*} [q(y)].$$



The upper and lower bounds on  $D$  stated in [Conjecture 3.4](#) may not be precise; what is important is that  $D$  not be too much larger than  $O(kd)$ . In support of this conjecture, we list several refutation problems for which the conjecture has been proven:  $k$ -CLIQUE [Barak, Hopkins, Kelner, P. Kothari, Moitra, and A. Potechin \[2016\]](#), TENSOR PCA [Hopkins, P. K. Kothari, A. Potechin, Raghavendra, Schramm, and Steurer \[2017\]](#), and random  $k$ -SAT and  $k$ -XOR [Grigoriev \[2001b\]](#) and [Schoenebeck \[2008\]](#). However, in each of these cases, the proofs have been somewhat ad hoc, and do not generalize well to other problems of interest, such as densest- $k$ -subgraph, community detection, and graph coloring.

Resolving this conjecture, which will likely involve discovering the “book” proof of the above results, is an open problem which we find especially compelling.

**Variations.** The incompleteness of our understanding of the pseudocalibration technique begs the question, is there a different choice of function  $\mu'(y, x)$  such that  $\mu'(y, \cdot)$  is a valid pseudodensity satisfying  $\mathcal{P}$  with high probability over  $y \sim \mathfrak{D}_\emptyset$ ? Indeed, already among the known constructions there is some variation in the implementation of the low-degree projection: the truncation threshold is not always a sharp degree  $D$ , and is sometimes done in a gradual fashion to ease the proofs (see e.g. [Barak, Hopkins, Kelner, P. Kothari, Moitra, and A. Potechin \[2016\]](#)). It is a necessary condition that  $\mu'$  and  $\mu_*$  agree at least on the moments of  $y$  which span the constraints of  $\mathcal{P}$ . However, there are alternative ways to ensure this, while also choosing  $\mu'$  to have higher entropy than  $\mu_*$ .

In [Hopkins, P. K. Kothari, A. Potechin, Raghavendra, Schramm, and Steurer \[2017\]](#), the authors give a different construction, in which rather than projecting  $\mu_*$  to the span of low-degree polynomials in  $y$ , they choose the function  $\mu'$  which minimizes energy under the constraint that  $\int x^{\otimes d} \mu'(y, x) dx$  is positive semidefinite for every  $y \in \text{supp}(\mathfrak{D}_\emptyset)$ , and that  $\mathbb{E}_y \mu'(y, x) p(y, x) = \mathbb{E}_y \mu_*(y, x) p(y, x)$  for every  $p(y, x)$  of degree at most  $D$  in  $y$ . Though in [Hopkins, P. K. Kothari, A. Potechin, Raghavendra, Schramm, and Steurer \[ibid.\]](#) this did not lead to unconditional lower bounds, it was used to obtain a characterization of sum-of-squares algorithms in terms of spectral algorithms.

**3.2 Example:  $k$ -CLIQUE.** In the remainder of this section, we will work out the pseudocalibration construction for the  $k$ -CLIQUE problem (see [Example 1.1](#) for a definition). We’ll follow the outline of the pseudocalibration recipe laid out in [Equation \(3-4\)](#), filling in the blanks as we go along.

**The null and structured distributions.** Recall that  $\mathfrak{D}_\emptyset$  is the uniform distribution over the hypercube  $\{\pm 1\}^{\binom{[n]}{2}}$ , corresponding to  $\mathbb{G}(n, 1/2)$ . For  $\mathfrak{J}_*$  we use the joint distribution over tuples of instance and solution variables  $(y^*, x^*)$  described in [Example 1.1](#), with a small twist designed to ease calculations: Rather than sampling  $x^*$  from  $\pi$  the uniform

distribution over the indicators  $\mathbf{1}_S \in \{0, 1\}^n$  for  $|S| = k$ , we sample  $x^*$  by choosing every coordinate to be 1 with probability  $\frac{2k}{n}$ , and 0 otherwise.

**Pseudomoments.** Instead of directly constructing the pseudodensity  $\bar{\mu}$ , it will be more convenient for us to work with the *pseudomoments*. So for each monomial  $x^A$  where the multiset  $A \subset [n]$  has cardinality at most  $d$ , we will directly define the function  $\tilde{\mathbb{E}}_{\bar{\mu}(y)}[x^A] : \{\pm 1\}^{\binom{[n]}{d}} \rightarrow \mathbb{R}$ . For convenience, and to emphasize the dependence on  $y$ , we will equivalently write  $\tilde{\mathbb{E}}[x^A](y)$ .

Let  $\mathcal{Q} \subseteq E^{\leq D}$  be a set of subsets of edges of cardinality at most  $D$  (we will specify  $D$  later). Following the pseudocalibration recipe from Equation (3-4), for each  $\alpha \in \mathcal{Q}$  we will set

$$\mathbb{E}_{y \sim \mathfrak{D}_\emptyset} \left[ y^\alpha \cdot \tilde{\mathbb{E}}[x^A](y) \right] = \sum_{y \in \{\pm 1\}^E} \int x^A \cdot \mu_*(y, x) dx = \mathbb{E}_{(y, x) \sim \mathfrak{D}_*} [y^\alpha x^A].$$

The right-hand side can be simplified further. For  $(y, x) \sim \mathfrak{D}_*$ , if any vertices of  $A$  are not chosen to be in the clique, then  $x^A$  is zero. Similarly, if any edge  $e \in \alpha$  has an endpoint not in the clique, then  $y^{\{e\}}$  is independent of  $y^{A \setminus \{e\}}$  and of expectation 0. Thus, the expression is equal to the probability that all vertices of  $\alpha$  and  $A$ , which we denote  $v(\alpha) \cup A$ , are contained in the clique:

$$\mathbb{E}_{(y, x) \sim \mathfrak{D}_*} [x^A y^\alpha] = \mathbb{P}_{x \sim \mathfrak{D}_*} [x_i = 1, \forall i \in v(\alpha) \cup A] = \left( \frac{2k}{n} \right)^{|v(\alpha) \cup A|}.$$

For convenience, we will let  $\lambda \stackrel{\text{def}}{=} \left( \frac{2k}{n} \right)$ . Now expressing  $\tilde{\mathbb{E}}[x^A](y)$  via its Fourier decomposition, we have

$$(3-5) \quad \tilde{\mathbb{E}}(y)[x^A] = \sum_{\alpha \in \mathcal{Q}} \left( \frac{2k}{n} \right)^{|v(\alpha) \cup A|} \cdot y^\alpha.$$

## 4 Connection to Spectral Algorithms

Sum-of-squares SDPs yield a systematic framework that capture and generalize a loosely defined class of algorithms often referred to as *spectral algorithms*. The term “spectral algorithm” refers to an algorithm that on an input  $x$  associates a matrix  $M(x)$  that can be easily computed from  $x$  and whose eigenvalues and eigenvectors manifestly yield a solution to the problem at hand. We will give a more concrete definition for the notion of a spectral algorithms a little later in this section.

Although spectral algorithms are typically subsumed by the sum-of-squares SDPs, the spectral versions tend to be simpler to implement and more efficient. Furthermore, in many cases such as the  $k$ -CLIQUE [Alon, Krivelevich, and Sudakov \[1998\]](#) and tensor decomposition [Harshman \[1970\]](#), the first algorithms discovered for the problem were spectral. From a theoretical standpoint, spectral algorithms are much simpler to study and could serve as stepping stones to understanding the limits of sum-of-squares SDPs.

In the worst case, sum-of-squares SDPs often yield strictly better guarantees than corresponding spectral algorithms. For instance, the Goemans-Williamson SDP yields a 0.878 approximation for MAX CUT [Goemans and Williamson \[1995\]](#), has no known analogues among spectral algorithms. Contrary to this, in many random settings, the best known sum-of-squares SDP algorithms yield guarantees that are no better than the corresponding spectral algorithms. Recent work explains this phenomena by showing an equivalence between spectral algorithms and their sum-of-squares SDP counterparts for a broad family of problems [Hopkins, P. K. Kothari, A. Potechin, Raghavendra, Schramm, and Steurer \[2017\]](#). To formally state this equivalence, we will need a few definitions. Let us begin by considering a classic example of a spectral algorithm for the  $k$ -CLIQUE problem. In a graph  $G = (V, E)$ , if a subset  $S \subset V$  of  $k$  vertices forms a clique then,

$$\left\langle 1_S, \left( A_G - \frac{J}{2} \right) 1_S \right\rangle = \frac{k(k-2)}{2}.$$

where  $J \in \mathbb{R}^{n \times n}$  denotes the  $n \times n$  matrix consisting of all ones. On the other hand, we can upper bound the right hand side by

$$\left\langle 1_S, \left( A_G - \frac{J}{2} \right) 1_S \right\rangle \leq \|1_S\|_2^2 \|A_G - \frac{J}{2}\|_{\text{op}} = k \cdot \lambda_{\max} \left( A_G - \frac{J}{2} \right).$$

thereby certifying an upper bound on the size of the clique  $k$ , namely,

$$k \leq 2\lambda_{\max} \left( A_G - \frac{J}{2} \right) + 2.$$

In particular, for a graph  $G$  drawn from the null distribution namely, Erdős-Rényi distribution  $G(n, \frac{1}{2})$ , the matrix  $A_G - \frac{J}{2}$  is a random matrix whose entries are i.i.d uniformly over  $\{\pm \frac{1}{2}\}$ . By Matrix Chernoff inequality [Tropp \[2015\]](#), we will have that  $\lambda_{\max}(A_G - J/2) = O(\sqrt{n})$  with high probability. Thus one can certify an upper bound of  $O(\sqrt{n})$  on the size of the clique in a random graph drawn from  $G(n, \frac{1}{2})$  by computing the largest eigenvalue of the associated matrix valued function  $P(G) = A_G - \frac{1}{2}J$ .

**Injective tensor norm.** Recall that the injective tensor norm (see [Example 1.2](#)) of a symmetric 4-tensor  $T \in \mathbb{R}^{[n] \times [n] \times [n] \times [n]}$  is given by  $\max_{\|x\| \leq 1} \langle x^{\otimes 4}, T \rangle$ . The injective

tensor norm  $\|T\|_{\text{inj}}$  is computationally intractable in the worst case [Hillar and Lim \[2013\]](#). We will now describe a sequence of spectral algorithms that certify tighter bounds for the injective tensor norm of a tensor  $T$  drawn from the null distribution, namely a tensor  $T$  whose entries are i.i.d Gaussian random variables from  $N(0, 1)$ .

Let  $T_{2,2}$  denotes the  $n^2 \times n^2$  matrix obtained by flattening the tensor  $T$  then,

$$\|T\|_{\text{inj}} = \operatorname{argmax}_{\|x\|_2 \leq 1} \langle T, x^{\otimes 4} \rangle = \operatorname{argmax}_{\|x\|_2 \leq 1} \langle x^{\otimes 2}, T_{2,2} x^{\otimes 2} \rangle \leq \lambda_{\max}(T_{2,2})$$

Thus  $\lambda_{\max}(T_{2,2})$  is a spectral upper bound on  $\|T\|_{\text{inj}}$ . Since each entry of  $T$  is drawn independently from  $N(0, 1)$ , we have  $\lambda_{\max}(T_{2,2}) \leq O(n)$  with high probability [Tropp \[2015\]](#). Note that the injective norm of a random  $N(0, 1)$  tensor  $T$  is at most  $O(\sqrt{n})$  with high probability [Tomioka and Suzuki \[2014\]](#) and [Montanari and Richard \[2014\]](#). In other words,  $\lambda_{\max}(T_{2,2})$  certifies an upper bound that is  $n^{1/2}$ -factor approximation to  $\|T\|_{\text{inj}}$ . We will now describe a sequence of improved approximations to the injective tensor norm via spectral methods. Fix a positive integer  $k \in \mathbb{N}$ . The polynomial  $T(x) = \langle x^{\otimes 4}, T \rangle$  can be written as,

$$T(x) = \langle x^{\otimes 2}, T_{2,2} x^{\otimes 2} \rangle = \langle x^{\otimes 2k}, T_{2,2}^{\otimes k} x^{\otimes 2k} \rangle^{1/k}.$$

The tensor  $x^{\otimes 2k}$  is symmetric, and is invariant under permutations of its modes. Let  $\Sigma_{2k}$  denote the set of all permutations of  $\{1, \dots, 2k\}$ . For a permutation  $\Pi \in \Sigma_{2k}$  and a  $2k$ -tensor  $A \in \mathbb{R}^{[n]^{2k}}$ , let  $\Pi \circ A$  denote the  $2k$ -tensor obtained by applying the permutation  $\Pi$  to the modes of  $A$ . By averaging over all permutations  $\Pi, \Pi' \in \Sigma_{2k}$ , we can write

$$\begin{aligned} T(x) &= \left( \mathbb{E}_{\Pi, \Pi' \in \Sigma_{2k}} \langle \Pi \circ x^{\otimes 2k}, T_{2,2}^{\otimes k} (\Pi' \circ x^{\otimes 2k}) \rangle \right)^{1/2k} \\ &= \left( \langle x^{\otimes 2k}, \left( \mathbb{E}_{\Pi, \Pi' \in \Sigma_{2k}} \Pi \circ T_{2,2}^{\otimes k} \circ \Pi' \right) x^{\otimes 2k} \rangle \right)^{1/2k} \\ (4-1) \quad &\leq \lambda_{\max} \left( \mathbb{E}_{\Pi, \Pi' \in \Sigma_{2k}} \Pi \circ T_{2,2}^{\otimes k} \circ \Pi' \right)^{1/2k} \cdot \|x\|_2^2. \end{aligned}$$

Therefore for every  $k \in \mathbb{N}$ , if we denote

$$P_k(T) \stackrel{\text{def}}{=} \mathbb{E}_{\Pi, \Pi' \in \Sigma_{2k}} \Pi \circ T_{2,2}^{\otimes k} \circ \Pi'$$

then  $\|T\|_{\text{inj}} \leq \lambda_{\max}(P_k(T))^{1/k}$ .

The entries of  $P_k(T)$  are degree  $k$  polynomials in the entries of  $T$ . For example, a generic entry of  $P_2(T)$  looks like,

$$P_2(T)_{ijk\ell, i'j'k'\ell'} =$$

$$= \frac{1}{(4!)^2} \cdot (T_{ij i' j'} \cdot T_{k \ell k' \ell'} + T_{ij i' k'} \cdot T_{k \ell j' \ell'} + T_{ij i' \ell'} \cdot T_{k \ell j' k'} + \dots 4!^2 \text{ terms } \dots) .$$

Thus a typical entry of  $P_k(T)$  with no repeated indices is an average of a super-exponentially large number, say  $N_k$ , of i.i.d. random variables. This implies that the variance of a typical entry of  $P_k(T)$  is equal to  $\frac{1}{N_k}$ . For the moment, let us assume that the spectrum of  $P_k(T)$  has a distribution that is similar to that of a random matrix with i.i.d. Gaussian entries with variance  $\frac{1}{N_k}$ . Then,  $\lambda_{\max}(P_k(T)) \leq O(n^k \cdot \frac{1}{N_k^{1/2}})$  with high probability, certifying that  $\|T\|_{\text{inj}} \leq \frac{n}{N_k^{1/2k}}$ . On accounting for the symmetries of  $T$ , it

is easy to see that  $N_k = k! \left( \frac{1}{2^k} \frac{2k!}{k!} \right)^2 \gg (k!)^2$ . Consequently, as per this heuristic argument,  $\lambda_{\max}(P_k(T))$  would certify an upper bound of  $\|T\|_{\text{inj}} \leq O(\frac{n}{k^{3/4}})$ .

Unfortunately, the entries of  $P_k(T)$  are not independent random variables and not all entries of  $P_k(T)$  are typical as described above. Although the heuristic bound on  $\lambda_{\max}(P_k(T))$  is not quite accurate, a careful analysis via the trace method shows that the upper bound  $\lambda_{\max}(P_k(T))^{1/k}$  decreases polynomially in  $k$  [Bhattiprolu, Guruswami, and Lee \[2017\]](#) and [Raghavendra, Rao, and Schramm \[2017\]](#).

**Theorem 4.1.** *Bhattiprolu, Guruswami, and Lee [2017] For  $4 \leq k \leq n^{2/3}$  if  $T$  is a symmetric 4-tensor with i.i.d. entries from a subgaussian measure then*

$$\lambda_{\max}(P_k(T))^{1/k} \leq \tilde{O}\left(\frac{n}{k^{1/2}}\right)$$

*then with probability  $1 - o(1)$ . Here  $\tilde{O}$  notation hides factors polylogarithmic in  $n$ .*

Thus the matrix polynomial  $P_k(T)$  yields a  $n^{O(k)}$ -time algorithm to certify an upper bound of  $\tilde{O}(n/k^{1/2})$  on the injective tensor norm of random 4-tensors with Gaussian entries.

Note that the upper bound certificate produced by the above spectral algorithm can be cast as a degree  $4k$  sum-of-squares proof. In particular, if  $\lambda_{\max}(P_k(T)) \leq B$  for some tensor  $T$  and  $B \in \mathbb{R}$  then,

$$\begin{aligned} B - T(x)^k &= B\|x\|_2^{4k} - \langle x^{\otimes 2k}, P_k(T)x^{\otimes 2k} \rangle + B(1 - \|x\|_2^{4k}) \\ &= \langle x^{\otimes 2k}, (B \cdot \text{Id} - P_k(T))x^{\otimes 2k} \rangle + B(1 - \|x\|_2^{4k}) \\ &= \langle x^{\otimes 2k}, (B \cdot \text{Id} - P_k(T))x^{\otimes 2k} \rangle + (1 - \|x\|_2^2) \left( B \cdot \sum_{i=0}^{2k-1} \|x\|_2^{2i} \right) \\ &= \sum_j s_j^2(x) + (1 - \|x\|_2^2) \left( B \cdot \sum_{i=0}^{2k-1} \|x\|_2^{2i} \right) \end{aligned}$$

The final step in the calculation uses the fact that if a matrix  $M \succeq 0$  is positive semidefinite, then the polynomial  $\langle x^{\otimes 2k}, Mx^{\otimes 2k} \rangle$  is a sum-of-squares. Therefore, the degree  $4k$  sum-of-squares SDP to obtain the same approximation guarantee at least as good as the somewhat adhoc spectral algorithm described above. This is a recurrent theme where the sum-of-squares SDP yields a unified and systematic algorithm that subsumes a vast majority of more adhoc approaches to algorithm design.

**Refuting Random CSPs.** The basic scheme used to upper bound the injective tensor norm (see Equation (4-1)) can be harnessed towards refuting random constraint satisfaction problems (CSPs). Fix a positive integer  $k \in \mathbb{N}$ . In general, a random  $k$ -CSP instance consists of a set of variables  $V$  over a finite domain, and a set of randomly sampled constraints each of which is on a subset of at most  $k$  variables. The problem of refuting random CSPs has been extensively studied for its numerous connections and applications Feige [2002], Ben-Sasson and Bilu [2002], Daniely, Linial, and Shalev-Shwartz [2014], Barak, Kindler, and Steurer [2013], and Crisanti, Leuzzi, and Parisi [2002]. For the sake of concreteness, let us consider the example of random 4-XOR.

**Example 4.2 (4-XOR).** In the 4-XOR problem, the input consists of a linear system over  $\mathbb{F}_2$ -valued variables  $\{X_1, \dots, X_n\}$  such that each equation has precisely 4 variables in it. A random 4-XOR instance is one where each equation is sampled uniformly at random (avoiding repetition). Let  $m$  denote the number of equations, and  $n$  the number of variables. For  $m \gg n$ , with high probability over the choice of the constraints, every assignment satisfies at most  $\frac{1}{2} + o(1)$  fraction of constraints. The goal of refutation algorithm is to certify that there no assignment that satisfies  $\frac{1}{2} + o(1)$  fraction of constraints. To formulate a polynomial system, we will use the natural  $\pm 1$ -encoding of  $\mathbb{F}_2$ , i.e.,  $x_i = 1 \iff X_i = 0$  and  $x_i = -1 \iff X_i = 1$ . An equation of the form  $X_i + X_j + X_k + X_\ell = 0/1$  translates in to  $x_i x_j x_k x_\ell = \pm 1$ . We can specify the instance using a symmetric 4-tensor  $\{T_{ijkl}\}_{i,j,k,\ell \in \binom{[n]}{4}}$ , with  $T_{ijkl} = \pm 1$  if we have the equation  $x_i x_j x_k x_\ell = \pm 1$ , and  $T_{ijk} = 0$  otherwise. To certify that no assignment satisfies more than  $\varepsilon m$  constraints, we will need to refute the following polynomial system.

$$(4-2) \quad \{x_i^2 - 1\}_{i \in [n]} \quad \text{and} \quad \{ \langle T, x^{\otimes 4} \rangle \geq \varepsilon \cdot m \}$$

This system is analogous to the injective tensor norm, except the maximization is over the boolean hypercube  $x \in \{\pm 1\}^n$ , as opposed to the unit ball. Unlike the case of random Gaussian tensors, the tensor  $T$  of interest in 4-XOR is a sparse tensor with about  $n^{1+o(1)}$  non-zero entries. While this poses a few technical challenges, the basic schema from Equation (4-1) can still be utilized to obtain the following refutation algorithm.

**Theorem 4.3.** *Raghavendra, Rao, and Schramm [2017]* For all  $\delta \in [0, 1)$ , the degree  $n^\delta$  sum-of-squares SDP can refute random 4-XOR instances with  $m > \tilde{\Omega}(n^{2-\delta})$  with high probability.

The refutation algorithm for XOR can be used as a building block to obtain sum-of-squares refutations for all random  $k$ -CSPs [Raghavendra, Rao, and Schramm \[ibid.\]](#). Moreover, these bounds on the degree of sum-of-squares refutations tightly match corresponding lower bounds for CSPs shown in [P. K. Kothari, Mori, O’Donnell, and Witmer \[2017\]](#) and [Barak, Chan, and P. K. Kothari \[2015\]](#).

**Defining spectral algorithms.** The above-described algorithms will serve as blue-prints for the class of spectral algorithms that we will formally define now. The problem setup that is most appropriate for our purposes is that of distinguishing problem. Recall that in a distinguishing problem, the input consists of a  $x$  sample drawn from one of two distributions say  $\mathfrak{D}_*$  or  $\mathfrak{D}_\emptyset$  and the algorithm’s goal is to identify the distribution the sample is drawn from. Furthermore, one of the distributions  $\mathfrak{D}_*$  is referred to as the structured distribution is guaranteed to have an underlying hidden structure that is planted within, while samples from the null distribution  $\mathfrak{D}_\emptyset$  typically do not.

A *spectral* algorithm  $\mathcal{Q}$  to distinguish between samples from a structured distribution  $\mathfrak{D}_*$  and a null distribution  $\mathfrak{D}_\emptyset$  proceeds as follows. Given an instance  $x$ , the algorithm  $\mathcal{Q}$  computes a matrix  $P(x)$  whose entries are given by low-degree polynomials in  $x$ , such that  $\lambda_{\max}(P(x)) > 0$  indicates whether  $x \sim \mathfrak{D}_*$  or  $x \sim \mathfrak{D}_\emptyset$ .

**Definition 4.4.** (Spectral Algorithm) A spectral algorithm  $\mathcal{Q}$  consists of a matrix valued polynomial  $P : \mathcal{P} \rightarrow \mathbb{R}^{N \times N}$ . The algorithm  $\mathcal{Q}$  is said to distinguish between samples from structured distribution  $\mathfrak{D}_*$  and a null distribution  $\mathfrak{D}_\emptyset$  if,

$$\mathbb{E}_{y \sim \mathfrak{D}_*} \lambda_{\max}^+(P(y)) \gg \mathbb{E}_{y \sim \mathfrak{D}_\emptyset} \lambda_{\max}^+(P(y))$$

where  $\lambda_{\max}^+(M) \stackrel{\text{def}}{=} \max(\lambda_{\max}(M), 0)$  for a matrix  $M$ .

In general, a spectral algorithm could conceivably use the entire spectrum of the matrix  $P(y)$  instead of the largest eigenvalue, and perform some additional computations on the spectrum. However, a broad range of spectral algorithms can be cast into this framework and as we will describe in this section, this restricted class of spectral algorithms already subsumes the sum-of-squares SDPs in a wide variety of settings.

Spectral algorithms as defined in [Definition 4.4](#) are a simple and highly structured class of algorithms, in contrast to algorithms for solving a sum-of-squares SDP. The feasible region for a sum-of-squares SDP is the intersection of the positive semidefinite cone with polynomially many constraints, some of which are equality constraints. Finding a feasible

solution to the SDP involves an iterated sequence of eigenvalue computations. Furthermore, the feasible solution returned by the SDP solver is by no-means guaranteed to be a low-degree function of the input instance. Instead a spectral algorithm involves exactly one eigenvalue computation of a matrix whose entries are low-degree polynomials in the instance. In spite of their apparent simplicity, we will now argue that they are no weaker than sum-of-squares SDPs for a wide variety of estimation problems.

**Robust Inference.** Many estimation problems share a useful property that we will refer to as “robust inference” property. Specifically, the structured distributions underlying these estimation problems are such that, a randomly chosen subsampling of the instance is sufficient to recover a non-negligible fraction of the planted structure. For example, consider the structured distribution  $\mathfrak{D}_*$  for the  $k$ -CLIQUE problem. A graph  $G \sim \mathfrak{D}_*$  consists of a  $k$ -clique embedded in to a Erdős-Rényi random graph. Suppose we subsample an induced subgraph  $G'$  of  $G$ , by randomly sampling a subset  $S \subset V$  of vertices of size  $|S| = \delta|V|$ . With high probability,  $G'$  contains  $\Omega(\delta \cdot k)$  of the planted clique in  $G$ . Therefore, the maximum clique in  $G'$  yields a clique of size  $\Omega(\delta \cdot k)$  in the original graph  $G$ . This is an example of *robust inference* property, where a random subsample  $G'$  can reveal non-trivial structure in the instance. While the subsample does not determine the planted clique in  $G$ , the information revealed is substantial. For example, as long as  $\delta \cdot k \gg 2 \log n$ ,  $G'$  is sufficient to distinguish whether  $G$  is sampled from the structured distribution  $\mathfrak{D}_*$  or the null distribution  $\mathfrak{D}_\emptyset$ . Moreover, the maximum clique in  $G'$  can be thought of as a feasible solution to a relaxed polynomial system where the clique size sought after is  $\delta \cdot k$ , instead of  $k$ .

Let  $\mathcal{P}$  denote a polynomial system defined on instance variables  $y \in \mathbb{R}^N$  and in solution variables  $x \in \mathbb{R}^n$ . Let  $\Upsilon$  denote the *subsampling distribution* namely, a probability distribution over subsets of instance variables  $[N]$ . Given an instance  $y \in \mathbb{R}^N$ , a subsample  $z$  can be sampled by first picking  $S \sim \Upsilon$  and setting  $z = y_S$ . Let  $\mathfrak{I}$  denote the collection of all instances, and  $\mathfrak{I}_\downarrow$  denote the collection of all sub-instances.

**Definition 4.5.** A polynomial system  $\mathcal{P}$  is  $\varepsilon$ -robustly inferable with respect to a subsampling distribution  $\Upsilon$  and a structured distribution  $\mathfrak{D}_*$ , if there exists a map  $\zeta : \mathfrak{I}_\downarrow \rightarrow \mathbb{R}^n$  such that,

$$\mathbb{P}_{\substack{y \sim \mathfrak{D}_* \\ S \sim \Upsilon}} [\zeta(y_S) \text{ is feasible for } \mathcal{P}] \geq 1 - \varepsilon$$

Robust inference property arises in a broad range of estimation problems including stochastic block models, densest  $k$ -subgraph, tensor PCA, sparse PCA and random CSPs (see Hopkins, P. K. Kothari, A. Potechin, Raghavendra, Schramm, and Steurer [2017] for a detailed discussion). The existence of robust inference property has a stark implication



on the power of low-degree sum-of-squares SDPs, namely they are no more powerful than spectral algorithms. This assertion is formalized in the following theorem.

**Theorem 4.6.** *Suppose  $\mathcal{P} = \{p_i(x, y) \geq 0\}_{i \in [m]}$  is a polynomial system with degree  $d_x$  and  $d_y$  over  $x$  and  $y$  respectively. Fix  $B \geq d_x \cdot d_y \in \mathbb{N}$ . If the degree  $d$  sum-of-squares SDP relaxation can be used to distinguish between the structured distribution  $\mathfrak{D}_*$  and the null distribution  $\mathfrak{D}_\emptyset$ , namely,*

- *For  $y \sim \mathfrak{D}_*$ , the polynomial system  $\mathcal{P}$  is not only satisfiable, but is  $1/n^{8B}$ -robustly inferable with respect to a sub-sampling distribution  $\Upsilon$ .*
- *For  $y \sim \mathfrak{D}_\emptyset$ , the polynomial system  $\mathcal{P}$  is not only infeasible but admits a degree  $d$  sum-of-squares refutation with numbers bounded by  $n^B$  with probability at least  $1 - 1/n^{8B}$ .*

Then, there exists a degree  $2D$  matrix polynomial  $Q : \mathfrak{L} \rightarrow \mathbb{R}^{[n]^{\leq d} \times [n]^{\leq d}}$  such that,

$$\frac{\mathbb{E}_{y \sim \mathfrak{D}_*} [\lambda_{\max}^+(Q(y))]}{\mathbb{E}_{y \sim \mathfrak{D}_\emptyset} [\lambda_{\max}^+(Q(y))]} \geq n^{B/2}$$

where  $D \in \mathbb{N}$  be smallest integer such that for every subset  $\alpha \subset [N]$  with  $|\alpha| \geq D - 2d_x d_y$ ,  $\mathbb{P}_{S \sim \Upsilon}[\alpha \subseteq S] \leq \frac{1}{n^{8B}}$ .

The degree  $D$  of the spectral distinguisher depends on the sub-sampling distribution. Intuitively, the more robustly inferable (a.k.a inferable from smaller subsamples) the problem is, the smaller the degree of the distinguisher  $D$ . For the  $k$ -CLIQUE problem with a clique size of  $n^{1/2-\varepsilon}$ , we have  $D = O(d/\varepsilon)$ . For random CSPs, community detection and densest subgraph we have  $D = O(d \log n)$  (see [Hopkins, P. K. Kothari, A. Potechin, Raghavendra, Schramm, and Steurer \[2017\]](#) for details).

From a practical standpoint, the above theorem shows that sum-of-squares SDPs can often be replaced by their more efficient spectral counterparts. From a theoretical standpoint, it reduces the task of showing lower bounds against the complicated algorithm namely the sum-of-squares SDP to that of understanding the spectrum of low-degree matrix polynomials over the two distributions.

**Future work.** The connection in [Theorem 4.6](#) could potentially be tightened, leading to a fine-grained understanding of the power of sum-of-squares SDPs. We will use a concrete example to expound on the questions laid open by [Theorem 4.6](#), but the discussion is applicable more broadly too.

Consider the problem of certifying an upper bound on the size of maximum independent sets in sparse random graphs. Formally, let  $G$  be a sparse random graph drawn from

$\mathbb{G}(n, k/n)$  by sampling each edge independently with probability  $k/n$ . There exists a constant  $\alpha_k \in (0, 1)$  such that the size of the largest independent set in  $G$  is  $(\alpha \pm o(1)) \cdot n$  with high probability. For every  $\beta \in (0, 1)$ , the existence of a size  $\beta \cdot n$ -independent set can be formulated as the following polynomial system.

$$\mathcal{P}_\beta(G) : \left\{ \begin{array}{l} \{x_i^2 - x_i = 0\}_{i \in [n]}, \quad \{x_i x_j = 0\}_{(i,j) \in E(G)}, \quad \sum_{i \in [n]} x_i \geq \beta \cdot n. \end{array} \right\}$$

For each degree  $d \in \mathbb{N}$  define

$$\alpha_d^{(k)} \stackrel{\text{def}}{=} \text{smallest } \beta \text{ such that } \lim_{n \rightarrow \infty} \mathbb{P}_{G \sim \mathbb{G}([n], k/n)} [\mathcal{P}_\beta \mid \frac{x}{d} \perp] = 1$$

It is natural to ask if the approximation obtained by the degree  $d$  sum-of-squares SDP steadily improves with  $k$ .

**Question 4.7.** Is  $\{\alpha_d^{(k)}\}_{d \in \mathbb{N}}$  a strictly decreasing sequence?

We can associate the following structured distribution  $\mathfrak{g}_\beta$  with the problem. For each subset  $S \in \binom{[n]}{\beta \cdot n}$ , define  $\mu_S$  as  $\mathbb{G}(n, k/n)$  conditioned on  $S$  being an independent set. For  $D \in \mathbb{N}$  define, Let  $\gamma_D^{(k)} \in (0, 1)$  be the largest value of  $\beta$  for which distribution of eigenvalues of low-degree matrix polynomials in the structured distribution  $\mathfrak{g}_\beta$  and null distribution  $\mathfrak{D}_\emptyset$  converge to each other in distribution. In other words,  $\gamma_D^{(k)}$  is the precise threshold of independent set size  $\beta$  below which the structured and the null distributions have same empirical distribution of eigenvalues. It is natural to conjecture that if the empirical distribution of eigenvalues look alike then the sum-of-squares SDP cannot distinguish between the two. Roughly speaking, the conjecture formalizes the notion that sum-of-squares SDPs are no more powerful than spectral algorithms.

**Question 4.8.** Is  $\alpha_d^{(k)} \geq \gamma_{O(d)}^{(k)}$ ?

## References

- Dimitris Achlioptas and Frank McSherry (2005). “On Spectral Learning of Mixtures of Distributions”. In: *COLT*. Vol. 3559. Lecture Notes in Computer Science. Springer, pp. 458–469 (cit. on p. 3423).
- Noga Alon, Michael Krivelevich, and Benny Sudakov (1998). “Finding a large hidden clique in a random graph”. In: *Proceedings of the Eighth International Conference “Random Structures and Algorithms” (Poznan, 1997)*. Vol. 13. 3-4, pp. 457–466. MR: 1662795 (cit. on p. 3432).

- Anima Anandkumar, Dean P. Foster, Daniel J. Hsu, Sham Kakade, and Yi-Kai Liu (2012). “A Spectral Algorithm for Latent Dirichlet Allocation”. In: *NIPS*, pp. 926–934 (cit. on p. 3420).
- Sanjeev Arora, Rong Ge, Tengyu Ma, and Andrej Risteski (2016). “Provable learning of Noisy-or Networks”. *CoRR* abs/1612.08795 (cit. on p. 3420).
- Sanjeev Arora and Ravi Kannan (2001). “Learning mixtures of arbitrary gaussians”. In: *STOC*. ACM, pp. 247–257 (cit. on p. 3423).
- Boaz Barak, Fernando G. S. L. Brandão, Aram Wettroth Harrow, Jonathan A. Kelner, David Steurer, and Yuan Zhou (2012). “Hypercontractivity, sum-of-squares proofs, and their applications”. In: *STOC*. ACM, pp. 307–326 (cit. on p. 3411).
- Boaz Barak, Siu On Chan, and Pravesh K. Kothari (2015). “Sum of squares lower bounds from pairwise independence [extended abstract]”. In: *STOC’15—Proceedings of the 2015 ACM Symposium on Theory of Computing*. ACM, New York, pp. 97–106. MR: 3388187 (cit. on p. 3436).
- Boaz Barak, Samuel B. Hopkins, Jonathan A. Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin (2016). “A Nearly Tight Sum-of-Squares Lower Bound for the Planted Clique Problem”. In: *FOCS*. IEEE Computer Society, pp. 428–437 (cit. on pp. 3413, 3425, 3426, 3430).
- Boaz Barak, Jonathan A. Kelner, and David Steurer (2014). “Rounding sum-of-squares relaxations”. In: *STOC*. ACM, pp. 31–40 (cit. on p. 3416).
- (2015). “Dictionary Learning and Tensor Decomposition via the Sum-of-Squares Method”. In: *STOC*. ACM, pp. 143–151 (cit. on pp. 3416, 3420, 3421).
- Boaz Barak, Guy Kindler, and David Steurer (2013). “On the optimality of semidefinite relaxations for average-case and generalized constraint satisfaction”. In: *ITCS*. ACM, pp. 197–214 (cit. on p. 3435).
- Boaz Barak and Ankur Moitra (2016). “Noisy Tensor Completion via the Sum-of-Squares Hierarchy”. In: *COLT*. Vol. 49. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 417–445 (cit. on pp. 3416, 3417, 3419).
- Boaz Barak and David Steurer (2014). “Sum-of-squares proofs and the quest toward optimal algorithms”. *Electronic Colloquium on Computational Complexity (ECCC)* 21, p. 59 (cit. on p. 3413).
- Mikhail Belkin and Kaushik Sinha (2010). “Toward Learning Gaussian Mixtures with Arbitrary Separation”. In: *COLT*. Omnipress, pp. 407–419 (cit. on p. 3423).
- Eli Ben-Sasson and Yonatan Bilu (2002). “A Gap in Average Proof Complexity”. *Electronic Colloquium on Computational Complexity (ECCC)* 003 (cit. on p. 3435).
- Aditya Bhaskara, Moses Charikar, Ankur Moitra, and Aravindan Vijayaraghavan (2014). “Smoothed analysis of tensor decompositions”. In: *STOC*. ACM, pp. 594–603 (cit. on p. 3420).

- Aditya Bhaskara, Moses Charikar, Aravindan Vijayaraghavan, Venkatesan Guruswami, and Yuan Zhou (2012). “Polynomial integrality gaps for strong SDP relaxations of Densest  $k$ -subgraph”. In: *SODA*. SIAM, pp. 388–405 (cit. on p. [3425](#)).
- Vijay Bhattiprolu, Venkatesan Guruswami, and Euiwoong Lee (2017). “Sum-of-Squares Certificates for Maxima of Random Tensors on the Sphere”. In: *APPROX-RANDOM*. Vol. 81. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, 31:1–31:20 (cit. on p. [3434](#)).
- Emmanuel J. Candès and Benjamin Recht (2009). “Exact Matrix Completion via Convex Optimization”. *Foundations of Computational Mathematics* 9.6, pp. 717–772 (cit. on pp. [3417](#), [3418](#)).
- Yudong Chen (2015). “Incoherence-Optimal Matrix Completion”. *IEEE Trans. Information Theory* 61.5, pp. 2909–2923 (cit. on pp. [3417](#), [3418](#)).
- Luca Chiantini and Giorgio Ottaviani (2012). “On Generic Identifiability of 3-Tensors of Small Rank”. *SIAM J. Matrix Analysis Applications* 33.3, pp. 1018–1037 (cit. on p. [3420](#)).
- Andrea Crisanti, Luca Leuzzi, and Giorgio Parisi (2002). “The 3-SAT problem with large number of clauses in the  $\infty$ -replica symmetry breaking scheme”. *Journal of Physics A: Mathematical and General* 35.3, p. 481 (cit. on p. [3435](#)).
- Amit Daniely, Nati Linial, and Shai Shalev-Shwartz (2014). “From average case complexity to improper learning complexity”. In: *STOC*. ACM, pp. 441–448 (cit. on p. [3435](#)).
- Sanjoy Dasgupta (1999). “Learning Mixtures of Gaussians”. In: *FOCS*. IEEE Computer Society, pp. 634–644 (cit. on p. [3423](#)).
- Yash Deshpande and Andrea Montanari (2015). “Improved Sum-of-Squares Lower Bounds for Hidden Clique and Hidden Submatrix Problems”. In: *COLT*. Vol. 40. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 523–562 (cit. on p. [3425](#)).
- Ilias Diakonikolas, Daniel M. Kane, and Alistair Stewart (2018). “List-Decodable Robust Mean Estimation and Learning Mixtures of Spherical Gaussians Mixture Models, Robustness, and Sum of Squares Proofs”. In: *STOC*. ACM, (to appear) (cit. on p. [3423](#)).
- Uriel Feige (2002). “Relations between average case complexity and approximation complexity”. In: *Proceedings of the Thirty-Fourth Annual ACM Symposium on Theory of Computing*. ACM, New York, pp. 534–543. MR: [2121179](#) (cit. on p. [3435](#)).
- Uriel Feige and Robert Krauthgamer (2000). “Finding and certifying a large hidden clique in a semirandom graph”. *Random Structures Algorithms* 16.2, pp. 195–208. MR: [1742351](#) (cit. on p. [3424](#)).
- Uriel Feige and Gideon Schechtman (2002). “On the optimality of the random hyperplane rounding technique for MAX CUT”. *Random Structures Algorithms* 20.3. Probabilistic methods in combinatorial optimization, pp. 403–440. MR: [1900615](#) (cit. on p. [3424](#)).

- Rong Ge and Tengyu Ma (2015). “Decomposing Overcomplete 3rd Order Tensors using Sum-of-Squares Algorithms”. In: *APPROX-RANDOM*. Vol. 40. LIPIcs. Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, pp. 829–849 (cit. on pp. [3421](#), [3423](#)).
- Sevag Gharibian (2010). “Strong NP-hardness of the quantum separability problem”. *Quantum Information & Computation* 10.3, pp. 343–360 (cit. on p. [3411](#)).
- Michel X. Goemans and David P. Williamson (1995). “Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming”. *J. Assoc. Comput. Mach.* 42.6, pp. 1115–1145. MR: [1412228](#) (cit. on p. [3432](#)).
- Dima Grigoriev (2001a). “Complexity of Positivstellensatz proofs for the knapsack”. *Computational Complexity* 10.2, pp. 139–154 (cit. on p. [3425](#)).
- (2001b). “Linear lower bound on degrees of Positivstellensatz calculus proofs for the parity”. *Theor. Comput. Sci.* 259.1-2, pp. 613–622 (cit. on pp. [3425](#), [3426](#), [3430](#)).
- David Gross (2011). “Recovering Low-Rank Matrices From Few Coefficients in Any Basis”. *IEEE Trans. Information Theory* 57.3, pp. 1548–1566 (cit. on pp. [3417](#), [3418](#)).
- Leonid Gurvits (2003). “Classical deterministic complexity of Edmonds’ Problem and quantum entanglement”. In: *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*. ACM, pp. 10–19 (cit. on p. [3411](#)).
- Richard A Harshman (1970). “Foundations of the PARAFAC procedure: Models and conditions for an” explanatory” multi-modal factor analysis” (cit. on pp. [3420](#), [3432](#)).
- Christopher J. Hillar and Lek-Heng Lim (2013). “Most tensor problems are NP-hard”. *J. ACM* 60.6, Art. 45, 39. MR: [3144915](#) (cit. on p. [3433](#)).
- Sam B. Hopkins and Jerry Li (2018). “Mixture Models, Robustness, and Sum of Squares Proofs”. In: *STOC*. ACM, (to appear) (cit. on pp. [3416](#), [3417](#), [3423](#)).
- Samuel B. Hopkins, Pravesh K. Kothari, Aaron Potechin, Prasad Raghavendra, Tselil Schramm, and David Steurer (2017). “The Power of Sum-of-Squares for Detecting Hidden Structures”. In: *FOCS*. IEEE Computer Society, pp. 720–731 (cit. on pp. [3430](#), [3432](#), [3437](#), [3438](#)).
- Samuel B. Hopkins, Pravesh Kothari, Aaron Henry Potechin, Prasad Raghavendra, and Tselil Schramm (2016). “On the Integrality Gap of Degree-4 Sum of Squares for Planted Clique”. In: *SODA*. SIAM, pp. 1079–1095 (cit. on p. [3425](#)).
- Samuel B. Hopkins, Tselil Schramm, Jonathan Shi, and David Steurer (2016). “Fast spectral algorithms from sum-of-squares proofs: tensor decomposition and planted sparse vectors”. In: *STOC*. ACM, pp. 178–191 (cit. on p. [3421](#)).
- Samuel B. Hopkins, Jonathan Shi, and David Steurer (2015). “Tensor principal component analysis via sum-of-square proofs”. In: *COLT*. Vol. 40. JMLR Workshop and Conference Proceedings. JMLR.org, pp. 956–1006 (cit. on p. [3416](#)).
- Daniel J. Hsu and Sham M. Kakade (2013). “Learning mixtures of spherical gaussians: moment methods and spectral decompositions”. In: *ITCS*. ACM, pp. 11–20 (cit. on p. [3420](#)).

- Adam Tauman Kalai, Ankur Moitra, and Gregory Valiant (2010). “Efficiently learning mixtures of two Gaussians”. In: *STOC*. ACM, pp. 553–562 (cit. on p. 3423).
- Subhash A. Khot and Nisheeth K. Vishnoi (2015). “The unique games conjecture, integrability gap for cut problems and embeddability of negative-type metrics into  $\ell_1$ ”. *J. ACM* 62.1, Art. 8, 39. MR: 3323774 (cit. on p. 3424).
- Pravesh K. Kothari, Ryuhei Mori, Ryan O’Donnell, and David Witmer (2017). “Sum of squares lower bounds for refuting any CSP”. In: *STOC*. ACM, pp. 132–145 (cit. on p. 3436).
- Pravesh K. Kothari, Jacob Steinhardt, and David Steurer (2018). “Robust moment estimation and improved clustering via sum-of-squares”. In: *STOC*. ACM, (to appear) (cit. on pp. 3416, 3417, 3423).
- Jean-Louis Krivine (1964). “Anneaux préordonnés”. *Journal d’analyse mathématique* 12.1, pp. 307–326 (cit. on p. 3408).
- Jean B. Lasserre (2000). “Global optimization with polynomials and the problem of moments”. *SIAM J. Optim.* 11.3, pp. 796–817. MR: 1814045 (cit. on p. 3412).
- Lieven De Lathauwer, Joséphine Castaing, and Jean-François Cardoso (2007). “Fourth-Order Cumulant-Based Blind Identification of Underdetermined Mixtures”. *IEEE Trans. Signal Processing* 55.6-2, pp. 2965–2973 (cit. on p. 3420).
- S. E. Leurgans, R. T. Ross, and R. B. Abel (1993). “A decomposition for three-way arrays”. *SIAM J. Matrix Anal. Appl.* 14.4, pp. 1064–1083. MR: 1238921 (cit. on p. 3420).
- Tengyu Ma, Jonathan Shi, and David Steurer (2016). “Polynomial-Time Tensor Decompositions with Sum-of-Squares”. In: *FOCS*. IEEE Computer Society, pp. 438–446 (cit. on pp. 3416, 3420, 3421).
- Raghu Meka, Aaron Potechin, and Avi Wigderson (2015). “Sum-of-squares Lower Bounds for Planted Clique”. In: *STOC*. ACM, pp. 87–96 (cit. on p. 3425).
- Ankur Moitra and Gregory Valiant (2010). “Settling the Polynomial Learnability of Mixtures of Gaussians”. In: *FOCS*. IEEE Computer Society, pp. 93–102 (cit. on p. 3423).
- Andrea Montanari and Emile Richard (2014). “A statistical model for tensor PCA”. *CoRR* abs/1411.1076 (cit. on p. 3433).
- Elchanan Mossel and Sébastien Roch (2005). “Learning nonsingular phylogenies and hidden Markov models”. In: *STOC*. ACM, pp. 366–375 (cit. on p. 3420).
- Pablo A Parrilo (2000). “Structured semidefinite programs and semialgebraic geometry methods in robustness and optimization”. PhD thesis. California Institute of Technology (cit. on p. 3412).
- Aaron Potechin and David Steurer (2017). “Exact tensor completion with sum-of-squares”. In: *COLT*. Vol. 65. Proceedings of Machine Learning Research. PMLR, pp. 1619–1673 (cit. on pp. 3416, 3417, 3419).

- Prasad Raghavendra, Satish Rao, and Tselil Schramm (2017). “Strongly refuting random CSPs below the spectral threshold”. In: *STOC*. ACM, pp. 121–131 (cit. on pp. [3434](#), [3436](#)).
- Benjamin Recht (2011). “A Simpler Approach to Matrix Completion”. *Journal of Machine Learning Research* 12, pp. 3413–3430 (cit. on pp. [3417](#), [3418](#)).
- Bruce Reznick (2000). “[Some concrete aspects of Hilbert’s 17th Problem](#)”. In: *Real algebraic geometry and ordered structures (Baton Rouge, LA, 1996)*. Vol. 253. Contemp. Math. Amer. Math. Soc., Providence, RI, pp. 251–272. MR: [1747589](#) (cit. on p. [3412](#)).
- Grant Schoenebeck (2008). “Linear Level Lasserre Lower Bounds for Certain k-CSPs”. In: *FOCS*. IEEE Computer Society, pp. 593–602 (cit. on pp. [3425](#), [3426](#), [3430](#)).
- Tselil Schramm and David Steurer (2017). “Fast and robust tensor decomposition with applications to dictionary learning”. In: *COLT*. Vol. 65. Proceedings of Machine Learning Research. PMLR, pp. 1760–1793 (cit. on p. [3421](#)).
- Gilbert Stengle (1974). “A Nullstellensatz and a Positivstellensatz in semialgebraic geometry”. *Mathematische Annalen* 207.2, pp. 87–97 (cit. on p. [3408](#)).
- Ryota Tomioka and Taiji Suzuki (2014). “[Spectral norm of random tensors](#)”. arXiv: [1407.1870](#) (cit. on p. [3433](#)).
- Luca Trevisan (2012). “[On Khot’s unique games conjecture](#)”. *Bull. Amer. Math. Soc. (N.S.)* 49.1, pp. 91–111. MR: [2869009](#) (cit. on p. [3413](#)).
- Joel A. Tropp (2015). “An Introduction to Matrix Concentration Inequalities”. *Foundations and Trends in Machine Learning* 8.1-2, pp. 1–230 (cit. on pp. [3432](#), [3433](#)).
- Madhur Tulsiani (2009). “CSP gaps and reductions in the lasserre hierarchy”. In: *STOC*. ACM, pp. 303–312 (cit. on p. [3425](#)).
- Santosh Vempala and Grant Wang (2004). “A spectral algorithm for learning mixture models”. *J. Comput. Syst. Sci.* 68.4, pp. 841–860 (cit. on pp. [3423](#), [3424](#)).

Received 2018-02-27.

PRASAD RAGHAVENDRA  
U. C. BERKELEY  
[nrprasad@gmail.com](mailto:nrprasad@gmail.com)

TSELIL SCHRAMM  
MIT AND HARVARD  
[tschramm@cs.berkeley.edu](mailto:tschramm@cs.berkeley.edu)

DAVID STEURER  
ETH ZÜRICH  
[dsteurer@gmail.com](mailto:dsteurer@gmail.com)

# LOWER BOUNDS FOR SUBGRAPH ISOMORPHISM

BENJAMIN ROSSMAN

## Abstract

We consider the problem of determining whether an Erdős–Rényi random graph contains a subgraph isomorphic to a fixed pattern, such as a clique or cycle of constant size. The computational complexity of this problem is tied to fundamental open questions including  $P$  vs.  $NP$  and  $NC^l$  vs.  $L$ . We give an overview of unconditional average-case lower bounds for this problem (and its colored variant) in a few important restricted classes of Boolean circuits.

## 1 Background and preliminaries

The *subgraph isomorphism problem* is the computational task of determining whether a “host” graph  $H$  contains a subgraph isomorphic to a “pattern” graph  $G$ . When both  $G$  and  $H$  are given as input, this is a classic  $NP$ -complete problem which generalizes both the MAXIMUM CLIQUE and HAMILTONIAN CYCLE problems [Karp \[1972\]](#). We refer to the  *$G$ -subgraph isomorphism problem* in the setting where the pattern  $G$  is fixed and  $H$  alone is given as input. As special cases, this includes the  $k$ -CLIQUE and  $k$ -CYCLE problems when  $G$  is a complete graph or cycle of order  $k$ .

For patterns  $G$  of order  $k$ , the  $G$ -subgraph isomorphism problem is solvable in time  $O(n^k)$  by the obvious exhaustive search.<sup>1</sup> This upper bound can be improved to  $O(n^{\alpha \lceil k/3 \rceil})$  using any  $O(n^\alpha)$  time algorithm for fast matrix multiplication [Nešetřil and Poljak \[1985\]](#) (the current record has  $\alpha < 2.38$  [Le Gall \[2014\]](#)). Additional upper bounds are tied to structural parameters of  $G$ , such as an  $O(n^{w+1})$  time algorithm for patterns  $G$  of tree-width  $w$  [Plehn and Voigt \[1990\]](#). (See [Marx and Pilipczuk \[2014\]](#) for a survey on upper bounds.)

---

The author’s work is supported by NSERC and a Sloan Research Fellowship.

MSC2010: primary 68Q17; secondary 05C60.

<sup>1</sup>Throughout this article, asymptotic notation ( $O(\cdot)$ ,  $\Omega(\cdot)$ , etc.), whenever bounding a function of  $n$ , hides constants that may depend on  $G$ .



The focus of this article are *lower bounds* which show that the  $G$ -subgraph isomorphism problem cannot be solved with insufficient computational resources. It is conjectured that the  $k$ -CLIQUE problem requires time  $n^{\Omega(k)}$  and that a colored version of  $G$ -subgraph isomorphism (described in [Section 2](#)) requires time  $n^{\Omega(w/\log w)}$  for patterns  $G$  of tree-width  $w$ . Conditionally, these lower bounds are known to follow from the Exponential Time Hypothesis [Chen, Chor, Fellows, Huang, Juedes, Kanj, and Xia \[2005\]](#) and [Marx \[2010\]](#). Proving such lower bounds unconditionally would separate  $P$  from  $NP$  in a very strong way. Since that goal is a long way off, we shall restrict attention to complexity measures much weaker than sequential time; specifically, we focus on restricted classes of Boolean circuits (described in [Section 1.2](#)).

**1.1 The average-case setting.** The lower bounds for the  $G$ -subgraph isomorphism problem described in this article are obtained in the natural average-case setting where the input is an Erdős–Rényi graph  $G_{n,p}$  (or  $G$ -colored version thereof). This is the random  $n$ -vertex graph in which each potential edge is included independently with probability  $p$ . For many patterns of interest including cliques and cycles,  $G_{n,p}$  is conjectured to be a source of hard-on-average instances at an appropriate threshold  $p$ . These conjectures are natural targets for the combinatorial and probabilistic approach of circuit complexity. Strong enough lower bounds for the average-case  $G$ -subgraph isomorphism problem would resolve  $P$  vs.  $NP$  and other fundamental questions, as we explain next.

In the average-case version of the  $k$ -CLIQUE problem, we are given an Erdős–Rényi graph  $G_{n,p}$  at the critical threshold  $p = \Theta(n^{-2/(k-1)})$  (where the existence of a  $k$ -clique occurs with probability bounded away from 0 and 1). Our task is to determine, asymptotically almost surely<sup>2</sup> correctly, whether or not the given graph contains a  $k$ -clique. One natural approach is to make several independent runs of the following *randomized greedy algorithm*: start with a uniform random vertex  $v_1$ , then select a vertex  $v_2$  uniformly at random from among the neighbors of  $v_1$ , next select a vertex  $v_3$  uniformly at random from among the common neighbors  $v_1$  and  $v_2$ , and so on until reaching a maximal (though not necessarily maximum) clique in the given graph. It is easy to show that a single run of the greedy algorithm on  $G_{n,p}$ , which only requires linear time with very high probability, almost surely produces a clique of size  $\lfloor \frac{k}{2} \rfloor$  or  $\lceil \frac{k}{2} \rceil$ . To find a clique of size  $\lfloor \frac{(1+\varepsilon)k}{2} \rfloor$  where  $\varepsilon < 1$ , it suffices to repeat the greedy algorithm  $n^{\varepsilon^2 k/4}$  times, while  $n^{k/4 + O(1/k)}$  iterations suffice to find a  $k$ -clique in  $G_{n,p}$  if any exists. The average-case  $k$ -CLIQUE problem is thus solvable in time  $n^{k/4 + O(1)}$ .

It is unknown whether this iterated greedy algorithm is optimal. In other words, is  $\Omega(n^{k/4})$  a lower bound on the complexity of the average-case  $k$ -CLIQUE problem? This

---

<sup>2</sup>Throughout this article, *asymptotically almost surely* (abbreviated as *a.a.s.*) means with probability  $1 - o(1)$ , that is, with probability that tends to 1 as  $n \rightarrow \infty$ .

question may be seen a scaled-down version of a famous open question of Karp [1976] concerning the uniform random graph  $G_{n,1/2}$ . It is well-known that  $G_{n,1/2}$  has expected maximum clique size  $\approx 2 \log n$ , while the randomized greedy algorithm almost surely finds a clique of size  $\approx \log n$ . Karp asked whether any polynomial-time algorithm a.a.s. succeeds in finding a clique of size  $(1 + \varepsilon) \log n$  for any constant  $\varepsilon > 0$ . Karp's question, together with a variant where  $G_{n,1/2}$  is augmented by a very large planted clique, have stimulated a great deal of research in theoretical computer science. The hardness of detecting planted cliques is used as a cryptographic assumption Juels and Peinado [2000], while lower bounds have been shown against specific algorithms such as the metropolis process Jerrum [1992], the sum-of-squares semidefinite programming hierarchy Barak, Hopkins, Kelner, Kothari, Moitra, and Potechin [2016], and a class of statistical query algorithms Feldman, Grigorescu, Reyzin, Vempala, and Xiao [2013].

The  $k$ -CYCLE problem is another instance where  $G_{n,p}$  at the critical threshold  $p = \Theta(1/n)$  is thought to be a source of hard-on-average instances. Compared to the  $k$ -CLIQUE problem, the average-case  $k$ -CYCLE problem has relatively low complexity: it is solvable in just  $n^{2+o(1)}$  time and moreover in logarithmic space. Nevertheless,  $G_{n,p}$  is believed to be hard-on-average with respect to formula size (a combinatorial complexity measure which we shall discuss shortly). The smallest known formulas solving the  $k$ -CYCLE problem have size  $n^{O(\log k)}$  and this upper bound is conjectured to be optimal even in the average-case. Proving such a lower bound unconditionally would separate complexity classes  $NC^I$  and  $L$ .

**1.2 Circuit complexity.** Circuit complexity is the quest for unconditional lower bounds in combinatorial models of computation. Among such models, Boolean circuits (acyclic networks of  $\wedge$ ,  $\vee$  and  $\neg$  gates) are the most basic and important. Every polynomial-time algorithm can be implemented by a sequence of polynomial-size Boolean circuits, one for each input length  $n$ . To separate  $P$  from  $NP$ , it therefore suffices to prove a super-polynomial lower bound on the minimum circuit size of any problem in  $NP$ , as represented by a sequence of Boolean functions  $\{0, 1\}^n \rightarrow \{0, 1\}$ .

Shannon et al. [1949] showed that *almost all* Boolean functions require circuits of exponential size. Yet after nearly 70 years of efforts, no one has yet proved a super-linear lower bound on the circuit size of any *explicit* Boolean function. In the meantime, the majority of research in circuit complexity has focused on restricted classes of Boolean circuits and other combinatorial models with the aim of developing sharper insights and techniques. Below, we describe three natural and important restricted settings: formulas (tree-like circuits), the  $AC^0$  setting (bounded alternation), and the monotone setting (the absence of negations).

**Definitions.** A *circuit* is a finite directed acyclic graph in which every node of in-degree 0 (“input”) is labeled by a literal (i.e., a variable  $x_i$  or its negation  $\neg x_i$ ), there is a unique node of out-degree 0 (the “output”), and each non-input (“gate”) has in-degree 2 and is labeled by  $\wedge$  or  $\vee$ . Every  $n$ -variable circuit computes a Boolean function  $\{0, 1\}^n \rightarrow \{0, 1\}$  in the obvious way.

The *size* of a circuit is the number of gates it contains. The complexity class  $P/poly$  consists of sequences of Boolean functions  $\{0, 1\}^n \rightarrow \{0, 1\}$  computable by  $n$ -variable circuits of polynomial size (i.e.,  $O(n^c)$  for any constant  $c$ ). (The more familiar class  $P$  is obtained by imposing a uniformity condition on the sequence of  $n$ -variable circuits.)

The *depth* of a circuit is the maximum number of gates on an input-to-output path. The class  $NC^l$  consists of Boolean functions computable by circuits of depth  $O(\log n)$ . Note that  $\text{size}(C) \leq 2^{\text{depth}(C)}$  for all circuits  $C$ , hence  $NC^l \subseteq P/poly$ . This containment is believed but not known to be proper.

The *alternation-depth* of a circuit is the maximum number of alternations between  $\wedge$  and  $\vee$  gates on an input-to-output path. The complexity class  $AC^0$  consists of Boolean functions computed by circuits of polynomial size and constant alternation-depth.<sup>3</sup> Breakthrough lower bounds of the 1980’s showed that  $AC^0$  is a proper subclass of  $NC^l$  [Ajtai \[1983\]](#) and [Furst, Saxe, and Sipser \[1984\]](#). Quantitatively, the strongest of these lower bounds shows that circuits with alternation-depth  $d$  require size  $2^{\Omega(n^{1/(d-1)})}$  to compute the  $n$ -variable PARITY function [Håstad \[1986\]](#).

Another important restricted class of circuits are *formulas*: circuits with the structure of a tree (i.e., in which every non-output node has out-degree 1). In the context of formulas, *size* and *depth* are closely related complexity measures, as every formula of size  $s$  is equivalent to a formula of depth  $O(\log s)$  [Spira \[1971\]](#). As a corollary,  $NC^l$  is equivalent to the class of Boolean functions computed by polynomial-size formulas.

In contrast to circuits, formulas are memoryless in the sense that the result of each sub-computation is only used once. However, despite this obvious weakness, the strongest lower bound on the formula size of an explicit Boolean function is only  $n^{3-o(1)}$  [Håstad \[1998\]](#) and [Tal \[2014\]](#). The challenge of proving a *super-polynomial formula-size lower bound* (i.e., showing that any explicit Boolean function is not in  $NC^l$ ) is one of the major frontiers in circuit complexity.

**1.3 The monotone setting.** Monotonicity is both a property of circuits and a property of Boolean functions. A circuit  $C$  is *monotone* if it has no negations (i.e., inputs are labeled by positive literals only). A Boolean function  $f : \{0, 1\}^n \rightarrow \{0, 1\}$  is *monotone* if  $f(x) \leq$

---

<sup>3</sup> $AC^0$  is usually defined in terms of constant depth circuits with AND and OR gates of unbounded in-degree. In this article, we adopt the equivalent definition in terms of alternation-depth, since the simplest version of our lower bounds naturally applies to binary  $\wedge$  and  $\vee$  gates.

$f(y)$  whenever  $x_i \leq y_i$  for all coordinates  $i$ . Note that the  $G$ -subgraph isomorphism problem is monotone when viewed as a sequence of functions  $\{0, 1\}^{\binom{n}{2}} \rightarrow \{0, 1\}$ .

It is natural to study the *monotone complexity* of monotone functions  $f$  (i.e., the minimum size of a monotone circuit or formula which computes  $f$ ). This has been an extremely fruitful restricted setting in circuit complexity beginning with celebrated results in the 1980's. In a groundbreaking paper which introduced the sunflower-plucking approximation method, Razborov [1985] showed that the  $k$ -CLIQUE problem requires monotone circuits of size  $\Omega(n^k / (\log n)^{2k})$  for any constant  $k$ .<sup>4</sup> By an entirely different technique based on communication complexity, Karchmer and Wigderson [1990] proved an  $n^{\Omega(\log k)}$  lower bound on the size of monotone formulas solving DISTANCE- $k$  ST-CONNECTIVITY, a problem which is equivalent to  $k$ -CYCLE up to a polynomial factor. These results and several others Grigni and Sipser [1992], Pitassi and Robere [2017], Potechin [2010], and Raz and McKenzie [1997] imply essentially all separations  $AC^0 \subset TC^0 \subset NC^1 \subset L \subset NL \subset P \subset NP$  in the monotone world (i.e., for the monotone versions of these classes), whereas in the non-monotone world it is open whether  $TC^0$  (the class of constant-depth threshold circuits) is equal to  $NP$ .

Unfortunately, it is unclear if any of the lower bound techniques developed in the monotone setting have the potential to extend to non-monotone classes. A “barrier” emerges from the observation that essentially all monotone lower bounds in the literature are obtained by pitting a class of *sparse 1-inputs* (e.g., isolated  $k$ -cliques or st-paths) against a class of *dense 0-inputs* (complete  $k - 1$ -partite graphs or st-cuts). In this circumstance, note that the sets of relevant 0- and 1-inputs are separable (in the anti-monotone direction) by a mere threshold function. No monotone lower bound with this property can therefore extend to  $TC^0$ .

This observation motivates the challenge of proving *average-case lower bounds under product distributions* in the monotone setting, in particular for problems like  $k$ -CLIQUE and  $k$ -CYCLE on Erdős–Rényi graphs. This challenge may be seen as a step toward non-monotone lower bounds insofar as product distributions like  $G_{n,p}$  resemble slice distributions like  $G_{n,m}$  (the random graph with exactly  $m$  edges), due to the fact that monotone and non-monotone complexity measures coincide on slice distributions up to a polynomial factor Berkowitz [1982].

**1.4 Outline of the article.** In the rest of this article, we give an overview of lower bounds which characterize the circuit size, as well as the formula size, of the average-case  $G$ -subgraph isomorphism problem in both the  $AC^0$  and monotone settings. The basic

---

<sup>4</sup>Note that this monotone lower bound is quantitatively stronger than the non-monotone  $O(n^{2.73\lceil k/3 \rceil})$  upper bound from fast matrix multiplication. This reveals a gap between monotone vs. non-monotone complexity (see Tardos [1988]).

technique originated in [Rossman \[2008\]](#) where it is shown that  $AC^0$  circuits solving the average-case  $k$ -CLIQUE problem require size  $\Omega(n^{k/4})$ , matching the upper bound from the greedy algorithm. This result improved the previous  $\Omega(n^{k/89d^2})$  lower bound of [Beame \[1990\]](#) for circuits of alternation-depth  $d$ . This is significant for eliminating the dependence on  $d$  in the exponent of  $n$  up to  $O(\log n/k^2 \log \log n)$ , at which point the technique breaks down (though the lower bound is conjectured to hold for unbounded  $d$ ).

[Amano \[2010\]](#) generalized the technique to the  $G$ -subgraph isomorphism problem for arbitrary patterns  $G$  and also gave an extension to hypergraphs. Subsequent work of [Li, Razborov, and Rossman \[2014\]](#) further generalized the technique to a colored variant of the  $G$ -subgraph isomorphism problem, obtaining an  $n^{\Omega(w/\log w)}$  lower bound for patterns of tree-width  $w$ . This result is presented in [Section 4](#).

The challenge of proving stronger lower bounds for formulas was addressed in [Rossman \[2014a\]](#) where it is shown that  $AC^0$  formulas solving the average-case  $k$ -CYCLE problem require size  $n^{\Omega(\log k)}$ . This result sharply separates the power of formulas vs. circuits in the  $AC^0$  setting, since  $k$ -CYCLE is solvable by  $AC^0$  circuits of size  $n^{O(1)}$ . A lower bound for arbitrary patterns  $G$  in terms of tree-depth (a graph invariant akin to tree-width) was subsequently shown using recent results in graph minor theory [Kawarabayashi and Rossman \[2018\]](#). These results are described in [Section 5](#).

These lower bounds in the  $AC^0$  setting apply more generally to any Boolean circuit (or formula) all of whose subcircuits (subformulas) have “low sensitivity with respect to planted subgraphs of  $G$ ” in a certain sense made precise in [Section 3](#). By considering a different notion of “sensitivity”, quantitatively similar lower bounds for *monotone* circuits and formulas are obtained in [Rossman \[2014b, 2015\]](#). For most patterns  $G$ , these lower bounds are merely average-case with respect to a non-product distribution (a convex combination of  $G_{n,p}$  and  $G_{n,p+o(p)}$ ). However, in the special case of the  $k$ -CYCLE problem, the technique produces an average-case lower bound under  $G_{n,p}$ . This is significant for being the first super-polynomial lower bound against monotone formulas under any product distribution.

It is hoped that the framework behind these lower bounds might eventually offer an approach to proving super-polynomial lower bounds for unrestricted Boolean formulas and circuits.

## 2 Colored $G$ -subgraph isomorphism

The main target problem for our lower bounds is actually a colored version of the  $G$ -subgraph isomorphism problem, which we denote by  $\text{SUB}(G)$ . In this problem, the input is a  $G$ -colored graph  $X$  with vertex set  $V(G) \times \{1, \dots, n\}$  and the task to determine whether

$X$  contains a copy of the pattern  $G$  that involves one vertex from each color class. Compared with the previously discussed *uncolored*  $G$ -subgraph isomorphism problem, which we denote by  $\text{SUB}_{\text{uncol}}(G)$ , the colored variant turns out to be better structured and admits a richer class of threshold distributions. All average-case lower bounds for  $\text{SUB}(G)$  in this article extend to the average-case  $\text{SUB}_{\text{uncol}}(G)$  as a special case (as we explain in [Example 2.6](#)).

**Definitions.** All *graphs* in this article are finite simple graphs without isolated vertices. Formally, a graph  $G$  consists of a set  $V(G)$  of vertices and a set  $E(G) \subseteq \binom{V(G)}{2}$  of unordered edges such that  $V(G) = \bigcup_{\{v,w\} \in E(G)} \{v, w\}$ . A *subgraph* of  $G$  is a graph  $H$  such that  $E(H) \subseteq E(G)$  (we simply write  $H \subseteq G$ ). A graph  $G$  thus has  $2^{|E(G)|}$  subgraphs, which are naturally identified with points in the hypercube  $\{0, 1\}^{|E(G)|}$ . An *isomorphism* between graphs  $G$  and  $G'$  is a bijection  $\pi : V(G) \rightarrow V(G')$  such that  $\{v, w\} \in E(G) \Leftrightarrow \{\pi(v), \pi(w)\} \in E(G')$  for all distinct vertices  $v, w$  of  $G$ .

The  $n$ -*blowup* of a graph  $G$ , denoted  $G^{\uparrow n}$ , has vertices  $v^{(1)}, \dots, v^{(n)}$  for each  $v \in V(G)$  and edges  $\{v^{(a)}, w^{(b)}\}$  for each  $\{v, w\} \in E(G)$  and  $a, b \in [n]$  ( $:= \{1, \dots, n\}$ ). We view  $G^{\uparrow n}$  and its subgraphs as “ $G$ -colored graphs” under the vertex-coloring  $v^{(a)} \mapsto v$ .

The *colored  $G$ -subgraph isomorphism problem*, denoted  $\text{SUB}(G)$  for short, is the computational task, given a  $G$ -colored graph  $X \subseteq G^{\uparrow n}$  as input, of determining whether  $X$  contains a subgraph that is isomorphic  $G$  via the map  $v^{(a)} \mapsto v$ . Formally, this problem is represented by a sequence of Boolean functions  $\{0, 1\}^{kn^2} \rightarrow \{0, 1\}$  where  $k = |E(G)|$  and  $kn^2 = |E(G^{\uparrow n})|$ .

Henceforth,  $H$  is always a subgraph of  $G$ , while  $X$  is a subgraph of  $G^{\uparrow n}$ . For an element  $\alpha \in [n]^{V(H)}$ , let  $H^{(\alpha)}$  denote the copy of  $H$  in  $G^{\uparrow n}$  with vertices  $v^{(\alpha_v)}$  for  $v \in V(H)$  and edges  $\{v^{(\alpha_v)}, w^{(\alpha_w)}\}$  for  $\{v, w\} \in E(H)$ . We refer to subgraphs of  $X$  of the form  $H^{(\alpha)}$  as  $H$ -*subgraphs* of  $X$ . Let  $\text{sub}_H(X)$  denote the number of  $H$ -subgraphs of  $X$ , that is,  $\text{sub}_H(X) := |\{\alpha \in [n]^{V(H)} : H^{(\alpha)} \subseteq X\}|$ .

**On the relationship between  $\text{SUB}_{\text{uncol}}(G)$  and  $\text{SUB}(G)$ .** For every pattern  $G$ , the color-coding method of [Alon, Yuster, and Zwick \[1995\]](#) provides an efficient many-one reduction from  $\text{SUB}_{\text{uncol}}(G)$  to  $\text{SUB}(G)$ . The colored version of  $G$ -subgraph isomorphism is therefore the harder problem in general. However, for many graphs  $G$  of interest such as cliques, these two problems are in fact equivalent. Namely, if  $G$  is a *core* (meaning every homomorphism  $G \rightarrow G$  is an isomorphism), then there is a trivial reduction from  $\text{SUB}(G)$  to  $\text{SUB}_{\text{uncol}}(G)$ , as the only subgraphs of  $G^{\uparrow n}$  that are isomorphic to  $G$  are those of the form  $G^{(\alpha)}$ .

**2.1 Threshold random graphs.** For the average-case analysis of the problem  $\text{SUB}(G)$ , it is natural to study a  $G$ -colored version of the Erdős–Rényi random graph. For a vector  $\vec{p} \in [0, 1]^{E(G)}$  of edge probabilities (one  $p_e \in [0, 1]$  for each  $e \in E(G)$ ), let  $\mathbf{G}_{n, \vec{p}}$  denote the random subgraph of  $G^{\uparrow n}$  which includes each potential edge  $\{v^{(a)}, w^{(b)}\}$  independently with probability  $p_{\{v, w\}}$ . The class of “threshold vectors” for the existence of  $G$ -subgraphs in  $\mathbf{G}_{n, \vec{p}}$  has a characterization in terms of certain edge-weightings on  $G$ .

**Definition 2.1** (Threshold weighting). Let  $G$  be a graph, let  $\theta$  be a function  $E(G) \rightarrow [0, 2]$ , and let  $\Delta_\theta$  be the function  $\{\text{subgraphs of } G\} \rightarrow \mathbb{R}_{\geq 0}$  defined by

$$\Delta_\theta(H) := |V(H)| - \sum_{e \in E(H)} \theta(e).$$

We say that  $\theta$  is a *threshold weighting* on  $G$  if  $\Delta_\theta(G) = 0$  and  $\Delta_\theta(H) \geq 0$  for all  $H \subseteq G$ . We say that  $\theta$  is *strict* if, moreover,  $\Delta_\theta(H) > 0$  for all proper subgraphs  $\emptyset \subset H \subset G$ .

The set of threshold weightings on  $G$  forms a convex polytope in  $[0, 2]^{E(G)}$ . For connected graphs  $G$ , the strict threshold weightings form the interior of this polytope. Note that only connected graphs admit strict threshold weightings, as it follows from the definition that  $\Delta_\theta(H) = 0$  whenever  $H$  is a union of connected components of  $G$ .

**Example 2.2.** For every graph  $G$ , the function  $\theta : E(G) \rightarrow [0, 2]$  defined by  $\theta(\{v, w\}) := \frac{1}{\deg(v)} + \frac{1}{\deg(w)}$  is a threshold weighting. In particular, if  $G$  is  $r$ -regular, then the constant function  $\theta = \frac{2}{r}$  is a threshold weighting. (Two additional constructions of threshold weightings are described at the end of this section.)

**Definition 2.3** (The random graph  $\mathbf{X}_\theta$ ). Every threshold weighting  $\theta$  on  $G$  gives rise to a sequence of random graphs  $\mathbf{X}_{n, \theta}$ , defined as the  $G$ -colored Erdős–Rényi graph  $\mathbf{G}_{n, \vec{p}}$  where  $\vec{p} \in [0, 1]^{E(G)}$  is the vector of edge probabilities  $p_e = n^{-\theta(e)}$ . That is,  $\mathbf{X}_{n, \theta}$  is the random subgraph of  $G^{\uparrow n}$  which includes each potential edge  $\{v^{(a)}, w^{(b)}\}$  independently with probability  $n^{-\theta(\{v, w\})}$ . To simplify notation, we will generally omit the parameter  $n$  and simply write  $\mathbf{X}_\theta$ .

Observe that the function  $\Delta_\theta$  characterizes the expected number of  $H$ -subgraphs in  $\mathbf{X}_\theta$ : for every  $H \subseteq G$ , we have  $\mathbb{E}[\text{sub}_H(\mathbf{X}_\theta)] = n^{\Delta_\theta(H)}$  by linearity of expectation. In particular,  $\text{sub}_G(\mathbf{X}_\theta)$  has expectation 1 (since  $\Delta_\theta(G) = 0$ ). Moreover, when  $\theta$  is strict,  $\text{sub}_G(\mathbf{X}_\theta)$  is asymptotically Poisson and  $\text{sub}_H(\mathbf{X}_\theta)$  is highly concentrated around its mean for all proper subgraphs  $H \subset G$ .

**Proposition 2.4.** *For every graph  $G$  and threshold weighting  $\theta$ , the probability that  $\mathbf{X}_\theta$  contains a  $G$ -subgraph converges to a limit in  $(0, 1)$ . When  $\theta$  is strict, this limit is  $1 - \frac{1}{e}$ .*

In light of [Proposition 2.4](#), it makes sense to study the average-case complexity of  $\text{SUB}(G)$  on  $\mathbf{X}_\theta$ , that is, the complexity of functions  $f : \{\text{subgraphs of } G^{\uparrow n}\} \rightarrow \{0, 1\}$

such that  $f(X_\theta) = 1 \Leftrightarrow \text{sub}_G(X_\theta) \geq 1$  holds asymptotically almost surely. We conclude this section with two constructions of threshold weightings.

**Example 2.5** (Threshold weightings from Markov chains). Let  $G$  be any graph and let  $M : V(G) \times V(G) \rightarrow [0, 1]$  be a *Markov chain* on  $G$  satisfying

- $M(v, w) > 0 \Rightarrow \{v, w\} \in E(G)$  and
- $\sum_w M(v, w) = 1$  for every  $v$ .

Then the function  $E(G) \rightarrow [0, 2]$  given by  $\{v, w\} \mapsto M(v, w) + M(w, v)$  is a threshold weighting on  $G$ . (This construction generalizes [Example 2.2](#), which corresponds to the Markov chain where  $M(v, w) = \frac{1}{\deg(v)}$  for every  $\{v, w\} \in E(G)$ .) The associated function  $\Delta_M$  has the property that  $\Delta_M(H)$  equals the amount of  $M$ -flow leaving the subgraph  $H$  (i.e.,  $\sum_{v,w} M(v, w)$  over pairs  $v, w$  with  $v \in V(H)$  and  $\{v, w\} \in E(G) \setminus E(H)$ ). In [Section 4.1](#) we use this construction of threshold weightings to bound the  $AC^0$  circuit size of  $\text{SUB}(G)$  in terms of the tree-width of  $G$ .

**Example 2.6** (The uncolored setting). The threshold for the existence of  $G$ -subgraphs in the Erdős–Rényi random graph  $G_{n,p}$  is well-known to be  $p = \Theta(n^{-c})$  where  $c$  is the constant  $\min_{H \subseteq G} \frac{|V(H)|}{|E(H)|}$  [Bollobás \[1981\]](#). For all intents and purposes, the average-case analysis of  $\text{SUB}_{\text{uncol}}(G)$  on  $G_{n,p}$  is equivalent to the average-case analysis of  $\text{SUB}(G)$  on  $X_{\theta_{\text{uncol}}}$  where  $\theta_{\text{uncol}} : E(G) \rightarrow \{0, c\}$  is the threshold weighting defined by  $\theta_{\text{uncol}}(e) = c \Leftrightarrow$  there exists  $H \subseteq G$  such that  $e \in E(H)$  and  $\frac{|V(H)|}{|E(H)|} = c$ . All lower and upper bounds described in this article translate easily between these two average-case settings, modulo insignificant constant factors as between  $n^{|V(G)|}$  and  $\binom{n}{|V(G)|}$ .

### 3 H-subgraph sensitivity

$AC^0$  functions are known to have *low average sensitivity* in the following sense [Boppana \[1997\]](#): for any  $AC^0$  function  $f : \{0, 1\}^n \rightarrow \{0, 1\}$  and independent uniform random  $x \in \{0, 1\}^n$  and  $i \in [n]$ , it holds that

$$\Pr_{x,i} [f(x) \neq f(x \text{ with its } i^{\text{th}} \text{ coordinate flipped})] \leq n^{-1+o(1)}.$$

Analogously, a key lemma in our lower bounds shows that  $AC^0$  functions  $f : \{\text{subgraphs of } G^{\uparrow n}\} \rightarrow \{0, 1\}$  have what might be termed “low average  $H$ -subgraph sensitivity on  $X_\theta$ ”.

**Definition 3.1.** For any graph  $F$ , let  $\mathbb{B}(F)$  denote the set of functions  $\{\text{subgraphs of } F\} \rightarrow \{0, 1\}$ .



We say that a function  $f \in \mathbb{B}(F)$  *depend on all coordinates* if for every  $e \in E(F)$ , there exists a subgraph  $F' \subseteq F$  such that  $f(F') \neq f(F' - e)$  where  $F' - e$  is the graph with edge set  $E(F') \setminus \{e\}$  (in other words, if  $f$  depends on all coordinates when viewed as a Boolean function  $\{0, 1\}^{|E(F)|} \rightarrow \{0, 1\}$ ).

For a function  $f \in \mathbb{B}(F)$  and graphs  $X, H \subseteq F$ ,

- let  $f^{\cup X} \in \mathbb{B}(F)$  denote the function  $f^{\cup X}(F') := f(X \cup F')$  and
- let  $f \upharpoonright_H \in \mathbb{B}(H)$  denote the restriction of  $f$  to domain  $\{\text{subgraphs of } H\}$ .

Note that the function  $f^{\cup X} \upharpoonright_H \in \mathbb{B}(H)$  depends on all coordinates if, and only if, for every  $e \in E(H)$ , there exists a subgraph  $H' \subseteq H$  such that  $f(X \cup H') \neq f(X \cup (H' - e))$ .

Fix any graph  $G$  and threshold weighting  $\theta$ . Consider any subgraph  $H \subseteq G$  and let  $\alpha$  be a uniform random element of  $[n]^{V(H)}$ , independent of  $X_\theta$ . For a function  $f \in \mathbb{B}(G^{\uparrow n})$ , we consider the randomly restricted function  $f^{\cup X_\theta} \upharpoonright_{H(\alpha)} \in \mathbb{B}(H(\alpha))$ . When  $f$  is  $AC^0$ -computable, the following lemma from [Li, Razborov, and Rossman \[2014\]](#) bounds the probability that  $f^{\cup X_\theta} \upharpoonright_{H(\alpha)}$  depends on all coordinates.

**Lemma 3.2** (*H*-subgraph sensitivity of  $AC^0$  functions). *Suppose  $f \in \mathbb{B}(G^{\uparrow n})$  is an  $AC^0$ -computable sequence of functions. Then for every subgraph  $H \subseteq G$ ,*

$$\Pr_{X_\theta, \alpha \in [n]^{V(H)}} [f^{\cup X_\theta} \upharpoonright_{H(\alpha)} \text{ depends on all coordinates} ] \leq n^{-\Delta_\theta(H) + o(1)}.$$

When  $\Delta_\theta(H) > 0$ , the  $n^{-\Delta_\theta(H) + o(1)}$  bound of [Lemma 3.2](#) is nontrivial and moreover tight. However, note that this lemma says nothing when  $\Delta_\theta(H) = 0$ , in particular when  $H = G$ . The main tools in the proof are the Switching Lemma of [Håstad \[1986\]](#), which shows that random restrictions simplify  $AC^0$  circuits, and Janson’s Inequality, which implies lower tail bounds for random variables  $\text{sub}_H(X_\theta)$  [Janson \[1990\]](#). The assumption that  $f$  is  $AC^0$ -computable is necessary, as for instance if  $f$  is the PARITY function (mapping  $X \subseteq G^{\uparrow n}$  to  $|E(X)| \bmod 2$ ), then the restricted function  $f^{\cup X_\theta} \upharpoonright_{H(\alpha)}$  depends on all coordinates with probability 1. (In the case that  $H$  is a single-edge subgraph of  $G$ , [Lemma 3.2](#) essentially equivalent to aforementioned bound on the average sensitivity of  $AC^0$  functions, only with respect to a product distribution rather than the uniform distribution.)

The next lemma from [Rossman \[2015\]](#) is an analogue of [Lemma 3.2](#) in the monotone setting. It shows that every monotone function, irrespective of its monotone circuit complexity, has “low average  $H$ -subgraph sensitivity of  $f$  on  $X_\theta$ ” in a different sense. Namely, we consider the event that  $H(\alpha)$  is a common minterm of  $f$  and  $f^{\cup X_\theta}$  (i.e.,  $f(H(\alpha)) = 1$  and  $f^{\cup X_\theta}(H(\alpha) - e) = 0$  for every  $e \in E(H(\alpha))$ ).

**Lemma 3.3** (*H*-subgraph sensitivity of monotone functions). *For every monotone function  $f \in \mathbb{B}(G^{\uparrow n})$  and subgraph  $H \subseteq G$ ,*

$$\Pr_{X_\theta, \alpha \in [n]^{V(H)}} [H^{(\alpha)} \text{ is a common minterm } f \text{ and } f^{\cup X_\theta}] \leq n^{-\Delta_\theta(H) + o(1)}.$$

In Section 5.3 we explain how Lemma 3.3 is used in place of Lemma 3.2 to derive lower bounds for monotone circuits and formulas using the same framework as our  $AC^0$  lower bounds.

## 4 The $AC^0$ circuit size of $\text{SUB}(G)$

This section presents results of Li, Razborov, and Rossman [2014] which characterize the average-case  $AC^0$  circuit size of  $\text{SUB}(G)$  on  $X_\theta$  for any  $G$  and  $\theta$  in terms of a combinatorial invariant  $\kappa_\theta(G)$ . This invariant is defined by dual min-max and max-min expressions.

**Definition 4.1.** A *union family* for a graph  $G$  is a set  $\mathcal{F}$  of subgraphs of  $G$  such that  $G \in \mathcal{F}$  and every  $F \in \mathcal{F}$  with at least two edges is the union of two proper subgraphs which both belong to  $\mathcal{F}$  (i.e., there exist proper subgraphs  $F_1, F_2 \subset F$  with  $F_1 \cup F_2 = F$  and  $F_1, F_2 \in \mathcal{F}$ ). Intuitively,  $\mathcal{F}$  is a blueprint for constructing  $G$  out of individual edges by taking pairwise unions of subgraphs.

A *hitting family* for  $G$  is a set  $\mathcal{H}$  of subgraphs of  $G$  such that  $\mathcal{F} \cap \mathcal{H} \neq \emptyset$  for every union family  $\mathcal{F}$  for  $G$ .

For any threshold weighting  $\theta$  on  $G$ , the invariant  $\kappa_\theta(G)$  is defined by the pair of dual expressions

$$\kappa_\theta(G) := \min_{\text{union families } \mathcal{F}} \max_{F \in \mathcal{F}} \Delta_\theta(F) = \max_{\text{hitting families } \mathcal{H}} \min_{H \in \mathcal{H}} \Delta_\theta(H).$$

**Example 4.2.** We illustrate these definitions by working through an example. Let  $K_k$  be the  $k$ -clique graph (i.e., the complete graph of order  $k \geq 2$ ) and let  $\theta$  be the constant threshold weighting  $\frac{2}{k-1}$ . We will show that  $\kappa_\theta(K_k) = \frac{k}{4} + O(\frac{1}{k})$  by constructing a union family  $\mathcal{F}$  and a hitting family  $\mathcal{H}$  that witness matching upper and lower bounds for  $\kappa_\theta(K_k)$ .

Let  $\mathcal{F}$  be the set of subgraphs  $F \subseteq K_k$  such that  $F$  is either a clique (i.e., a complete subgraph  $K_I \subseteq K_k$  where  $I \subseteq [k]$  with  $|I| \geq 2$ ) or a clique minus a single edge. Note that  $\mathcal{F}$  is a union family for  $K_k$ , as  $K_k \in \mathcal{F}$  and every graph in  $\mathcal{F}$  with at least two edges is the union of two proper subgraphs in  $\mathcal{F}$  (e.g.,  $K_{\{1, \dots, j\}}$  minus the edge  $\{1, j\}$  is the union of  $K_{\{1, \dots, j-1\}}$  and  $K_{\{2, \dots, j\}}$ ). A straightforward calculation shows  $\kappa_\theta(K_k) \leq \max_{F \in \mathcal{F}} \Delta_\theta(F) = \max_{F \in \mathcal{F}} |V(F)| - \frac{2}{k-1} |E(F)| = \frac{k}{4} + O(\frac{1}{k})$ , where this maximum over  $F \in \mathcal{F}$  is attained by a clique of size  $\lceil \frac{k}{2} \rceil$  minus a single edge.

To obtain a matching lower bound on  $\kappa_\theta(K_k)$ , we consider the hitting family  $\mathcal{H}$  consisting of subgraphs  $H \subseteq K_k$  such that  $|V(H)| \geq \frac{k}{2}$  and  $H = H_1 \cup H_2$  for some  $H_1, H_2$  satisfying  $|V(H_1)|, |V(H_2)| < \frac{k}{2}$ . The minimum of  $\Delta_\theta(H)$  over  $H \in \mathcal{H}$  is again attained by a clique of size  $\lceil \frac{k}{2} \rceil$  minus a single edge. This shows that the  $\frac{k}{4} + O(\frac{1}{k})$  upper bound coming from  $\mathcal{F}$  is tight.

**Example 4.3.** If  $G$  is an  $r$ -regular expander and  $\theta = \frac{2}{r}$ , then we obtain a lower bound  $\kappa_\theta(G) = \Omega(|V(G)|)$  (for a constant depending on the edge-expansion of  $G$ ) by considering the hitting family  $\{H \subseteq G : \frac{1}{3} \leq \frac{|V(H)|}{|V(G)|} < \frac{2}{3}\}$ .

We next state the main theorem of [Li, Razborov, and Rossman \[2014\]](#) and outline its proof.

**Theorem 4.4.** *For every graph  $G$  and threshold weighting  $\theta$ , the average-case  $AC^0$  circuit size of  $\text{SUB}(G)$  on  $X_\theta$  is at least  $n^{\kappa_\theta(G)-o(1)}$  and at most  $n^{2\kappa_\theta(G)+O(1)}$ .*

Theorem 4.4 together with Examples 2.6 and 4.2 imply a lower bound of  $\Omega(n^{k/4})$  on the  $AC^0$  circuit size of the average-case  $k$ -CLIQUE problem on  $\mathbf{G}_{n,p}$  at the threshold  $p = \Theta(n^{-2/(k-1)})$ .

**The upper bound.** We give a high-level description of an algorithm that solves  $\text{SUB}(G)$  a.a.s. correctly on  $X_\theta$  in time  $n^{2\kappa_\theta(G)+O(1)}$ , omitting details of the implementation by  $AC^0$  circuits. We use the fact that, with high probability,  $\text{sub}_H(X_\theta)$  is at most  $n^{\Delta_\theta(H)+o(1)}$  for all  $H \subseteq G$  (by Markov's inequality). Fix an optimal union family  $\mathcal{F}$  such that  $\kappa_\theta(G) = \max_{F \in \mathcal{F}} \Delta_\theta(F)$ . Also fix an enumeration  $F_1, \dots, F_m$  of graphs in  $\mathcal{F}$  such that  $F_m = G$  and each  $F_i$  is either a single edge or the union of two previous graphs in the sequence. In order for  $k = 1, \dots, m$ , the algorithm will compile a list of all  $F_k$ -subgraphs in  $X_\theta$ . When  $F_k$  is a single edge, this takes time  $O(n^2)$ . When  $F_k = F_i \cup F_j$  for  $i, j < k$ , this is done by examining each pair of subgraphs  $F_i^{(\alpha)} \subseteq X_\theta$  and  $F_j^{(\beta)} \subseteq X_\theta$  from the previously compiled lists: if  $\alpha_v = \beta_v$  for all  $v \in V(F_i) \cap V(F_j)$ , then  $F_k^{(\alpha \cup \beta)}$  is added to the list of  $F_k$ -subgraphs. Compiling this list therefore takes time  $O(\text{sub}_{F_i}(X_\theta) \cdot \text{sub}_{F_j}(X_\theta))$ , which with high probability is at most  $n^{\Delta_\theta(F_i)+\Delta_\theta(F_j)+o(1)} \leq n^{2\kappa_\theta(G)+o(1)}$ . Since there are only  $O(1)$  (at most  $2^{|E(G)|}$ ) lists to compute and nonemptiness of the final list determines whether  $X_\theta$  contains a  $G$ -subgraph, this algorithm has expected time  $n^{2\kappa_\theta(G)+O(1)}$ .

**The lower bound.** Let  $C$  be a sequence of  $AC^0$  circuits of size  $n^{\kappa_\theta(G)-\Omega(1)}$  which compute functions  $f \in \mathbb{B}(G \uparrow^n)$ . Our goal is to show that  $f$  does not agree with  $\text{SUB}(G)$  a.a.s. on  $X_\theta$ . We consider the randomly restricted function  $f^{\cup X_\theta} \upharpoonright_{G^{(\alpha)}}$  where  $\alpha$  is a uniform

random element of  $[n]^{V(G)}$  independent of  $X_\theta$ . We will show that

$$(4-1) \quad \Pr[f^{\cup X_\theta} \upharpoonright_{G^{(\alpha)}} \text{ depends on all coordinates }] = o(1).$$

Inequality 4-1 uses Lemma 3.2 on the “ $H$ -subgraph sensitivity” of  $AC^0$  functions. However, (4-1) does not follow by directly applying Lemma 3.2 to  $f$  with  $H = G$  (as the  $n^{-\Delta_\theta(H)+o(1)}$  bound of Lemma 3.2 is trivial when  $H = G$ ). Rather, we apply Lemma 3.2 to all functions  $g$  computed by subcircuits of  $C$  with respect to all subgraphs  $H \subseteq G$  which come from an optimal hitting family for  $G$ . We present the argument in detail in a moment.

On the other hand, we show that every function  $f \in \mathbb{B}(G^{\uparrow n})$  which agrees with  $\text{SUB}(G)$  a.a.s. on  $X_\theta$  satisfies

$$(4-2) \quad \Pr[f^{\cup X_\theta} \upharpoonright_{G^{(\alpha)}} \text{ depends on all coordinates }] = \Omega(1).$$

Since (4-1) and (4-2) are contradictory for sufficiently large  $n$ , we conclude that functions  $f$  computed by  $C$  do not solve  $\text{SUB}(G)$  on  $X_\theta$ .

We first justify (4-2), which is the more straightforward inequality. To illustrate the general idea, we make the stronger assumption that  $f$  coincides with  $\text{SUB}(G)$  on all inputs and we further assume that  $\theta$  is strict. In this case, Proposition 2.4 implies that  $X_\theta$  has no  $G$ -subgraph with probability  $\frac{1}{e} - o(1)$ . A straightforward union bound shows that, a.a.s., if  $X_\theta$  has no  $G$ -subgraph, then neither does  $X_\theta \cup H^{(\alpha)}$  for any proper subgraph  $H \subset G$ . (By “ $H^{(\alpha)}$ ” we mean  $H^{(\alpha \vee (H))}$ , which is a uniform random  $H$ -subgraph of  $G^{\uparrow n}$  independent of  $X_\theta$ .) It follows that, with probability  $\frac{1}{e} - o(1)$ , the randomly restricted function  $f^{\cup X_\theta} \upharpoonright_{G^{(\alpha)}} \in \mathbb{B}(G^{(\alpha)})$  outputs 1 on  $G$  and 0 on every  $H \subset G$  (i.e.,  $f^{\cup X_\theta} \upharpoonright_{G^{(\alpha)}}$  is the AND function over coordinates  $G^{(\alpha)}$ ). Since this function depends on all coordinates, inequality 4-2 follows. (When we only assume that  $f$  agrees with  $\text{SUB}(G)$  a.a.s. on  $X_\theta$ , this argument additionally requires showing that the total variation distance between random graphs  $X_\theta$  and  $X_\theta \cup H^{(\alpha)}$  is  $1 - \Omega(1)$  for every  $H \subseteq G$ .)

Onto the more interesting inequality (4-1), showing that a.a.s.  $f^{X_\theta} \upharpoonright_{G^{(\alpha)}}$  does not depend on all coordinates. Let  $\mathcal{G} \subseteq \mathbb{B}(G^{\uparrow n})$  be the set of functions computed by subcircuits of  $C$ . For every  $g \in \mathcal{G}$  and  $H \subseteq G$ , Lemma 3.2 implies that the randomly restricted function  $g^{\cup X_\theta} \upharpoonright_{H^{(\alpha)}}$  depends on all coordinates with probability at most  $n^{-\Delta_\theta(H)+o(1)}$ . Let us now fix an optimal hitting family  $\mathcal{H} \subseteq \{\text{subgraphs of } G\}$  such that  $\kappa_\theta(G) = \min_{H \in \mathcal{H}} \Delta_\theta(H)$ . Taking a union bound over  $g \in \mathcal{G}$  and  $H \in \mathcal{H}$ , we have

$$(4-3) \quad \Pr[(\exists g \in \mathcal{G})(\exists H \in \mathcal{H}) g^{\cup X_\theta} \upharpoonright_{H^{(\alpha)}} \text{ depends on all coordinates}] \leq |\mathcal{G}| \cdot |\mathcal{H}| \cdot n^{-\kappa_\theta(G)+o(1)} = o(1)$$

since  $|\mathcal{G}| \leq \text{size}(C) = n^{\kappa_\theta(G)-\Omega(1)}$  and  $|\mathcal{H}| \leq 2^{|E(G)|} = O(1)$ . Inequality (4-1) now follows by combining (4-3) with the following non-probabilistic claim.

**Claim 4.5.** *For any  $X \subseteq G^{\uparrow n}$  and  $\alpha \in [n]^{V(G)}$ , if  $f^{\cup X} \upharpoonright_{G^{(\alpha)}}$  depends on all coordinates, then there exist  $g \in \mathcal{G}$  and  $H \in \mathcal{H}$  such that  $g^{\cup X} \upharpoonright_{H^{(\alpha)}}$  depends on all coordinates.*

To prove [Claim 4.5](#), assume  $f^{\cup X} \upharpoonright_{G^{(\alpha)}}$  depends on all coordinates. Let  $\mathcal{F}$  be the family of subgraphs  $F \subseteq G$  for which there exists  $g \in \mathcal{G}$  such that  $g^{\cup X} \upharpoonright_{F^{(\alpha)}}$  depends on all coordinates. It suffices to show that  $\mathcal{F}$  is a union family for  $G$ . The claim then follows from the fact that  $\mathcal{F} \cap \mathcal{H}$  is nonempty (since  $\mathcal{H}$  is a hitting family for  $G$ ). To show that  $\mathcal{F}$  is a union family, we first note that  $G \in \mathcal{F}$  (by the assumption that  $f^{\cup X} \upharpoonright_{G^{(\alpha)}}$  depends on all coordinates).

Now consider any  $F \in \mathcal{F}$  with  $\geq 2$  edges. It remains to show that  $F$  is the union of two proper subgraphs which belong to  $\mathcal{F}$ . By definition of  $\mathcal{F}$  and  $\mathcal{G}$ , there exists a function  $g \in \mathbb{B}(G^{\uparrow n})$  computed by a subcircuit of  $C$  such that  $g^{\cup X} \upharpoonright_{F^{(\alpha)}}$  depends on all coordinates. Fix a choice of  $g$  computed by a subcircuit of minimal depth in  $C$ . Since  $g^{\cup X} \upharpoonright_{F^{(\alpha)}}$  depends on  $\geq 2$  coordinates (namely all edges of  $F^{(\alpha)}$ ), it cannot correspond to an input of  $C$  and must therefore come from a gate of  $C$ . Let  $g_1$  and  $g_2$  be the functions computed by the two subcircuits feeding into this gate. The function  $g$  is thus either  $g_1 \wedge g_2$  or  $g_1 \vee g_2$ .

For  $i = 1, 2$ , let  $F_i$  be the graph consisting of edges  $\{v, w\} \in E(F)$  such that  $g_i^{\cup X} \upharpoonright_{F^{(\alpha)}}$  depends on the corresponding edge  $\{v^{(\alpha_v)}, w^{(\alpha_w)}\} \in E(F^{(\alpha)})$ . Observe that the function  $g_i^{\cup X} \upharpoonright_{F_i^{(\alpha)}} \in \mathbb{B}(F_i^{(\alpha)})$  depends on all coordinates. Therefore,  $F_i \in \mathcal{F}$ . Next, note that  $F_i$  must be proper subgraph of  $F$  by the minimality in our choice of  $g$ . Finally, observe that  $F = F_1 \cup F_2$  (since if  $g$  depends on a given coordinate in  $E(F)$ , then so must one or both of  $g_1$  and  $g_2$ , and the same is true after applying the restriction  $^{\cup X} \upharpoonright_{F^{(\alpha)}}$  to all three functions). As we have shown that  $F$  is the union of two proper subgraphs which belong to  $\mathcal{F}$ , this completes the proof.  $\square$

By a similar argument, we obtain a similar  $n^{\kappa_\theta(G)-o(1)}$  lower bound on the *monotone* circuit size of  $\text{SUB}(G)$ . In this argument, [Lemma 3.3](#) plays the role of [Lemma 3.2](#) in bounding the “ $H$ -subgraph sensitivity” of each subcircuit. However, as we explain in [Section 5.3](#), for most patterns  $G$ , the lower bound we obtain in the monotone setting is only worst-case, or average-case under a non-product distribution.

**4.1 Tree-width.** Tree-width, denoted  $\text{tw}(G)$ , is an important invariant that arises frequently in parameterized complexity and several areas of graph theory. Roughly speaking, it measures the extent to which a graph is “tree-like”: trees and forests have tree-width 1, while the complete graph of order  $k$  has tree-width  $k - 1$ .

In the introduction, it was mentioned that the  $G$ -subgraph isomorphism problem is solvable in time  $O(n^{\text{tw}(G)+1})$  for all patterns  $G$ . In fact,  $\text{SUB}(G)$  is solvable by monotone  $AC^0$  circuits of size  $O(n^{\text{tw}(G)+1})$ . If we compare this upper bound to the lower bound of

**Theorem 4.4**, we see that  $\max_{\theta} \kappa_{\theta}(G) \leq \text{tw}(G) + 1$ . The next proposition shows that this inequality is nearly tight.

**Proposition 4.6.** *Every graph  $G$  admits a threshold weighting  $\theta$  such that  $\kappa_{\theta}(G) = \Omega(\text{tw}(G)/\log \text{tw}(G))$ .*

*Proof.* We will use a lemma of [Grohe and Marx \[2009\]](#) which states that, for every  $G$  with tree-width  $k$ , there exists a set  $W \subseteq V(G)$  of size  $|W| = \Omega(k)$  together with a concurrent flow on  $G$  with vertex-capacity 1 which routes  $\Omega(\frac{1}{k \log k})$  units of flow between every pair of vertices in  $W$ . This concurrent flow is easily transformed to a Markov chain  $M$  on  $G$  (in the sense of [Example 2.5](#)) with the property that  $\Delta_M(H) = \Omega(\frac{|V(H) \cap W| \cdot |V(H) \setminus W|}{k \log k})$  for all  $H \subseteq G$ . We now consider the hitting family  $\mathcal{H}$  consisting of subgraphs  $H \subseteq G$  such that  $\frac{1}{3} \leq \frac{|V(H) \cap W|}{|W|} < \frac{2}{3}$  (similar to [Example 4.3](#)). This gives the bound  $\kappa_M(G) \geq \min_{H \in \mathcal{H}} \Delta_M(H) = \Omega(\frac{k}{\log k})$  with respect to the threshold weighting  $\{v, w\} \mapsto M(v, w) + M(w, v)$  induced by  $M$ .  $\square$

We remark that the upper bound  $\max_{\theta} \kappa_{\theta}(G) \leq \text{tw}(G) + 1$  has a direct proof that does not appeal to [Theorem 4.4](#). In fact, the next proposition shows that  $\max_{\theta} \kappa_{\theta}(G)$  is at most the *branch-width* of  $G$ , an invariant that is related to tree-width by  $\text{bw}(G) \leq \text{tw}(G) + 1 \leq \frac{3}{2}\text{bw}(G)$  [Robertson and Seymour \[1991\]](#).

**Proposition 4.7.**  $\kappa_{\theta}(G) \leq \text{bw}(G)$  for every threshold weighting  $\theta$  on  $G$ .

*Proof.* Branch-width admits a simple characterization in terms of union families:

$$\text{bw}(G) = \min_{\text{complement-closed union families } \mathcal{F}} \max_{F \in \mathcal{F}} |V(F) \cap V(\overline{F})|.$$

Here *complement-closed* means  $F \in \mathcal{F} \Rightarrow \overline{F} \in \mathcal{F}$  where  $\overline{F}$  is the graph with  $E(\overline{F}) = E(G) \setminus E(F)$ . It follows from the definition of threshold weighting that  $\Delta_{\theta}(F) \leq \Delta_{\theta}(F) + \Delta_{\theta}(\overline{F}) = |V(F) \cap V(\overline{F})|$  for every threshold weighting  $\theta$  and subgraph  $F \subseteq G$ . Therefore,  $\kappa_{\theta}(G) = \min_{\text{union families } \mathcal{F}} \max_{F \in \mathcal{F}} \Delta_{\theta}(F) \leq \text{bw}(G)$ .  $\square$

## 5 The restricted formula size of $\text{SUB}(G)$

In this section we sketch an extension the lower bound technique that yields quantitatively stronger lower bounds for formulas vis-à-vis circuits in both the  $AC^0$  and monotone settings. The improvement is significant for patterns of constant tree-width such as paths and cycles where  $\text{SUB}(G)$  is computable by polynomial-size circuits but is conjectured to require super-polynomial size formulas.

An outline of this section is as follows. In [Section 5.1](#) we introduce the key notion of *pathsets* (relations  $\mathcal{Q} \subseteq [n]^{V(H)}$ ) that satisfy certain density constraints related to the

bounds on “ $H$ -subgraph sensitivity” given by Lemmas 3.2 and 3.3). We then define *pathset formulas*, which are a tree-like model for constructing pathsets. In Section 5.2 we describe a randomized reduction which transforms any  $AC^0$  formula that solves average-case  $\text{SUB}(G)$  on  $X_\theta$  into a pathset formula that computes a dense subset of  $[n]^{V(G)}$ . In Section 5.3 we outline a similar transformation for monotone formulas.

In Section 5.4 we arrive at the combinatorial heart of the technique: an  $n^{\tau_\theta(G)-o(1)}$  lower bound on the size of pathset formulas that compute a dense subset of  $[n]^{V(G)}$ . Here  $\tau_\theta(G)$  is an invariant of the threshold-weighted graphs, which plays an analogous role to  $\kappa_\theta(G)$  in the context of formulas. Although  $\tau_\theta(G)$  turns out to be much harder to compute, we are able to bound  $\tau_\theta(G)$  in a few special cases of interest, such as when  $G$  is a cycle, path or complete binary tree. Finally, in Section 5.5 we discuss a relationship between  $\max_\theta \tau_\theta(G)$  and the tree-depth of  $G$ .

**5.1 Pathset formulas.** In what follows, we fix a graph  $G$  and a threshold weighting  $\theta$ , as well as  $n \in \mathbb{N}$  and an arbitrary “density parameter”  $\varepsilon \in [0, 1]$ . (In our applications, we take  $\varepsilon$  to be  $n^{1-o(1)}$  and later  $n^{1/2-o(1)}$ .)

**Definition 5.1.** Let  $\mathcal{Q} \subseteq [n]^V$  where  $V$  is any finite set. (We regard  $\mathcal{Q}$  as a “ $V$ -ary relation with universe  $[n]$ ”.) The *density* of  $\mathcal{Q}$  is defined by

$$\mu(\mathcal{Q}) := \Pr_{\alpha \in [n]^V} [\alpha \in \mathcal{Q}] (= |\mathcal{Q}|/n^{|V|}).$$

For  $S \subseteq V$  and  $\beta \in [n]^S$ , the *conditional density* of  $\mathcal{Q}$  on  $\beta$  is defined by

$$\mu(\mathcal{Q} \mid \beta) := \Pr_{\alpha \in [n]^V} [\alpha \in \mathcal{Q} \mid \alpha_S = \beta].$$

The *join* of relations  $\mathcal{Q} \subseteq [n]^V$  and  $\mathcal{B} \subseteq [n]^W$  is the relation  $\mathcal{Q} \bowtie \mathcal{B} \subseteq [n]^{V \cup W}$  consisting of  $\gamma \in [n]^{V \cup W}$  such that  $\gamma_V \in \mathcal{Q}$  and  $\gamma_W \in \mathcal{B}$ .

**Definition 5.2.** Let  $H$  be a subgraph of  $G$ . An  $H$ -*pathset* (with respect to  $G, \theta, n, \varepsilon$ ) is a relation  $\mathcal{Q} \subseteq [n]^{V(H)}$  satisfying density constraints

$$(5-1) \quad \mu(\mathcal{Q} \mid \beta) \leq \varepsilon^{\Delta_\theta(H_1)} \quad \text{for all } H_1 \uplus H_2 = H \text{ and } \beta \in [n]^{V(H_2)}.$$

Here the pair  $H_1, H_2$  range over vertex-disjoint partitions of  $H$  (such that  $H_1 \cup H_2 = H$  and  $V(H_1) \cap V(H_2) = \emptyset$ ). Thus, if  $H$  has  $t$  connected components, then (5-1) includes  $2^t$  separate inequalities. Note that the inequality corresponding to  $H_1 = H$  and  $H_2 = \emptyset$  (the empty graph) is  $\mu(\mathcal{Q}) \leq \varepsilon^{\Delta_\theta(H)}$ , while the inequality corresponding to  $H_1 = \emptyset$  and  $H_2 = H$  is vacuous since  $\Delta_\theta(\emptyset) = 0$ . If  $H$  is connected, it follows that a relation  $\mathcal{Q} \subseteq [n]^{V(H)}$  is an  $H$ -pathset if and only if  $\mu(\mathcal{Q}) \leq \varepsilon^{\Delta_\theta(H)}$ . Finally, note that every relation  $\mathcal{Q} \subseteq [n]^{V(G)}$  is a  $G$ -pathset since  $\Delta_\theta(G_1) = 0$  whenever  $G_1$  is a union of connected components of  $G$ .

**Definition 5.3.** A *pathset formula* (with respect to  $G, \theta, n, \varepsilon$ ) is a rooted binary tree  $F$  together with an indexed family of relations  $\{\mathcal{Q}_{f,H} \subseteq [n]^{V(H)}\}_{f \in V(F), H \subseteq G}$  subject to three conditions:

- (i)  $\mathcal{Q}_{f,H}$  is a  $H$ -pathset,
- (ii) if  $f$  is a leaf and  $|E(H)| \geq 2$ , then  $\mathcal{Q}_{f,H} = \emptyset$ ,
- (iii) if  $f$  is a non-leaf with children  $f_1$  and  $f_2$ , then

$$\mathcal{Q}_{f,H} \subseteq \bigcup_{H_1, H_2 \subseteq H : H_1 \cup H_2 = H} (\mathcal{Q}_{f_1, H_1} \bowtie \mathcal{Q}_{f_2, H_2}).$$

We view  $F$  as “computing” the family of pathsets  $\{\mathcal{Q}_{f_{\text{out}}, H}\}_{H \subseteq G}$  (and in particular the  $G$ -pathset  $\mathcal{Q}_{f_{\text{out}}, G}$ ) where  $f_{\text{out}}$  is the root of  $F$ .

**5.2 Transforming  $AC^0$  formulas to pathset formulas.** For any Boolean function  $f \in \mathbb{B}(G^{\uparrow n})$  and a subgraph  $H \subseteq G$ , let  $\mathcal{Q}_{f,H}^{X_\theta} \subseteq [n]^{V(H)}$  be the random relation defined by

$$\mathcal{Q}_{f,H}^{X_\theta} := \{\alpha \in [n]^{V(H)} : f^{\cup X_\theta} \upharpoonright_{H^{(\alpha)}} \text{ depends on all coordinates}\}.$$

When  $f$  is  $AC^0$ -computable, Lemma 3.2 is equivalent to the expectation bound  $\mathbb{E}[\mu(\mathcal{Q}_{f,H}^{X_\theta})] \leq n^{-\Delta_\theta(H) + o(1)}$ . This can be extended to show that  $\mu(\mathcal{Q}_{f,H}^{X_\theta}) > n^{-(1-\delta)\Delta_\theta(H)}$  with exponentially small probability for any constant  $\delta > 0$  (i.e., with probability  $\exp(-\Omega(n^c))$  where  $c > 0$  depends on  $\delta$  and the minimum nonzero value of  $\Delta_\theta$ ). It is a small additional step to show that  $\mathcal{Q}_{f,H}^{X_\theta}$  fails to be an  $H$ -pathset (with respect to  $G, \theta, n$  and  $\varepsilon = n^{-1+\delta}$ ) with exponentially small probability.

If  $F$  is an  $AC^0$  formula, it follows that the family of relations  $\mathcal{Q}_{f,H}^{X_\theta} \subseteq [n]^{V(H)}$  (indexed by subformulas  $f$  of  $F$  and subgraphs  $H \subseteq G$ ) a.a.s. constitutes a pathset formula. Condition (i) of Definition 5.3 is established by taking a union bound, over the  $n^{O(1)}$  pairs of  $f$  and  $H$ , of the exponentially small probability that  $\mathcal{Q}_{f,H}^{X_\theta}$  fails to be an  $H$ -pathset. Conditions (ii) and (iii) both hold with probability 1 (by observations which appeared earlier in the proof of Claim 4.5). Finally, if  $F$  solves SUB( $G$ ) a.a.s. correctly on  $X_\theta$ , it follows that the  $G$ -pathset computed by  $F$  is .99-dense with constant probability by an argument similar to inequality (4-2).

**5.3 Transforming monotone formulas to pathset formulas.** Let  $F$  be a monotone formula of polynomial size. As a first attempt to transform  $F$  to a pathset formula, for each subformula  $f$  of  $F$  and subgraph  $H \subseteq G$ , let  $\mathfrak{M}_{f,H}^{X_\theta} \subseteq [n]^{V(H)}$  be the relation consisting of  $\alpha \in [n]^{V(H)}$  such that  $H^{(\alpha)}$  is a minterm of  $f^{\cup X_\theta}$ . This family of relations satisfies



conditions (ii) and (iii) of [Definition 5.3](#) with probability 1. (Condition (iii) follows from the elementary fact that every minterm of  $f_1 \vee f_2$  is a minterm of  $f_1$  or  $f_2$ , while every minterm of  $f_1 \wedge f_2$  is the union of a minterm of  $f_1$  and a minterm of  $f_2$ .) However, the relation  $\mathfrak{M}_{f,H}^{X_\theta}$  can fail to be an  $H$ -pathsets with probability  $\Omega(1/n)$  (e.g., if  $f$  is the monotone threshold function  $f(X) = 1 \Leftrightarrow |E(X)| \geq \sum_{e \in E(G)} n^{2-\theta(e)}$ ). This failure probability is too large for us to establish condition (i) by taking a union bound over pairs  $f$  and  $H$ .

To get around this issue, we consider different relations defined in terms of an increasing sequence  $\vec{X}_\theta$  of random graphs  $X_\theta^0 \subseteq \dots \subseteq X_\theta^m$  where  $m = n^{o(1)}$ . This sequence is generated as  $X_\theta^0 := X_\theta$  and  $X_\theta^i := X_\theta^{i-1} \cup Y^i$  where  $Y^i$  is an independent copy of the  $G$ -colored Erdős–Rényi graph  $\mathbf{G}_{n,p}$  where  $p_e := n^{-(1-\delta)\theta(e)}$  for a small constant  $\delta > 0$  (i.e., each  $Y^i$  is a sparse version of  $X_\theta$ ). If  $f$  is a depth- $d$  subformula of  $F$ , we say that  $H^{(\alpha)}$  is a *persistent minterm* of  $f \cup \vec{X}_\theta$  if it is a common minterm of  $f \cup X_\theta^i$  and  $f \cup X_\theta^j$  for some  $0 \leq i < j \leq m$  with  $j - i = \binom{d+|E(H)|}{|E(H)|}$ . Finally, we consider relations

$$\mathcal{P}_{f,H}^{\vec{X}_\theta} := \{\alpha \in [n]^{V(H)} : H^{(\alpha)} \text{ is a persistent minterm of } f \cup \vec{X}_\theta\}.$$

The definition of persistent minterms ensures that, just like  $\mathfrak{M}_{f,H}^{X_\theta}$ , this family of relations satisfies conditions (ii) and (iii) of [Definition 5.3](#) with probability 1. An extension of [Lemma 3.3](#) shows that  $\mathcal{P}_{f,H}^{X_\theta}$  fails to be an  $H$ -pathset (with respect to  $\varepsilon = n^{-1+2\delta}$ ) with exponentially small probability. A union bound now shows that this family of relations a.a.s. satisfies condition (i), thus transforming  $F$  to a pathset formula.

In order for this pathset formula to compute a .99-dense  $G$ -pathset with constant probability, we require two additional assumptions: first, that  $F$  has depth  $O(\log n)$  so that  $\binom{\text{depth}(F)+|E(G)|}{|E(G)|} \leq m = n^{o(1)}$  (this is without loss of generality by [Spira \[1971\]](#)); and second, that  $F$  solves  $\text{SUB}(G)$  a.a.s. on both  $X_\theta$  and  $X_\theta^m (= X_\theta \cup Y^1 \cup \dots \cup Y^m)$ . This is akin to solving  $\text{SUB}_{\text{uncol}}(G)$  a.a.s. correctly on both  $\mathbf{G}_{n,p}$  and  $\mathbf{G}_{n,p+p^{1+\delta}}$ , or alternatively on a convex combination of these random graphs. The lower bounds that we obtain in the monotone setting are therefore merely worst-case, or average-case under a non-product distribution.

However, in the special case of  $G = C_k$  and  $\theta = 1$  (corresponding to the average-case  $k$ -CYCLE problem on  $\mathbf{G}_{n,p}$  at the threshold  $p = \Theta(1/n)$ ), we may take each  $Y^i$  to be the union of  $n^{1/2-\delta}$  random paths of length  $k$ . In this case we are able to show that relations  $\mathcal{P}_{f,H}^{\vec{X}_\theta}$  are pathsets with respect to density parameter  $\varepsilon = n^{1/2-2\delta}$ . Moreover, random graphs  $X_\theta$  and  $X_\theta^m$  have total variation distance  $o(1)$ . As a result, we obtain an average-case lower bound for  $\text{SUB}(G)$  on  $X_\theta$  alone.

**5.4 Pathset complexity.** At this point, we are left with the task of proving lower bounds on the size of pathset formulas computing dense  $G$ -pathsets. This is by far the hardest part of the overall technique. Here we present only a brief outline. We introduce a family of complexity measures, each associated with different union family. However, rather than viewing a union family as a set of subgraphs of  $G$ , we explicitly consider the underlying tree structure.

**Definition 5.4.** A *union tree*  $A$  is a rooted binary tree whose leaves are labeled by edges of  $G$ . We denote by  $G_A$  the subgraph of  $G$  formed by the edges that label the leafs in  $A$ . We say that  $A$  is an  $H$ -union tree if  $G_A = H$ . For union trees  $A$  and  $B$ , let  $\langle A, B \rangle$  denote the union tree consisting of a root attached to  $A$  and  $B$  (with  $G_{\langle A, B \rangle} = G_A \cup G_B$ ). Notation  $A \preceq B$  denotes that  $A$  is a subtree of  $B$  formed by a node of  $B$  together with all of its descendants.

**Definition 5.5.** *Pathset complexity* (with respect to  $G, \theta, n, \varepsilon$ ) is the unique pointwise maximal family of functions  $\chi_A : \{G_A\text{-pathsets}\} \rightarrow \mathbb{N}$ , one for each union tree  $A$ , subject to the following inequalities:

- $\chi_A(\mathcal{Q}) \leq 1$  whenever  $A$  is a union tree of size 1,
- $\chi_A(\mathcal{Q}) \leq \sum_i \chi_A(\mathcal{Q}_i)$  whenever  $\mathcal{Q} \subseteq \bigcup_i \mathcal{Q}_i$ ,
- $\chi_A(\mathcal{Q}) \leq \max\{\chi_B(\mathcal{B}), \chi_C(\mathcal{C})\}$  whenever  $A = \langle B, C \rangle$  and  $\mathcal{Q} \subseteq \mathcal{B} \bowtie \mathcal{C}$ .

Pathset complexity gives lower bounds on pathset formula size (and by extension lower bounds on  $AC^0$  formula size and monotone formula size). We describe the relationship between pathset formula size and pathset complexity in terms of a parameter  $\tau_\theta(G)$ , which plays an analogous role to  $\kappa_\theta(G)$  in our formula lower bounds.

**Definition 5.6.** For each union tree  $A$ , let  $\Phi_A$  be the maximum constant (depending on  $G$  and  $\theta$  alone) such that the inequality  $\chi_A(\mathcal{Q}) \geq (1/\varepsilon)^{\Phi_A} \cdot \mu(\mathcal{Q})$  holds for every  $G_A$ -pathset  $\mathcal{Q}$  and every setting of parameters  $n$  and  $\varepsilon$ . The invariant  $\tau_\theta(G)$  is defined as the minimum value of  $\Phi_A$  over  $G$ -union trees  $A$ .

For comparison, note that the invariant  $\kappa_\theta(G)$  equals the minimum value of  $\max_{A' \preceq A} \Delta_{A'}$  over  $G$ -union trees  $A$ , writing  $\Delta_{A'}$  to abbreviate  $\Delta_\theta(G_{A'})$ . The constant  $\Phi_A$  thus plays a similar role in our formula lower bounds as  $\max_{A' \preceq A} \Delta_{A'}$  in our circuit lower bounds.

It follows from the above definitions, though not entirely straightforwardly, that any pathset formula  $F$  computing a .99-dense  $G$ -pathset (i.e., such that  $\mu(\mathcal{R}_{f_{\text{out}}, G}) \geq .99$ ) must have size  $\Omega((1/\varepsilon)^{\tau_\theta(G)})$ . (This  $\Omega(\cdot)$  hides a factor of  $(1/2)^{2^{|E(G)|}}$ , which arises from partitioning  $\mathcal{R}_{f_{\text{out}}, G}$  according to a union tree that accounts for the construction of each of

its elements in  $F$ .) Combined with the reduction outlined in [Section 5.2](#), this implies the following lower bound, which is a version of [Theorem 4.4](#) for  $AC^0$  formulas.

**Theorem 5.7** ([Rossman \[2014a\]](#)). *The average-case  $AC^0$  formula size of  $\text{SUB}(G)$  on  $X_\theta$  is at least  $n^{\tau_\theta(G)-o(1)}$ .*

Using the reduction outlined in [Section 5.3](#), we get the following lower bounds in the monotone setting.

**Theorem 5.8** ([Rossman \[2015\]](#)). *For all  $G$  and  $\theta$ , the worst-case monotone formula (resp. circuit) size  $\text{SUB}(G)$  is at least  $n^{\tau_\theta(G)-o(1)}$  (resp.  $n^{\kappa_\theta(G)-o(1)}$ ). In the case of  $G = C_k$  and  $\theta = 1$ , the average-case monotone formula size of  $\text{SUB}(G)$  on  $X_\theta$  is at least  $n^{\frac{1}{2}\tau_\theta(G)-o(1)}$ .*

It remains to prove lower bounds on  $\tau_\theta(G)$ , especially in cases of interest like  $G = C_k$  and  $\theta = 1$ . This requires us to prove lower bounds on constants  $\Phi_A$  for every possible  $G$ -union tree  $A$ . In principle, this is a problem in the realm of graph theory, since  $\Phi_A$  depends on  $G$  and  $\theta$  alone. Unfortunately, we do not have any nice expression for  $\Phi_A$ , nor even an efficient method of computing these constants. Nevertheless, we are able to deduce some useful inequalities. For starters, it is simple to show that  $\Phi_A \geq \Delta_A$  and moreover  $\Phi_A \geq \Delta_{A'}$  for every  $A' \preceq A$ . However, this merely amounts to the inequality  $\tau_\theta(G) \geq \kappa_\theta(G)$ , which is the unsurprising fact that our formula lower bounds are not weaker than our circuit lower bounds.

To derive stronger lower bounds on  $\Phi_A$ , we make use of structural properties of pathset complexity:

- **(projection lemma)**  $\chi_{A'}(\text{proj}_{A'}(\mathcal{Q})) \leq \chi_A(\mathcal{Q})$  for all union trees  $A' \preceq A$  and every  $G_A$ -pathset  $\mathcal{Q}$ , where  $\text{proj}_{A'}(\mathcal{Q}) \subseteq [n]^{V(G_{A'})}$  is the projection of  $\mathcal{Q}$  to coordinates in  $V(G_{A'})$ ,
- **(restriction lemma)**  $\chi_{A \upharpoonright H_1}(\mathcal{Q} \upharpoonright \beta) \leq \chi_A(\mathcal{Q})$  for every vertex-disjoint partition  $G_A = H_1 \uplus H_2$  and  $\beta \in [n]^{V(H_2)}$ , where  $A \upharpoonright H_1$  is the union tree obtained from  $A$  by deleting every leaf that is labeled by an edge of  $H_2$ .

These lemmas allow us to derive two useful inequalities on constants  $\Phi_A$ : for all union trees  $A = \langle B, C \rangle$  and  $B' \preceq B$  and  $C' \preceq C$ ,

$$(5-2) \quad \Phi_A \geq \Phi_{B'} + \Delta_C + \Delta_{A \ominus C},$$

$$(5-3) \quad \Phi_A \geq \frac{1}{2}(\Phi_{B'} + \Phi_{C' \ominus B'} + \Delta_A + \Delta_{A \ominus \langle B', C' \rangle}).$$

Here  $\ominus$  is the following operation on union trees:  $A \ominus B$  is the union tree obtained from  $A$  by deleting every leaf that is labeled an edge whose connected component in  $G_A$  contains any vertex of  $G_B$ .

In the case of  $G = C_k$  and  $\theta = 1$ , inequalities (5-2) and (5-3) can be used to show that  $\kappa_\theta(G) \geq \frac{1}{6} \log_2(k)$ . This yields the following corollary of Theorems 5.7 and 5.8.

**Corollary 5.9.**  *$AC^0$  formulas, as well as monotone formulas, which solve the average-case  $k$ -CYCLE problem on  $G_{n,p}$  at the threshold  $p = \Theta(1/n)$  require size  $n^{\Omega(\log k)}$ .*

In unpublished work in progress, we explore an additional inequality on constant  $\Phi_A$ . Consider any root-to-leaf branch in a union tree  $A$ , and let  $A_1, \dots, A_m$  enumerate the union trees hanging off this branch in any order. For example, we might have  $A = \langle A_3, \langle \langle A_1, \langle A_5, A_2 \rangle \rangle, A_4 \rangle \rangle$ . For all such  $A$  and  $A_1, \dots, A_m$ , there is an inequality

$$(5-4) \quad \Phi_A \geq \Delta_{A_1} + \Delta_{A_2 \ominus A_1} + \Delta_{A_3 \ominus (A_1 \cup A_2)} + \dots + \Delta_{A_m \ominus (A_1 \cup \dots \cup A_{m-1})}.$$

Again in the case  $G = C_k$  and  $\theta = 1$ , using (5-4) we can show that if  $A$  is a  $G$ -union tree with *left-depth*  $d$  (i.e., no root-to-leaf branch in  $A$  descends to the left more than  $d$  times), then  $\Phi_A \geq \Omega(dk^{1/d}) - O(d)$ . This in turn leads to nearly tight tradeoffs between the size and alternation-depth of  $AC^0$  formulas solving the average-case  $k$ -CYCLE problem. Inequality (5-4) is also useful in bounding  $\tau_\theta(G)$  for additional patterns of interest, such as complete binary trees.

**5.5 Tree-depth.** The *tree-depth* of a graph  $G$ , denoted  $\text{td}(G)$ , is the minimum height of a forest  $F$  with the property that every edge of  $G$  connects a pair of vertices that have an ancestor-descendant relationship to each other in  $F$  (see Nešetřil and Ossona de Mendez [2006]). Analogous to the relationship between tree-width and the circuit size, it turns out that  $\text{SUB}(G)$  is solvable by monotone  $AC^0$  formulas of size  $O(n^{\text{td}(G)})$ . Comparing this upper bound to the lower bound of Theorem 5.7, it follows that  $\max_\theta \tau_\theta(G) \leq \text{td}(G)$ .

Using a recent result in graph minor theory of Kawarabayashi and Rossman [2018], we are able to show that  $\max_\theta \tau_\theta(G) \geq \text{td}(G)^c$  for all patterns  $G$  where  $c > 0$  is an absolute constant. This result reduces this inequality to three special cases when the pattern  $G$  is a grid, a path, or a complete binary tree. By bounding  $\max_\theta \tau_\theta(G)$  in these three cases, we obtain an  $\Omega(n^{\text{td}(G)^c})$  lower bound on both the  $AC^0$  and monotone formula size of  $\text{SUB}(G)$  for arbitrary patterns  $G$ .

## References

- Miklós Ajtai (1983). “ $\Sigma_1^1$  formulae on finite structures”. *Annals of Pure and Applied Logic* 24, pp. 1–48 (cit. on p. 3446).
- Noga Alon, Raphael Yuster, and Uri Zwick (1995). “Color-coding”. *Journal of the ACM* 42.4, pp. 844–856 (cit. on p. 3449).

- Kazuyuki Amano (2010). “ $k$ -Subgraph isomorphism on  $AC^0$  circuits”. *Computational Complexity* 19.2, pp. 183–210 (cit. on p. 3448).
- Boaz Barak, Samuel B Hopkins, Jonathan Kelner, Pravesh Kothari, Ankur Moitra, and Aaron Potechin (2016). “A nearly tight sum-of-squares lower bound for the planted clique problem”. In: *57th IEEE Symposium on Foundations of Computer Science*, pp. 428–437 (cit. on p. 3445).
- Paul Beame (1990). “Lower bounds for recognizing small cliques on CRCW PRAM’s”. *Discrete Appl. Math.* 29.1, pp. 3–20 (cit. on p. 3448).
- Stuart J. Berkowitz (1982). *On some relationships between monotone and nonmonotone circuit complexity*. Tech. rep. Department of Computer Science, University of Toronto (cit. on p. 3447).
- Béla Bollobás (1981). “Threshold functions for small subgraphs”. *Math. Proc. Camb. Phil. Soc.* 90, pp. 197–206 (cit. on p. 3451).
- Ravi B Boppana (1997). “The average sensitivity of bounded-depth circuits”. *Information Processing Letters* 63.5, pp. 257–261 (cit. on p. 3451).
- Jianer Chen, Benny Chor, Mike Fellows, Xiuzhen Huang, David Juedes, Iyad A Kanj, and Ge Xia (2005). “Tight lower bounds for certain parameterized NP-hard problems”. *Information and Computation* 201.2, pp. 216–231 (cit. on p. 3444).
- Vitaly Feldman, Elena Grigorescu, Lev Reyzin, Santosh Vempala, and Ying Xiao (2013). “Statistical algorithms and a lower bound for detecting planted cliques”. In: *45th ACM Symposium on Theory of Computing*, pp. 655–664 (cit. on p. 3445).
- Merrick L. Furst, James B. Saxe, and Michael Sipser (1984). “Parity, Circuits, and the Polynomial-Time Hierarchy”. *Mathematical Systems Theory* 17, pp. 13–27 (cit. on p. 3446).
- Michelangelo Grigni and Michael Sipser (1992). “Monotone complexity”. *Boolean function complexity* 169, pp. 57–75 (cit. on p. 3447).
- Martin Grohe and Dániel Marx (2009). “On tree width, bramble size, and expansion”. *Journal of Combinatorial Theory, Series B* 99.1, pp. 218–228 (cit. on p. 3457).
- Johan Håstad (1986). “Almost optimal lower bounds for small depth circuits”. In: *18th ACM Symposium on Theory of Computing*, pp. 6–20 (cit. on pp. 3446, 3452).
- (1998). “The shrinkage exponent of de Morgan formulas is 2”. *SIAM Journal on Computing* 27.1, pp. 48–64 (cit. on p. 3446).
- Svante Janson (1990). “Poisson approximation for large deviations”. *Random Structures and Algorithms* 1.2, pp. 221–229 (cit. on p. 3452).
- Mark Jerrum (1992). “Large cliques elude the metropolis process”. *Random Structures and Algorithms* 3.4, pp. 347–359 (cit. on p. 3445).
- Ari Juels and Marcus Peinado (2000). “Hiding Cliques for Cryptographic Security”. *Des. Codes Cryptography* 20.3, pp. 269–280 (cit. on p. 3445).

- Mauricio Karchmer and Avi Wigderson (1990). “Monotone circuits for connectivity require super-logarithmic depth”. *SIAM Journal on Discrete Mathematics* 3.2, pp. 255–265 (cit. on p. 3447).
- Richard M. Karp (1972). “Reducibility among combinatorial problems”. In: *Complexity of computer computations*. Springer, pp. 85–103 (cit. on p. 3443).
- (1976). “The probabilistic analysis of some combinatorial search algorithms”. *Algorithms and complexity: New directions and recent results* 1, p. 19 (cit. on p. 3445).
- Ken-ichi Kawarabayashi and Benjamin Rossman (2018). “A Polynomial Excluded-Minor Characterization of Treedepth”. In: *29th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 234–246 (cit. on pp. 3448, 3463).
- François Le Gall (2014). “Powers of tensors and fast matrix multiplication”. In: *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation*. ACM, pp. 296–303 (cit. on p. 3443).
- Yuan Li, Alexander A. Razborov, and Benjamin Rossman (2014). “On the  $AC^0$  complexity of subgraph isomorphism”. In: *55th IEEE Symposium on Foundations of Computer Science*, pp. 344–353 (cit. on pp. 3448, 3452–3454).
- Dániel Marx (2010). “Can You Beat Treewidth?” *Theory of Computing* 6, pp. 85–112 (cit. on p. 3444).
- Dániel Marx and Michał Pilipczuk (2014). “Everything you always wanted to know about the parameterized complexity of Subgraph Isomorphism (but were afraid to ask)”. In: *31st International Symposium on Theoretical Aspects of Computer Science*, p. 542 (cit. on p. 3443).
- Jaroslav Nešetřil and Patrice Ossona de Mendez (2006). “Tree depth, subgraph coloring and homomorphism bounds”. *European J. Combin.* 27.6, pp. 1022–1041 (cit. on p. 3463).
- Jaroslav Nešetřil and Svatopluk Poljak (1985). “On the complexity of the subgraph problem”. *Comment. Math. Univ. Carolinae*. 26.2, pp. 415–419 (cit. on p. 3443).
- Toniann Pitassi and Robert Robere (2017). “Strongly exponential lower bounds for monotone computation”. In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*. ACM, pp. 1246–1255 (cit. on p. 3447).
- Jürgen Plehn and Bernd Voigt (1990). “Finding minimally weighted subgraphs”. In: *International Workshop on Graph-Theoretic Concepts in Computer Science*. Springer, pp. 18–29 (cit. on p. 3443).
- Aaron Potechin (2010). “Bounds on Monotone Switching Networks for Directed Connectivity”. In: *51st IEEE Symposium on Foundations of Computer Science*, pp. 553–562 (cit. on p. 3447).
- Ran Raz and Pierre McKenzie (1997). “Separation of the monotone NC hierarchy”. In: *38th Annual Symposium on Foundations of Computer Science*. IEEE, pp. 234–243 (cit. on p. 3447).

- Alexander A Razborov (1985). “Lower bounds on the monotone complexity of some Boolean functions”. *Dokl. Akad. Nauk SSSR* 281.4, pp. 798–801 (cit. on p. [3447](#)).
- Neil Robertson and Paul D Seymour (1991). “Graph minors. X. Obstructions to tree-decomposition”. *Journal of Combinatorial Theory, Series B* 52.2, pp. 153–190 (cit. on p. [3457](#)).
- Benjamin Rossman (2008). “On the constant-depth complexity of  $k$ -clique”. In: *40th ACM Symposium on Theory of Computing*, pp. 721–730 (cit. on p. [3448](#)).
- (2014a). “Formulas vs. circuits for small distance connectivity”. In: *46th ACM Symposium on Theory of Computing*, pp. 203–212 (cit. on pp. [3448](#), [3462](#)).
  - (2014b). “The monotone complexity of  $k$ -clique on random graphs”. *SIAM Journal on Computing* 43.1, pp. 256–279 (cit. on p. [3448](#)).
  - (2015). “Correlation Bounds Against Monotone NC<sup>1</sup>”. In: *LIPICs-Leibniz International Proceedings in Informatics*. Vol. 33. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik (cit. on pp. [3448](#), [3452](#), [3462](#)).
- Claude Shannon et al. (1949). “The Synthesis of Two-Terminal Switching Circuits”. *Bell System Technical Journal* 28.1, pp. 59–98 (cit. on p. [3445](#)).
- P.M. Spira (1971). “On time-hardware complexity tradeoffs for Boolean functions”. In: *4th Hawaii Symposium on System Sciences*, pp. 525–527 (cit. on pp. [3446](#), [3460](#)).
- Avishay Tal (2014). “Shrinkage of De Morgan Formulae by Spectral Techniques”. In: *55th IEEE Foundations of Computer Science*, pp. 551–560 (cit. on p. [3446](#)).
- Éva Tardos (1988). “The gap between monotone and non-monotone circuit complexity is exponential”. *Combinatorica* 8.1, pp. 141–142 (cit. on p. [3447](#)).

Received 2017-12-22.

BENJAMIN ROSSMAN

[ben.rossman@utoronto.ca](mailto:ben.rossman@utoronto.ca)

# ON SOME FINE-GRAINED QUESTIONS IN ALGORITHMS AND COMPLEXITY

VIRGINIA VASSILEVSKA WILLIAMS

## Abstract

In recent years, a new “fine-grained” theory of computational hardness has been developed, based on “fine-grained reductions” that focus on exact running times for problems. Mimicking NP-hardness, the approach is to (1) select a key problem  $X$  that for some function  $t$ , is conjectured to not be solvable by any  $O(t(n)^{1-\varepsilon})$  time algorithm for  $\varepsilon > 0$ , and (2) reduce  $X$  in a fine-grained way to many important problems, thus giving tight conditional time lower bounds for them. This approach has led to the discovery of many meaningful relationships between problems, and to equivalence classes.

The main key problems used to base hardness on have been: the 3-SUM problem, the CNF-SAT problem (based on the Strong Exponential Time Hypothesis (SETH)) and the All Pairs Shortest Paths Problem. Research on SETH-based lower bounds has flourished in particular in recent years showing that the classical algorithms are optimal for problems such as Approximate Diameter, Edit Distance, Frechet Distance and Longest Common Subsequence.

This paper surveys the current progress in this area, and highlights some exciting new developments.

## 1 Introduction

Arguably the main goal of the theory of algorithms is to study the worst case time complexity of fundamental computational problems. When considering a problem  $P$ , we fix a computational model, such as a Random Access Machine (RAM) or a Turing machine (TM). Then we strive to develop an efficient algorithm that solves  $P$  and to prove that for a (hopefully slow growing) function  $t(n)$ , the algorithm solves  $P$  on instances of size  $n$  in  $O(t(n))$  time in that computational model. The gold standard for the running time  $t(n)$  is linear time,  $O(n)$ ; to solve most problems, one needs to at least read the input, and so linear time is necessary.



The theory of algorithms has developed a wide variety of techniques. These have yielded near-linear time algorithms for many diverse problems. For instance, it is known since the 1960s and 70s (e.g. [Tarjan \[1971, 1972, 1974\]](#) and [Hopcroft and Tarjan \[1974\]](#)) that Depth-First Search (DFS) and Breadth-First Search (BFS) run in linear time in graphs, and that using these techniques one can obtain linear time algorithms (on a RAM) for many interesting graph problems: Single-Source Shortest paths, Topological Sort of a Directed Acyclic Graph, Strongly Connected Components, Testing Graph Planarity etc. More recent work has shown that even more complex problems such as Approximate Max Flow, Maximum Bipartite Matching, Linear Systems on Structured Matrices, and many others, admit close to linear time algorithms, by combining combinatorial and linear algebraic techniques (see e.g. [Spielman and Teng \[2004\]](#), [Christiano, Kelner, Madry, Spielman, and Teng \[2011\]](#), [Spielman and Teng \[2014\]](#), [Madry \[2013, 2016\]](#), [Cohen, Madry, Sankowski, and Vladu \[2017\]](#), [Cohen, Madry, Tsipras, and Vladu \[2017\]](#), [Cohen, Kelner, Peebles, Peng, Rao, Sidford, and Vladu \[2017\]](#), [Cohen, Y. T. Lee, Miller, Pachocki, and Sidford \[2016\]](#), and [Y. T. Lee and Sidford \[2014\]](#)).

Nevertheless, for most problems of interest, the fastest known algorithms run much slower than linear time. This is perhaps not too surprising. Time hierarchy theorems show that for most computational models, for any computable function  $t(n) \geq n$ , there exist problems that are solvable in  $O(t(n))$  time but are NOT solvable in  $O(t(n)^{1-\varepsilon})$  time for  $\varepsilon > 0$  (this was first proven for TMs [Hartmanis and Stearns \[1965\]](#), see [Papadimitriou \[1994\]](#) for more).

Time hierarchy theorems are proven by the diagonalization method pioneered by Cantor in the 19th century. Unfortunately, however, these theorems say almost nothing about *particular* problems of interest. Consider for instance the ubiquitous Boolean Satisfiability (SAT) problem: given a Boolean expression  $F$  over  $n$  variables and Boolean operators AND, OR and NOT, is there a Boolean assignment to the variables that makes  $F$  evaluate to true?

A simple algorithm to solve SAT is to try all possible  $2^n$  assignments and evaluate  $F$  on each of them. The runtime depends on how  $F$  is represented. In Circuit-SAT,  $F$  is given as a (directed acyclic) circuit with AND, OR and NOT gates,  $n$  input gates representing the variables and a designated output gate. The evaluation of a circuit can be performed in  $O(m + n)$  time, where  $m$  is the number of gates and wires, by evaluating its gates in topological order. A much more structured version of SAT is CNF-SAT. Here,  $F$  is given as a Boolean expression in Conjunctive Normal Form (CNF): an AND of  $m$  *clauses* that are ORs of *literals* (variables and their negations), i.e. one needs to satisfy every clause by setting at least one literal to TRUE. A CNF-Formula can be evaluated in  $O(m + n)$  time. Regardless, of the representation, Circuit or CNF, the enumeration of all  $2^n$  assignments dominates if  $m$  is, say, subexponential.

When the maximum clause length is a constant  $k$ , CNF-SAT can be solved in  $O^*(2^{n-cn/k})$  time for constant  $c$  independent of  $n$  and  $k$  (see e.g., [Hirsch \[1998\]](#), [Monien and Speckenmeyer \[1985\]](#), [Paturi, Pudlák, and Zane \[1999\]](#), [Paturi, Pudlák, Saks, and Zane \[2005\]](#), [Schiermeyer \[1992\]](#), and [Schöning \[1999\]](#)). Nevertheless, as  $k$  grows, this runtime approaches  $2^n$ , and the exhaustive search algorithm is essentially the best known for general CNF-SAT. For general Circuit-SAT, there is *no better* algorithm known than exhaustive search.

A natural question then is, can one prove, for a robust model of computation, that this  $2^n$  runtime dependence is inherent to the problem? Unfortunately, such a result is very far from current techniques in computational complexity. In fact, it is not even known whether SAT can be solved in *linear time*!

The only known superlinear runtime lower bounds for SAT are obtained by restricting the algorithms, for instance, to use only a small amount of space. The best of these is by R. Williams [R. Williams \[2008\]](#) who showed that if an algorithm running on a RAM uses  $n^{o(1)}$  space, then it requires at least  $n^{2\cos(\pi/7)-o(1)} \geq \Omega(n^{1.8})$  time to solve SAT on  $n$  variables. This runtime lower bound is very far from the  $2^n$  upper bound, and in fact, Buss and Williams showed [Buss and R. Williams \[2012\]](#) that this is the best result one can obtain with known techniques.

Since unconditional lower bounds seem so challenging to derive, the computer science community has long resorted to lower bounds that are conditioned on plausible, but so far unproven hypotheses. One of the most commonly used hardness hypotheses is  $P \neq NP$ . The hypothesis is formally about *decision* problems — problems whose outputs are binary — YES or NO. E.g. CNF-SAT is the decision problem that asks whether the given CNF formula is satisfiable.  $P$  is the set of decision problems that can be decided by a polynomial time algorithm<sup>1</sup> on a TM.  $NP$  is the set of decision problems that have polynomial time algorithms (on a TM) that can verify a polynomial sized solution to the instance<sup>2</sup>: e.g. CNF-SAT is in  $NP$  because we can check in polynomial time if any given Boolean assignment satisfies the formula.

$P$  vs  $NP$  asks whether all decision problems that can be verified in polynomial time (in the sense of the above paragraph), can also be decided in polynomial time.  $P$  vs  $NP$  is one of the most famous open problems in computer science. It is one of the Clay Millennium problems. While current techniques seem very far from resolving this problem, most researchers believe that  $P \neq NP$ .

The most fascinating implication of the  $P \neq NP$  hypothesis is that many problems such as SAT *cannot* be solved in polynomial time. A problem  $A$  is  $NP$ -hard if every instance of

<sup>1</sup>When we say polynomial, we mean  $O(n^c)$  for constant  $c > 0$ , where  $n$  is the size of the input instance.

<sup>2</sup>More formally,  $\pi \in NP$  if there is a polynomial time algorithm  $V$  such that if  $x$  is a YES instance of  $\pi$ , then there is a string  $y$  of size  $O(|x|^c)$  for some constant  $c$ , such that  $V(x, y)$  returns YES, and if  $x$  is a NO instance,  $V(x, y)$  returns NO for all  $y$ .

every problem in NP can be encoded in polynomial time as an instance of  $A$ . A problem in NP which is NP-hard is called NP-Complete.

Clearly, if an NP-hard problem has a polynomial time algorithm, then  $P = NP$ . Thus, if we assume that  $P \neq NP$ , no NP-hard problem can have a polynomial time algorithm. Starting with the work of Cook and Levin (who showed that SAT is NP-complete) and Karp (who added 21 more NP-complete problems), NP-hardness took off. Now there are many thousands of problems known to be NP-hard.

NP-hardness is arguably the biggest export of theoretical computer science (TCS) to other disciplines. It is routinely used to explain why it is so hard to find efficient algorithms for problems occurring in practice, and why one should probably use specialized algorithms and heuristics to solve them.

$P$  and NP are defined for TMs. However, due to polynomial time reductions between computational models (see [van Emde Boas \[1990\]](#) for a thorough treatment), whether we consider a TM or a RAM in the definition of  $P$  and NP does not actually matter.  $P$  vs NP is essentially model-independent. This model-independence was one of the reasons to focus on *polynomial* time as a model of efficiency. (Another simple reason is that polynomials compose into polynomials.) Nevertheless, no-one would argue that all polynomial runtimes are actually efficient. In fact, for today's large inputs, even *quadratic* time is inefficient.

There are many fundamental problems for which the fastest known algorithms run in quadratic time or slower. A simple example is the *Edit Distance* problem, with many diverse applications from computational biology to linguistics: given two strings  $\alpha$  and  $\beta$ , over some finite alphabet, what is the smallest sequence of symbol insertions, deletions and substitutions that can be performed on  $\alpha$  to obtain  $\beta$ ?

The problem has a long history: a classical dynamic programming algorithm by [Wagner and Fischer \[1974\]](#) runs in  $O(n^2)$  time, and despite many efforts, the best known algorithm [Masek and Paterson \[1980\]](#) only shaves a  $\log^2 n$  factor. On inputs where  $n$  is in the billions (such as the human genome), quadratic runtime is prohibitive.

Another simple problem, from Computational Geometry, asks for a given set of points in the plane, are any three colinear; conversely, are the points in general position. Before running a computational geometry algorithm, one typically needs to check this important primitive. Unfortunately, the best known algorithms for this Colinearity question for  $n$  points run in  $n^{2-o(1)}$ , i.e. quadratic time.

There are many such examples within  $P$ , from a vast variety of research areas. Why has it been so hard to find faster algorithms for such problems? Addressing this is impossible using  $P \neq NP$  as an assumption: no problem that is already in  $P$  can be NP-Complete, unless  $P = NP$ . We need a different approach.

**The Fine-Grained Question.** Let us delve further into the issue described above. From now on let us fix the computational model to a word-RAM with  $O(\log n)$  bit words. Informally, this is a RAM machine that can read from memory, write to memory and perform operations on  $O(\log n)$  bit chunks of data in constant time. We can fix any computational model; we pick the word-RAM because it is simple to work with.

When faced with a computational problem  $P$ , we can usually apply well-known algorithmic techniques (such as dynamic programming, greedy, divide and conquer etc.) and come up with a simple algorithm that runs in  $O(t(n))$  time on inputs of size  $n$ .

Often this algorithm is obtained by the brute-force approach — enumerate all candidate solutions in a search space. This is the case for SAT, but also for a large variety of other problems.

Sometimes the simple algorithm is not brute-force but uses textbook techniques in the natural way. Consider for instance the Longest Common Subsequence problem (LCS), a simplified version of Edit Distance: given two sequences  $A$  and  $B$  of length  $n$  over an alphabet  $\Sigma$ , determine the maximum length sequence  $S$  that appears in both  $A$  and  $B$ , with symbols in the same order, but possibly not consecutively. For instance, an LCS of  $(b, b, c, a, d, e)$  and  $(b, d, c, b, e)$  is  $(b, c, e)$ . The textbook approach here is to apply dynamic programming, concluding that  $\text{LCS}((b, b, c, a, d, e), (b, d, c, b, e))$  is the longest of  $\text{LCS}((b, b, c, a, d), (b, d, c, b)) \odot e$ ,  $\text{LCS}((b, b, c, a, d, e), (b, d, c, b))$ , and  $\text{LCS}((b, b, c, a, d), (b, d, c, b, e))$ . The runtime is  $O(n^2)$  since throughout the computation, one at most needs to memoize the computed longest common subsequences of  $n^2$  pairs of prefixes. (The textbook algorithm for Edit Distance is similar.)

More often than not, the obtained “textbook” running time seems difficult to improve upon: improvements have been sought after for decades, and the simple algorithm has stood almost unchallenged. We mentioned earlier that this is the case for CNF-SAT, Co-linearity and Edit Distance. The situation is similar for LCS and Edit Distance (fastest runtime  $O(n^2 / \log^2 n)$  [Masek and Paterson \[ibid.\]](#) for constant size alphabet and otherwise  $O(n^2 \log \log n / \log^2 n)$  [Bille and Farach-Colton \[2008\]](#) and [Grabowski \[2014\]](#)), and for a large variety of other problems from all over computer science and beyond. The central question that needs to be addressed is:

*For each of the problems of interest with textbook runtime  $O(t(n))$  and nothing much better known, is there a barrier to obtaining an  $O(t(n)^{1-\varepsilon})$  time algorithm for  $\varepsilon > 0$ ?*

Relatedly, is the reason for this difficulty the same for all problems of interest?

## 2 Fine-Grained Complexity (and algorithms)

We would like to mimic NP-Completeness. The approach will be as follows.

1. We will identify some believable *fine-grained* hardness hypotheses. These will be about specific conjectured running times for very well-studied computational problems.
2. Using *fine-grained* reductions we will show that for a problem with textbook running time  $t(n)$ , obtaining an  $O(t(n)^{1-\varepsilon})$  time algorithm for  $\varepsilon > 0$  would violate one or more of the hypotheses. The reductions we employ cannot be mere polynomial time reductions - they would have to be tailored to the specific textbook runtime  $t(n)$ . As we will see, they will differ in other ways as well from most reductions used in traditional complexity.

We would also like to give equivalences, i.e. to show that problem  $A$  with textbook running time  $a(n)$  and problem  $B$  with textbook running time  $b(n)$  are equivalent in the sense that if  $A$  admits an  $a(n)^{1-\varepsilon}$  time algorithm for  $\varepsilon > 0$ , then  $B$  admits an  $b(n)^{1-\varepsilon'}$  time algorithm for some  $\varepsilon' > 0$ . This would mean that the reason why it has been hard to improve on  $A$  and on  $B$  is the same.

In the following we will discuss some of the most prominent hardness hypotheses in fine-grained complexity, and the reductions we employ to achieve fine-grained hardness.

**2.1 Key Hypotheses.** Much of fine-grained complexity is based on hypotheses of the time complexity of three problems: CNF-SAT, All-Pairs Shortest Paths (APSP) and 3-SUM. Below we will introduce these, and a few more related hypotheses. There are no known reductions between CNF-SAT, APSP and 3-SUM: they are potentially unrelated. All hypotheses are about the word-RAM model of computation with  $O(\log n)$  bit words, where  $n$  is the size of the input.

**SETH..** [Impagliazzo and Paturi \[2001\]](#) introduced the Strong Exponential Time Hypothesis (SETH) to address the complexity of CNF-SAT. At the time they only considered deterministic algorithms, but nowadays it is common to extend SETH to allow randomization.

**Hypothesis 1** (Strong Exponential Time Hypothesis (SETH)). *For every  $\varepsilon > 0$  there exists an integer  $k \geq 3$  such that CNF-SAT on formulas with clause size at most  $k$  (the so called  $k$ -SAT problem) and  $n$  variables cannot be solved in  $O(2^{(1-\varepsilon)n})$  time even by a randomized algorithm.*

As the clause size  $k$  grows, the lower bound given by SETH converges to  $2^n$ . SETH also implies that general CNF-SAT on formulas with  $n$  variables and  $m$  clauses requires  $2^{n-o(n)}$  poly( $m$ ) time.

SETH is motivated by the lack of fast algorithms for  $k$ -SAT as  $k$  grows. It is a much stronger assumption than  $P \neq NP$  which assumes that SAT requires superpolynomial time.

A weaker version, the Exponential Time Hypothesis (ETH) asserts that there is some constant  $\delta > 0$  such that CNF-SAT requires  $\Omega(2^{\delta n})$ .

Both ETH and SETH are used within Fixed Parameter and Exponential Time algorithms as hardness hypotheses, and they imply meaningful hardness results for a variety of problems (see e.g. [Cygan, Fomin, Kowalik, Lokshtanov, Marx, Pilipczuk, Pilipczuk, and Saurabh \[2015\]](#)). Because we are concerned with tight, fine-grained, runtime bounds, we focus on SETH as opposed to ETH.

**3-SUM Hypothesis.** The 3-SUM problem is as follows: given a set  $S$  of  $n$  integers from  $\{-n^c, \dots, n^c\}$  for some constant  $c$ , determine whether there are  $x, y, z \in S$  such that  $x + y + z = 0$ . A standard hashing trick allows us to assume that  $c \leq 3 + \delta$  for any  $\delta > 0$ .<sup>3</sup>

**Hypothesis 2 (3-SUM Hypothesis).** *3-SUM on  $n$  integers in  $\{-n^4, \dots, n^4\}$  cannot be solved in  $O(n^{2-\varepsilon})$  time for any  $\varepsilon > 0$  by a randomized algorithm.*

The hypothesis was introduced by [Gajentaan and M. Overmars \[1995\]](#) and [Gajentaan and M. H. Overmars \[2012\]](#) who used it to show that many problems in computational geometry require quadratic time, assuming that 3-SUM does. Quadratic lower bounds for 3-SUM are known in restricted models of computation such as the linear decision tree model in which each decision is based on the sign of an affine combination of at most 3 inputs (see e.g. [Erickson and Seidel \[1995\]](#) and [Erickson \[1995\]](#)). However, in the more general linear decision tree model, Kane et al. [Kane, Lovett, and Moran \[2017\]](#) show that  $O(n \log^2 n)$  queries suffice to solve 3-SUM, so that such lower bounds should be taken with a grain of salt.

The 3-SUM problem is very simple and has been studied extensively. The textbook algorithm is a simple  $O(n^2 \log n)$  time enumeration algorithm: sort  $S$  and then for every  $x, y \in S$ , check if  $-z \in S$  using binary search. An  $O(n^2)$  runtime can be obtained by traversing the sorted order of  $S$  in both directions. Baran, Demaine and Pătraşcu [Baran, E. Demaine, and Pătraşcu \[2008\]](#) improved this running time to  $O(n^2 (\log \log n)^2 / \log^2 n)$  time. If the input numbers are real numbers instead of integers (now in the Real-RAM model of computation), [Jørgensen and Pettie \[2014\]](#) gave an  $O(n^2 (\log \log n)^{2/3} / \log^{2/3} n)$  time algorithm. This runtime was recently improved by [Chan \[2018\]](#) to  $n^2 (\log \log n)^{O(1)} / \log^2 n$ , almost matching the known running time for integer inputs.

---

<sup>3</sup>One can pick a random prime  $p$  that is between  $n$  and  $n^{3+\delta}$ . The number of distinct primes in this range that can divide any particular sum of three input integers is  $O(1)$ , and hence the total number of distinct primes that can divide the sum of some three input integers is  $O(n^3)$ . However, there are  $\Omega(n^{3+\delta'})$  primes in the interval between  $n$  and  $n^{3+\delta}$ , for any  $0 < \delta' < \delta$ , and the probability that  $p$  divides one of the sums from  $S$  is  $\leq O(1/n^{\delta'})$ . We can then reduce 3-SUM mod  $p$  to three instances of the original 3-SUM problem with integers in the range  $\{-2p, \dots, p-1\}$  — checking if  $x, y, z$  sum to 0,  $p$  or  $2p$ .

**All-Pairs Shortest Paths (APSP)..** The APSP problem is as follows: given an  $n$  node graph  $G = (V, E)$ , and integer edge weights  $w : E \rightarrow \{-M, \dots, M\}$  for some  $M = \text{poly}(n)$ , compute for every  $u, v \in V$ , the (shortest path) distance  $d(u, v)$  in  $G$  from  $u$  to  $v$ , i.e. the minimum over all paths from  $u$  to  $v$  of the total weight sum of the edges of the path.  $G$  is assumed to contain no negative weight cycles.

The textbook algorithm for APSP is the  $O(n^3)$  time Floyd-Warshall algorithm from the 1960s based on dynamic programming. Many other algorithms run in the same time. For instance, one can run Dijkstra’s algorithm from every vertex, after computing new nonnegative edge weights using Johnson’s trick [Johnson \[1977\]](#). Following many polylogarithmic improvements (e.g. [chan06](#); [Fredman \[1976\]](#)), the current best APSP running time is a breakthrough  $n^3 / \exp(\sqrt{\log n})$  runtime by R. [Williams \[2014\]](#). Despite the long history, the cubic runtime of the textbook algorithm has remained unchallenged. This motivates the APSP Hypothesis below, implicitly used in many papers (e.g. [Roditty and Zwick \[2004\]](#)). Its first explicit use as a hardness hypothesis is in [Vassilevska Williams and R. Williams \[2010\]](#).

**Hypothesis 3** (APSP Hypothesis). *No randomized algorithm can solve APSP in  $O(n^{3-\varepsilon})$  time for  $\varepsilon > 0$  on  $n$  node graphs with edge weights in  $\{-n^c, \dots, n^c\}$  and no negative cycles for large enough  $c$ .*

**2.2 Fine-grained reductions.** Our goal is as follows. Consider problem  $A$  with textbook runtime  $a(n)$  and problem  $B$  with textbook runtime  $b(n)$ . Given a supposed  $O(b(n)^{1-\varepsilon})$  time algorithm for  $B$  for  $\varepsilon > 0$ , we would like to compose it with another algorithm (the reduction) that transforms instances of  $A$  into instances of  $B$ , to obtain an algorithm for  $A$  running in time  $O(a(n)^{1-\varepsilon'})$  time for  $\varepsilon' > 0$  (a function of  $\varepsilon$ ).

The most common reductions used in complexity are polynomial time (or sometimes logspace) reductions. For our purposes such reductions are not sufficient since we truly care about the runtimes  $a(n)$  and  $b(n)$  that we are trying to relate, and our reductions need to run faster than  $a(n)$  time for sure; merely polynomial time does not suffice. In turn, if  $a(n)$  is super-polynomial, we would like to allow ourselves super-polynomial time in the reduction – there is no reason to restrict the reduction runtime to a polynomial.

Beyond the time restriction, reductions differ in whether they are Karp or Turing reductions. Karp (also called many-one) reductions transform an instance of  $A$  into a single instance of  $B$ . Turing reductions are allowed to produce multiple instances, i.e. oracle calls to  $B$ . If we restrict ourselves to Karp-style reductions, then we wouldn’t be able to reduce a search problem to any decision problem: decision problems return a single bit and if we only make one oracle call to a decision problem, in general we would not get enough information to solve the original search problem. We hence use Turing-style reductions.

The most general definition is:

**Definition 2.1** (Fine-grained reduction). *Assume that  $A$  and  $B$  are computational problems and  $a(n)$  and  $b(n)$  are their conjectured running time lower bounds, respectively. Then we say  $A$   $(a, b)$ -reduces to  $B$ ,  $A \leq_{a,b} B$ , if for every  $\varepsilon > 0$ , there exists  $\delta > 0$ , and an algorithm  $R$  for  $A$  that runs in time  $a(n)^{1-\delta}$  on inputs of length  $n$ , making  $q$  calls to an oracle for  $B$  with query lengths  $n_1, \dots, n_q$ , where*

$$\sum_{i=1}^q (b(n_i))^{1-\varepsilon} \leq (a(n))^{1-\delta}.$$

*If  $A \leq_{a,b} B$  and  $B \leq_{b,a} A$ , we say that  $A$  and  $B$  are fine-grained equivalent,  $A \equiv_{a,b} B$ .*

The definition implies that if  $A \leq_{a,b} B$  and  $B$  has an algorithm with running time  $O(b(n)^{1-\varepsilon})$ , then,  $A$  can be solved by replacing the oracle calls by the corresponding runs of the algorithm, obtaining a runtime of  $O(a(n)^{1-\delta})$  for  $A$  for some  $\delta > 0$ . If  $A \equiv_{a,b} B$ , then arguably the reason why we have not been able to improve upon the runtimes  $a(n)$  and  $b(n)$  for  $A$  and  $B$ , respectively, is the same.

Notice that the oracle calls in the definition need not be independent — the  $i$ th oracle call might be adaptively chosen, according to the outcomes of the first  $i - 1$  oracle calls.

### 3 Hardness results from SETH

SETH was first used to give conditional hardness for other NP-hard problems. For instance, Cygan et al. [Cygan, Dell, Lokshtanov, Marx, Nederlof, Okamoto, Paturi, Saurabh, and Wahlström \[2016\]](#) show that several other problems (such as  $k$ -Hitting Set and  $k$ -NAE-SAT) are equivalent to  $k$ -SAT, in that an  $O(2^{(1-\varepsilon)n})$  time algorithm for  $\varepsilon > 0$  for one of them (for all  $k$ ) would imply such an algorithm for all of them, and would refute SETH.

The introduction of SETH as a hardness hypothesis for polynomial time problems was initiated by R. Williams [R. Williams \[2005\]](#). Among other things, Williams shows that the so called *Orthogonal Vectors* (OV) problem, a problem in quadratic time, requires quadratic time under SETH. We will describe the reduction shortly.

**Orthogonal Vectors.** The OV problem, and its generalization  $k$ -OV, form the basis of many fine-grained hardness results for problems in P.

The *OV problem* is defined as follows: Let  $d = \omega(\log n)$ ; given two sets  $A, B \subseteq \{0, 1\}^d$  with  $|A| = |B| = n$ , determine whether there exist  $a \in A, b \in B$  so that  $a \cdot b = 0$  where  $a \cdot b = \sum_{i=1}^d a[i] \cdot b[i]$ .



The  $k$ -OV problem for constant  $k \geq 2$  is the generalization of OV to  $k$  sets: Let  $d = \omega(\log n)$ ; given  $k$  sets  $A_1, \dots, A_k \subseteq \{0, 1\}^d$  with  $|A_i| = n$  for all  $i$ , determine whether there exist  $a_1 \in A_1, \dots, a_k \in A_k$  so that  $a_1 \cdot \dots \cdot a_k = 0$  where  $a_1 \cdot \dots \cdot a_k := \sum_{i=1}^d \prod_{j=1}^k a_j[i]$ .

OV is a special case of Hopcroft's problem: given two sets  $R$  and  $B$  of  $n$  vectors each in  $\mathbb{R}^d$ , detect  $r \in R, b \in B$  such that  $(r, b) = 0$  (an equivalent version that Hopcroft posed is when we are given points and hyperplanes through the origin, and we want to detect a point lying on one of the hyperplanes). The fastest algorithms for Hopcroft's problem for general  $d$  run in  $2^{O(d)} n^{2-\Theta(d)}$  time [Matoušek \[1993\]](#) and [Chazelle \[1993\]](#).

OV is equivalent to the Batch Subset Query problem from databases [Ramasamy, Patel, Naughton, and Kaushik \[2000\]](#), [Goel and Gupta \[2010\]](#), [Agrawal, Arasu, and Kaushik \[2010\]](#), and [Melnik and Garcia-Molina \[2003\]](#): given two sets  $S$  and  $T$  of sets over  $[d]$ , check if there is some  $s \in S, t \in T$  such that  $s \subseteq t$ . It is also known to be equivalent to the classical Partial Match problem.

It is not hard to solve  $k$ -OV in  $O(n^k d)$  time by exhaustive search, for any  $k \geq 2$ . The fastest known algorithms for the problem run in time  $n^{k-1/\Theta(\log(d/\log n))}$  [Abboud, R. R. Williams, and Yu \[2015\]](#) and [Chan and R. Williams \[2016\]](#). It seems that  $n^{k-o(1)}$  is necessary. This motivates the now widely used  $k$ -OV Hypothesis.

**Hypothesis 4** ( $k$ -OV Hypothesis). *No randomized algorithm can solve  $k$ -OV on instances of size  $n$  in  $n^{k-\varepsilon} \text{poly}(d)$  time for constant  $\varepsilon > 0$ .*

Interestingly, Williams and Yu [R. Williams and Yu \[2014\]](#) show that the 2-OV Hypothesis is false when operations are over the ring  $\mathbb{Z}_m$ , or over the field  $\mathbb{F}_m$  for any prime power  $m = p^k$ . In the first case, OV can be solved in  $O(nd^{m-1})$  time, and in the second case, in  $O(nd^{p(k-1)})$  time. Although the problem is easier in these cases, [R. Williams and Yu \[ibid.\]](#) actually also show that these runtimes cannot be improved very much, unless SETH fails, so there is still some hidden hardness. Over  $\mathbb{Z}_6$ , it turns out that OV does still require quadratic time under SETH: no  $n^{2-\varepsilon} d^{o_d(\log d / \log \log d)}$  time algorithm  $\varepsilon > 0$  can exist.

Gao et al. [Gao, Impagliazzo, Kolokolova, and R. R. Williams \[2017\]](#) show that OV is complete for a large class of problems: the class of all first order properties. They consider properties expressible by a first-order formula with  $k+1$  quantifiers on a given structure with  $m$  records; checking if any such property holds can easily be done in  $O(m^k)$  time, and [Gao, Impagliazzo, Kolokolova, and R. R. Williams \[ibid.\]](#) give an improved  $m^k / 2^{\Theta(\sqrt{\log m})}$  time algorithm. The completeness of OV is as follows. The First-Order Property Conjecture (FOPC) [Gao, Impagliazzo, Kolokolova, and R. R. Williams \[ibid.\]](#) asserts that there is some  $k \geq 2$  s.t. for all  $\varepsilon > 0$  there is a first order property on  $k+1$  quantifiers that cannot be decided in  $O(m^{k-\varepsilon})$  time. Gao et al. [Gao, Impagliazzo, Kolokolova, and R. R. Williams \[ibid.\]](#) show that FOPC is equivalent to the 2-OV hypothesis.

Here we present Williams' [R. Williams \[2005\]](#) result that  $k$ -OV requires essentially  $n^k$  time, under SETH. Afterwards we will see some applications of this result.

**Theorem 3.1** ([R. Williams \[ibid.\]](#)). *If  $k$ -OV on sets with  $N$  vectors from  $\{0, 1\}^m$  can be solved in  $N^{k-\varepsilon} \text{poly}(m)$  time for any  $\varepsilon > 0$ , then CNF-SAT on  $n$  variables and  $m$  clauses can be solved in  $2^{(1-\varepsilon')n} \text{poly}(m)$  time for some  $\varepsilon' > 0$  and SETH is false.*

*Proof.* We present a fine-grained reduction from CNF-SAT to  $k$ -OV. Let the given formula  $F$  have  $n$  variables and  $m$  clauses. Split variables into  $k$  parts  $V_1, \dots, V_k$  on  $n/k$  variables each. For every  $j = 1, \dots, k$  create a set  $A_j$  containing a length  $m$  binary vector  $a^j(\phi)$  for every one of the  $N = 2^{n/k}$  Boolean assignments  $\phi$  to the variables in  $V_j$ , where

$$a^j(\phi)[c] = 0 \text{ if the } c\text{th clause of } F \text{ is satisfied by } \phi, \text{ and } 1 \text{ otherwise.}$$

The instance of  $k$ -OV formed by  $A_1, \dots, A_k$  has all  $|A_j| = N = 2^{n/k}$ .

Suppose that for some  $a_1(\phi_1) \in A_1, \dots, a_k(\phi_k) \in A_k$ , we have  $\sum_c \prod_j a_j(\phi_j)[c] = 0$ , then for every clause  $c$ , there is some vector  $a_j(\phi_j)$  that is 0 in clause  $c$ , and hence the Boolean assignment  $\phi_j$  to the variables in  $V_j$  satisfies clause  $c$ . Thus, the concatenation  $\phi_1 \odot \dots \odot \phi_k$  is a Boolean assignment to all variables  $V$  of  $F$  that satisfies all clauses. Conversely, if  $\phi$  satisfies all clauses, then we define  $\phi_j$  to be the restriction of  $\phi$  to  $V_j$ , and we see that  $\sum_c \prod_j a_j(\phi_j)[c] = 0$ , as every clause must be satisfied by some  $\phi_j$ .

If  $k$ -OV on  $k$  sets of  $N$  vectors each in  $\{0, 1\}^m$  can be solved in  $N^{k-\varepsilon} \text{poly}(m)$  time, then CNF-SAT on  $n$  variables and  $m$  clauses can be solved in time  $(2^{n/k})^{k-\varepsilon} \text{poly}(m) = 2^{n-\varepsilon'} \text{poly}(m)$  time for  $\varepsilon' = \varepsilon/k > 0$ . This contradicts SETH.  $\square$

We note that due to the Sparsification Lemma [Impagliazzo and Paturi \[2001\]](#), one can assume that the  $n$ -variable  $\ell$ -CNF instance that one reduces to  $k$ -OV has  $O(n)$  clauses. Thus, to refute SETH, one only needs to obtain an  $N^{k-\varepsilon} \text{poly}(d)$  time algorithm for  $\varepsilon > 0$  for  $k$ -OV where the dimension  $d$  of the vectors is any slowly growing function of  $N$  that is  $\omega(\log N)$ , for instance  $d = \log^2 N$ .

Besides  $k$ -OV, Williams also considers the  $k$ -Dominating set problem: for a fixed constant  $k$ , given an  $n$  node graph  $G = (V, E)$ , determine whether there is a subset  $S \subseteq V$  of size  $k$  so that for every  $v \in V$  there is some  $s \in S$  so that  $(s, v) \in E$ . Williams ([R. Williams \[2007b\]](#), later in [Pătraşcu and R. Williams \[2010\]](#)) shows via a reduction from CNF-SAT,  $k$ -Dominating set requires  $n^{k-o(1)}$  time. The reduction from CNF-SAT to  $k$ -OV can be routed through  $k$ -Dominating Set, showing that that problem is in a sense between CNF-SAT and  $k$ -OV.

The  $k$ -OV problem is the basis for most reductions from CNF-SAT to problems within Polynomial Time. We will give two examples, and will then give a short summary of most known results.

It is simple to reduce  $k$ -OV to  $(k - 1)$ -OV: go over all vectors in the first set, and solve a  $(k - 1)$ -OV instance for each. Hence 2-OV is the hardest out of all  $k$ -OV problems. Also,  $k$ -OV is potentially strictly harder than SETH. Thus, even if SETH turns out to be false, the  $k$ -OV Hypothesis might still hold.

**Orthogonal Vectors and Graph Diameter.** Arguably the first reduction from SETH to a graph problem in P is from a paper by [Roditty and Vassilevska Williams \[2013\]](#) that considers the Diameter problem: given an  $n$  node,  $m$  edge graph  $G = (V, E)$ , determine its diameter, i.e.  $\max_{u,v \in V} d(u, v)$ .

For directed or undirected graphs with arbitrary (real) edge weights, the fastest known algorithm for the Diameter problem computes all the pairwise distances in  $G$ , solving APSP. As mentioned earlier, the fastest known algorithm for APSP in dense graphs runs in  $n^3 / \exp(\sqrt{\log n})$  time. For sparser graphs, the fastest known algorithms run in  $\tilde{O}(mn)$  time<sup>4</sup> [Pettie and Ramachandran \[2005\]](#), [Pettie \[2004\]](#), and [Pettie \[2008\]](#).

If the graph is unweighted, one can solve Diameter in  $\tilde{O}(n^\omega)$  time, where  $\omega < 2.373$  [Vassilevska Williams \[2012\]](#) and [Le Gall \[2014\]](#) is the exponent of square matrix multiplication. If the graph has small integer edge weights in  $\{0, \dots, M\}$ , Diameter is in  $\tilde{O}(Mn^\omega)$  time (see e.g. [Cygan, Gabow, and Sankowski \[2012\]](#)). However, since  $\omega \geq 2$ , all known algorithms for Diameter run in  $\Omega(n^2)$  time, even when the graph is unweighted, and undirected<sup>5</sup>, and has  $m \leq O(n)$  edges.

With a simple reduction, [Roditty and Vassilevska Williams \[2013\]](#) show that under SETH, the Diameter problem in undirected unweighted graphs with  $n$  nodes and  $O(n)$  edges requires  $n^{2-o(1)}$  time. Moreover, their reduction shows that under SETH, even distinguishing between graphs with Diameter 2 and 3 requires  $n^{2-o(1)}$  time, and hence no  $3/2 - \varepsilon$  approximation algorithm can run in  $O(n^{2-\delta})$  time for  $\varepsilon, \delta > 0$  even on sparse graphs.

[Chechik, Larkin, Roditty, Schoenebeck, Tarjan, and Vassilevska Williams \[2014\]](#) obtained a  $3/2$ -approximation algorithm for Diameter that runs in  $\tilde{O}(m^{3/2})$  time in  $m$ -edge graphs; their algorithm was based on a previous  $\tilde{O}(m\sqrt{n})$  time algorithm from [Roditty and Vassilevska Williams \[2013\]](#) that is a  $3/2$  approximation when the diameter is divisible by 3 (and slightly worse otherwise). This algorithm is thus in a sense optimal, under SETH: it runs in truly subquadratic time in sparse graphs, but if one wants to improve upon the approximation factor even slightly, all of a sudden  $n^{2-o(1)}$  time is needed. An  $\tilde{O}(n^2)$  runtime in sparse graphs is very easy to achieve: just solve APSP by running BFS

<sup>4</sup>  $\tilde{O}(f(n))$  denotes  $f(n)\text{polylog}(n)$ .

<sup>5</sup> All shortest paths problems, including Diameter, are at least as hard in directed graphs as they are in undirected graphs; similarly, they are at least as hard in weighted graphs as they are in unweighted graphs, and at least as hard in denser graphs than they are in sparser graphs.

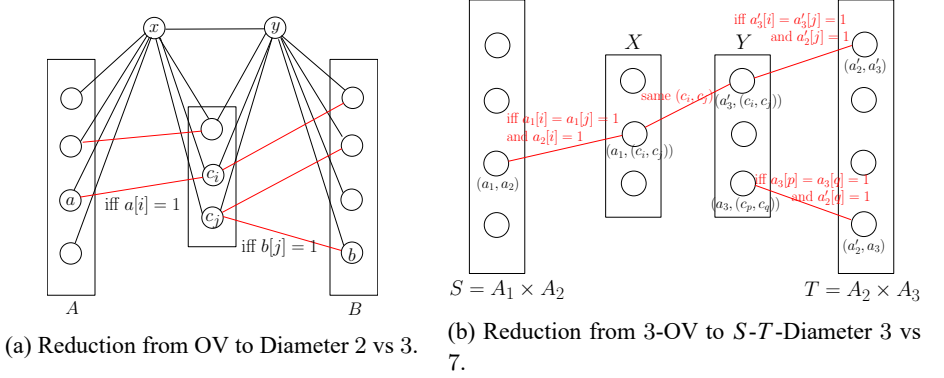


Figure 1: Two reductions to Diameter Problems.

from every node! Thus, under SETH, there is essentially nothing more to do besides the easy algorithm, for approximation below  $3/2$ .

**Theorem 3.2.** *If one can distinguish between Diameter 2 and 3 in an undirected unweighted graph with  $O(N)$  nodes and edges in  $O(N^{2-\varepsilon})$  time for some  $\varepsilon > 0$ , then 2-OV on two sets of  $n$  vectors in  $d$  dimensions can be solved in  $n^{2-\varepsilon}\text{poly}(d)$  time and SETH is false.*

*Proof.* Suppose we are given an instance of 2-OV,  $A, B$  of  $n$  vectors each in  $\{0, 1\}^d$ , where  $|A| = |B| = n$ . Let's create a graph  $G$  as follows. See Figure 1a.

For every vector  $a \in A$ , create a node  $a$  of  $G$ . For every vector  $b \in B$ , create a node  $b$  of  $G$ . For every  $i \in [d]$ , create a node  $c_i$ . We add two additional nodes  $x$  and  $y$ .

The edges are as follows. For every  $a \in A$  and  $i \in [d]$ , if  $a[i] = 1$ , add an edge between  $a$  and  $c_i$  in  $G$ . Similarly, for every  $b \in B$  and  $i \in [d]$ , if  $b[i] = 1$ , add an edge between  $b$  and  $c_i$  in  $G$ .

The edges incident to  $x$  are as follows:  $(x, a)$  for every  $a \in A$ ,  $(x, c_i)$  for every  $i \in [d]$  and  $(x, y)$ . The edges incident to  $y$  are as follows:  $(y, b)$  for every  $b \in B$  and  $(y, c_i)$  for every  $i \in [d]$  (and  $(x, y)$ ).

Now, if  $a \in A$  and  $b \in B$  are not orthogonal, then there is some  $i$  such that  $a[i] = b[i] = 1$ , and so  $d(a, b) = 2$  via the path through  $c_i$ . Otherwise, if  $a$  and  $b$  are orthogonal, then there is no such  $c_i$  and the shortest  $a - b$  path goes through  $(x, y)$ , and  $d(a, b) = 3$ . All nodes in the graph are at distance at most 2 to  $x, y$ , and each  $c_i$ , and hence the Diameter is 3 if there is an orthogonal pair, and 2 otherwise.

Let  $N = nd$ . The number of nodes and edges is at most  $O(N)$ . If Diameter 2 vs 3 can be solved in  $O(N^{2-\varepsilon})$  time for some  $\varepsilon > 0$ , then 2-OV is in  $O((nd)^{2-\varepsilon}) \leq n^{2-\varepsilon}\text{poly}(d)$  time for  $\varepsilon > 0$ .  $\square$

By the above result, we get that under the 2-OV Hypothesis, improving upon the approximation factor of the known Diameter algorithms [Roditty and Vassilevska Williams \[2013\]](#) and [Chechik, Larkin, Roditty, Schoenebeck, Tarjan, and Vassilevska Williams \[2014\]](#) is impossible without blowing up the running time to  $n^{2-o(1)}$ . However, all known  $3/2$ -approximation algorithms run in  $\tilde{O}(n^{1.5})$  time in sparse graphs. Can this runtime be improved? Can it be made linear?

Cairo et al. [Cairo, Grossi, and Rizzi \[2016\]](#) presented faster approximation algorithms for Diameter. Generalizing [Roditty and Vassilevska Williams \[2013\]](#) and [Chechik, Larkin, Roditty, Schoenebeck, Tarjan, and Vassilevska Williams \[2014\]](#), they presented for every integer  $k \geq 1$ , an  $\tilde{O}(mn^{1/(k+1)})$  time algorithm that is a  $2 - 1/2^k$ -approximation to the Diameter of undirected graphs (if it is divisible by  $2^{k+1} - 1$ , and slightly worse otherwise). Unfortunately, the approximation quality degrades as the runtime decreases. Thus their results do not answer the question of whether there are faster  $3/2$ -approximation algorithms.

In recent work, Backurs et al. [Backurs, Roditty, Segal, Vassilevska Williams, and Wein \[2018\]](#) show that unless the 3-OV Hypothesis (and hence SETH) is false, any  $3/2$ -approximation algorithm to the Diameter in sparse graphs needs  $n^{1.5-o(1)}$  time, thus resolving the question. They also obtain a variety of other tight conditional lower bounds based on  $k$ -OV for different  $k$  for graph Eccentricities, and variants of Diameter.

The hardness result for  $3/2$ -approximate Diameter is based on a hardness construction for a slightly more difficult problem called  $S$ - $T$  Diameter. In it, one is given a graph  $G = (V, E)$  and two subsets  $S, T \subseteq V$  and is asked to compute  $D_{S,T} := \max_{s \in S, t \in T} d(s, t)$ , the so called  $S$ - $T$  Diameter which is the largest distance between a node of  $S$  and a node of  $T$ .

When it comes to exact computation in sparse weighted graphs,  $S$ - $T$  Diameter is  $(n^2, n^2)$ -equivalent to Diameter (see [Backurs, Roditty, Segal, Vassilevska Williams, and Wein \[ibid.\]](#)). When it comes to approximation, the problems differ a bit. In linear time, Diameter admits a 2-approximation, while  $S$ - $T$  Diameter admits a 3-approximation. In  $\tilde{O}(m^{3/2})$  time, Diameter admits a  $3/2$ -approximation, whereas  $S$ - $T$  Diameter admits a 2-approximation. Thus, the starting point of the hardness for  $3/2$ -approximate Diameter is a hardness construction for 2-approximate  $S$ - $T$  Diameter.

**Theorem 3.3** ([Backurs, Roditty, Segal, Vassilevska Williams, and Wein \[ibid.\]](#)). *Under the 3-OV Hypothesis, no  $O(N^{1.5-\varepsilon})$  time algorithm for  $\varepsilon > 0$ , can distinguish between  $S$ - $T$  Diameter 3 and 7 in graphs with at most  $N$  nodes and edges.*

Since any 2-approximation algorithm can distinguish between  $S$ - $T$  Diameter 3 and 7, the Theorem above implies that  $n^{1.5-o(1)}$  time is needed to 2-approximate the  $S$ - $T$  Diameter of a sparse graph. We will present the proof of [Theorem 3.3](#). To complete the reduction to Diameter, some extra gadgets are needed; these create a graph in which the

Diameter is either 5 or 9 and thus give hardness for  $3/2$ -Diameter approximation. We refer the reader to the presentation in [Backurs, Roditty, Segal, Vassilevska Williams, and Wein \[ibid.\]](#). [Theorem 3.2](#) and the extension to [Theorem 3.3](#) to Diameter can be generalized to a reduction from  $k$ -OV for arbitrary  $k$  to Diameter, thus showing a time/approximation tradeoff lower bound [Backurs, Roditty, Segal, Vassilevska Williams, and Wein \[ibid.\]](#).

*Proof Sketch of Theorem 3.3.* Let  $A_1, A_2, A_3 \subseteq \{0, 1\}^d$  be the  $n$ -sets forming the 3-OV instance.

For every pair of vectors  $a_1 \in A_1, a_2 \in A_2$ , we create a node  $(a_1, a_2)$  in a set  $S$ . For every pair of vectors  $a_2 \in A_2, a_3 \in A_3$ , we create a node  $(a_2, a_3)$  in a set  $T$ .

For every node  $a_1 \in A_1$  and every pair of coordinates  $i, j \in [d]$ , create a node  $(a_1, x_i, x_j)$  in a set  $X$ . For every node  $a_3 \in A_3$  and every pair of coordinates  $i, j \in [d]$ , create a node  $(a_3, x_i, x_j)$  in a set  $Y$ .

See [Figure 1b](#). The edges are as follows.

For every  $i, j \in [d]$ , and every  $a_1 \in A_1, a_3 \in A_3$ , add an edge between  $(a_1, x_i, x_j)$  and  $(a_3, x_i, x_j)$ ; we get bicliques in  $X \times Y$  corresponding to each pair of coordinates  $i, j$ .

For each  $(a_1, a_2) \in S$ , we add an edge to  $(a_1, x_i, x_j) \in X$  if and only if  $a_1[i] = a_1[j] = 1$  and  $a_2[i] = 1$ . For each  $(a_2, a_3) \in T$ , we add an edge to  $(a_3, x_i, x_j) \in Y$  if and only if  $a_3[i] = a_3[j] = 1$  and  $a_2[j] = 1$ .

Suppose that there is no 3-OV solution. Then, for every  $a_1 \in A_1, a_2 \in A_2, a_3 \in A_3$ , there exists a coordinate  $k$  such that  $a_1[k] = a_2[k] = a_3[k] = 1$ . Consider an arbitrary  $(a_1, a_2) \in S$  and  $(a'_2, a_3) \in T$ . There is a coordinate  $i$  for which  $a_1[i] = a_2[i] = a_3[i] = 1$  and a coordinate  $j$  for which  $a_1[j] = a'_2[j] = a_3[j] = 1$ . By construction, there is an edge between  $(a_1, a_2) \in S$  and  $(a_1, x_i, x_j) \in X$  and between  $(a'_2, a_3) \in T$  and  $(a_3, x_i, x_j) \in Y$ . Together with the edge between  $(a_1, x_i, x_j)$  and  $(a_3, x_i, x_j)$ , we get that the distance between  $(a_1, a_2) \in S$  and  $(a'_2, a_3) \in T$  is 3. Thus the  $S$ - $T$ -Diameter is 3.

Suppose now that there is a 3-OV solution,  $a_1 \in A_1, a_2 \in A_2, a_3 \in A_3$ . Then one can show that if  $d((a_1, a_2), (a_2, a_3)) \leq 5$ , then there is a coordinate  $i$  such that  $a_1[i] = a_2[i] = a_3[i] = 1$ , giving a contradiction. Because the graph is bipartite, the distance must be  $\geq 7$ , and we can conclude.

Thus, the  $S$ - $T$  Diameter is 3 if there is no 3-OV solution or  $\geq 7$  if there is one. The number of vertices is  $O(n^2 + nd^2)$  and the number of edges is  $O(n^2d^2)$ . Let  $N = n^2d^2$ . If there is an  $O(N^{3/2-\varepsilon})$  time algorithm distinguishing 3 and 7 for  $\varepsilon > 0$ , then 3-OV can be solved in  $n^{3-2\varepsilon}\text{poly}(d)$  time.  $\square$

**Other known hardness results under SETH and  $k$ -OV.** In recent years, there has been an explosion of conditional hardness results based on OV and hence SETH:

1. Tight lower bounds for approximating the Graph Diameter and Graph Eccentricities [Roditty and Vassilevska Williams \[2013\]](#), [Chechik, Larkin, Roditty, Schoenebeck, Tarjan, and Vassilevska Williams \[2014\]](#), [Abboud, Vassilevska Williams, and Wang \[2016\]](#), and [Backurs, Roditty, Segal, Vassilevska Williams, and Wein \[2018\]](#).
2. Tight quadratic lower bounds for the Local Alignment problem [Abboud, Vassilevska Williams, and Weimann \[2014\]](#).
3. Tight lower bounds for dynamic problems. The first comprehensive paper to consider multiple hardness hypotheses to explain the difficulty of dynamic problems was by [Abboud and Vassilevska Williams \[2014\]](#). Under SETH, the main hardness results concern the following dynamic problems: maintaining under edge insertions and deletions, the strongly connected components of a graph, the number of nodes reachable from a fixed source, a 1.3 approximation of the graph diameter, or given fixed node sets  $S$  and  $T$ , whether there is a pair of nodes  $s \in S, t \in T$  so that  $s$  can reach  $t$ .
4. Strong hardness for the All Pairs Max Flow problem [Krauthgamer and Trabelsi \[2017\]](#): in  $n$  node,  $m$  edge graphs  $mn^{2-o(1)}$  time is needed. Lower bounds from OV and from Max-CNF-SAT. These results are based on previous hardness for variants of the Max Flow problem under SETH by Abboud et al. [Abboud, Vassilevska Williams, and Yu \[2015\]](#).
5. Lower bounds for incremental and decremental Max-Flow [Dahlgaard \[2016\]](#) following [Abboud, Vassilevska Williams, and Yu \[2015\]](#) and [Abboud and Vassilevska Williams \[2014\]](#). This is among the handful of lower bounds that address *amortized* runtimes for partially dynamic algorithms. The prior techniques could only provide worst case lower bounds here.
6. Lower bounds for sensitivity problems. Sensitivity problems are similar to dynamic problems in that they need to preprocess the input and prepare a data structure that answers queries after some sequence of updates. The difference is that once the queries are answered, the updates must be rolled back to the original state of the input. That is, the sensitivity problem is to prepare for any set of small changes and be able to answer queries on them. [Henzinger, Lincoln, Neumann, and Vassilevska Williams \[2017\]](#) give lower bounds under SETH for sensitivity data structures for graph problems such as answering for any small (constant) size set of edge insertions, approximate Graph Diameter queries or queries about the number of nodes reachable from a fixed source node.
7. Closest Pair in  $d$ -dimensional Hamming Space cannot be solved in  $n^{2-\varepsilon}2^{o(d)}$  time for  $\varepsilon > 0$  [Alman and R. Williams \[2015\]](#). The best algorithm for this problem and

- several others (e.g. offline bichromatic furthest neighbors) is by [Alman, Chan, and R. R. Williams \[2016\]](#) and runs in  $n^{2-1/O(c \log^2(c))}$  time for  $d = c \log n$ .
8. Quadratic lower bounds for LCS [Abboud, Backurs, and Vassilevska Williams \[2015b\]](#) and [Bringmann and Künnemann \[2015b\]](#), Edit Distance [Backurs and Indyk \[2015\]](#), Frechet Distance [Bringmann \[2014\]](#). [Abboud, Backurs, and Vassilevska Williams \[2015b\]](#) also give an  $n^{k-o(1)}$  lower bound for computing the LCS of  $k$  strings for any  $k \geq 2$ .
  9. Tight lower bounds for problems like LCS and RNA-Folding where the input strings are represented as a context free grammar whose only output is the input string [Abboud, Backurs, Bringmann, and Künnemann \[2017\]](#). Some of the lower bounds are also based on Hypotheses about the complexity of  $k$ -Clique and  $k$ -SUM.
  10. Subset Sum on  $n$  integers and target  $T$ , cannot be solved in  $T^{1-\varepsilon} 2^{o(n)}$  time for any  $\varepsilon > 0$  [Abboud, Bringmann, Hermelin, and Shabtay \[2017\]](#). Similar results apply to the Bicriteria Path problem.
  11. Tight lower bounds for the Subtree Isomorphism problem: given rooted trees on  $n$  total nodes  $T$  and  $T'$ , is  $T$  a subtree of  $T'$ ? [Abboud, Backurs, Hansen, Vassilevska Williams, and Zamir \[2016\]](#) show that truly subquadratic algorithms for the following refute the OV Hypothesis: for binary, rooted trees, or for rooted trees of depth  $O(\log \log n)$ . Conversely, for every constant  $d$ , there is a constant  $\varepsilon_d > 0$  and a randomized, truly subquadratic algorithm for degree- $d$  rooted trees of depth at most  $(1 + \varepsilon_d) \log_d n$ .
  12. Frechet distance on  $n$ -length strings requires  $n^{2-o(1)}$  time [Bringmann \[2014\]](#), and is hard to approximate [Bringmann and Künnemann \[2015a\]](#) and [Bringmann and Mulzer \[2016\]](#).
  13. Tight results for regular expression matching ([Backurs and Indyk \[2016\]](#) and [Bringmann, Grønlund, and Larsen \[2017\]](#)): here one is given a pattern of length  $m$ , text of length  $n$ , and the pattern involves concatenation, OR, Kleene star and Kleene plus. Under SETH, there is a dichotomy of problems (proven for depth 2 by [Backurs and Indyk \[2016\]](#) and for  $> 2$  by [Bringmann, Grønlund, and Larsen \[2017\]](#)): either they are solvable in near-linear time, or they require  $mn^{1-o(1)}$  time. There is a single exception: the Word Break problem solvable in  $\tilde{O}(m + nm^{1/3})$  time [Bringmann, Grønlund, and Larsen \[ibid.\]](#).
  14. Tight lower bounds for problems in model checking: for Büchi objectives [Chatterjee, Dvorák, Henzinger, and Loitzenbauer \[2016a\]](#) and others [Chatterjee, Dvorák, Henzinger, and Loitzenbauer \[2016b\]](#).
  15. Tight lower bounds for succinct stable matching [Moeller, Paturi, and Schneider \[2016\]](#).



16. Quadratic hardness results for problems in Machine Learning [Backurs, Indyk, and Schmidt \[2017\]](#).
17. Tight hardness for some one dimensional Dynamic Programming problems [Künne-mann, Paturi, and Schneider \[2017\]](#).
18. Furthest pair in  $R^d$  ( $\ell_2$ ) on  $n$  vectors, when  $d = \omega(\log \log n)$  requires  $n^{2-o(1)}$  time [R. Williams \[2018\]](#). This is to be contrasted with Closest Pair in the same dimensions which can be solved in  $n^{1+o(1)}$  time.
19. Very strong inapproximability several problems via the introduction of *Distributed PCPs* for Fine-Grained Hardness of Approximation [Abboud, Rubinstein, and R. R. Williams \[2017\]](#): Bichromatic Max-Inner Product on  $N$  vectors in  $\{0, 1\}^d$  cannot be approximated better than a factor of  $2^{(\log N)^{1-o(1)}}$  if you do not spend  $N^{2-o(1)}$  time. Similar inapproximability for approximation versions of Subset Query, Bichromatic LCS Closest Pair, Regular Expression Matching and Diameter in Product Metrics.

## 4 Hardness results from 3-SUM

A seminal paper by [Gajentaan and M. Overmars \[1995\]](#) and [Gajentaan and M. H. Overmars \[2012\]](#) from the 1990s introduces the 3-SUM Hypothesis and proves that a large set of problems in computational geometry require quadratic time, under this hypothesis:

1. Given a set of points in the plane, decide whether any three are colinear (Colinearity / 3 Points on Line).
2. Given a set of lines in the plane, decide whether any three of them pass through the same point (Point on 3 Lines).
3. Given a set of non-intersecting, axis-parallel line segments, decide whether some line separates them into two non-empty subsets (Separator).
4. Given a set of (infinite) strips in the plane and a rectangle, decide whether they fully cover the rectangle (Strips Cover Box).
5. Given a set of triangles in the plane, compute their measure (Triangle Measure).
6. Given a set of horizontal opaque triangles in three dimensional space, a view point  $p$  and another triangle  $T$ , decide whether there is a point on  $T$  that can be seen from  $p$  (Visible Triangle).
7. Given a set of non-intersecting, axis-parallel line segment obstacles in the plane, a rod and a source and a destination point, decide whether the rod can be moved by translations and rotations from the source to the destination without colliding with the obstacles (Planar Motion Planning).

8. Given a set of horizontal triangle obstacles in three dimensional space, a vertical rod, and a source and destination, decide whether the rod can be translated (without rotation) from the source to the destination without colliding with the obstacles (3D Motion Planning).

The notion of 3-SUM hardness reduction used in [Gajentaan and M. Overmars \[1995\]](#) and [Gajentaan and M. H. Overmars \[2012\]](#) is more restrictive than the fine-grained reduction defined later on. It only allows the creation of  $O(1)$  number of instances, each of no more than linear size. Even though the reduction notion is limited, it is still possible to obtain all of the above hardness results using more or less simple algebraic transformations. The paper inspired many other 3-SUM hardness results in computational geometry. Some of these include polygon containment [Barequet and Har-Peled \[2001\]](#), testing whether a dihedral rotation will cause a chain to self-intersect [Soss, Erickson, and M. H. Overmars \[2003\]](#) and many others [de Berg, de Groot, and M. H. Overmars \[1997\]](#), [Erickson \[1999\]](#), [Aronov and Har-Peled \[2008\]](#), [Cheong, Efrat, and Har-Peled \[2007\]](#), [Bose, van Kreveld, and Toussaint \[1998\]](#), [Erickson, Har-Peled, and Mount \[2006\]](#), [Arkin, Chiang, Held, Mitchell, Sacristán, Skiena, and Yang \[1998\]](#), [Archambault, Evans, and Kirkpatrick \[2005\]](#), and [Abellanas, Hurtado, Icking, Klein, Langetepe, Ma, Palop, and Sacristán \[2001\]](#).

A transformative paper in 3-SUM research by Pătraşcu [Pătraşcu \[2010\]](#) shows that 3-SUM is equivalent (under subquadratic reductions) to a slightly simpler looking problem, *3-SUM Convolution*: Given three length  $n$  arrays  $A$ ,  $B$  and  $C$  of integers, decide whether there exist  $i, k$  such that  $C[k] = A[i] + B[k - i]$ .

Unlike for 3-SUM,  $O(n^2)$  is the brute-force algorithm runtime for 3-SUM Convolution (for 3-SUM the trivial runtime is  $O(n^3)$ ). This makes it easier to reduce 3-SUM Convolution to other problems whose best known algorithm is the brute-force one. Also, because now the search is reduced to finding two indices  $i, k$ , as opposed to searching for a sum of two integers, one can use 3-SUM Convolution in reductions to problems that are more combinatorial in nature. Pătraşcu reduces 3-SUM Convolution to problems such as Listing Triangles in a graph. He shows that listing up to  $m$  triangles in an  $m$ -edge graph requires  $m^{4/3-o(1)}$  time under the 3-SUM Hypothesis. This is the first hardness result for a truly combinatorial problem (no numbers in the instance).

Prior to Pătraşcu's results, there is one other 3-SUM hardness result for a problem outside computational geometry, by [Vassilevska and R. Williams \[2009\]](#). They show that under the 3-SUM Hypothesis, the following *Exact Triangle* problem requires  $n^{2.5-o(1)}$  time on an  $n$  node edge-weighted graph  $G$ : determine whether there is a triangle  $G$  whose edge weights sum to 0. Pătraşcu's equivalence between 3-SUM and 3-SUM Convolution allows this hardness to be improved to  $n^{3-o(1)}$ , thus showing that the brute-force cubic algorithm for the problem might be optimal [Vassilevska Williams and R. Williams \[2013\]](#).

After [Pătraşcu \[2010\]](#) and [Vassilevska Williams and R. Williams \[2013\]](#), many other combinatorial problems were proven to be 3-SUM hard: [Abboud and Vassilevska Williams \[2014\]](#) continue Pătraşcu’s work, giving lower bounds for many dynamic problems under the 3-SUM hypothesis. Example graph problems of consideration are to maintain under edge deletions and insertions:  $s - t$  Reach (whether a given fixed source can reach a given fixed destination in a directed graph), SCC (the strongly connected components of a graph, or even just their number), BPMatch (whether a bipartite graph has a perfect matching), and many others. Kopelowitz et al. [Kopelowitz, Pettie, and Porat \[2016\]](#) improve Pătraşcu’s reduction to triangle listing and show that the known algorithms for listing triangles in graphs [Björklund, Pagh, Vassilevska Williams, and Zwick \[2014\]](#) are optimal if  $\omega = 2$  and under the 3-SUM Hypothesis. They also give amortized conditional lower bound for maintaining a maximum matching in a graph under edge insertions. [Abboud, Vassilevska Williams, and Weimann \[2014\]](#) show that the Local Alignment requires quadratic time under 3-SUM. The following other problems are also known to be hard under 3-SUM: jumbled indexing [Amir, Chan, M. Lewenstein, and N. Lewenstein \[2014\]](#), online pattern matching with gaps [Amir, Kopelowitz, Levy, Pettie, Porat, and Shalom \[2016\]](#), partial matrix multiplication, and witness reporting versions of convolution problems [Goldstein, Kopelowitz, M. Lewenstein, and Porat \[2016\]](#), and others.

## 5 Hardness results from APSP

APSP is now known to be equivalent to many other problems on  $n$  node graphs and  $n \times n$  matrices so that either all these problems admit  $O(n^{3-\varepsilon})$  time algorithms for  $\varepsilon > 0$ , or none of them do. A partial list of these equivalent problems is below. The main references are the original paper by [Vassilevska Williams and R. Williams \[2010, 2018\]](#) (bullets 1-9), and also [Backurs, Dikkala, and Tzamos \[2016\]](#) (bullet 9), [Abboud, Grandoni, and Vassilevska Williams \[2015\]](#) (bullets 10-12), and [Lincoln, Vassilevska Williams, and R. R. Williams \[2018\]](#) (bullet 13).

1. The all-pairs shortest paths problem on weighted digraphs (APSP).
2. Detecting if an edge-weighted graph has a triangle of negative total edge weight (Negative Triangle).
3. Listing up to  $n^{2.99}$  negative triangles in an edge-weighted graph (Triangle listing).
4. Finding a minimum weight cycle in a graph of non-negative edge weights (Shortest Cycle).
5. The replacement paths problem on weighted digraphs (RP).
6. Finding the second shortest simple path between two nodes in a weighted digraph (2nd Shortest Path).

7. Checking whether a given matrix defines a metric (Metricity).
8. Verifying a matrix product over the  $(\min, +)$ -semiring (Distance Product Verification).
9. Finding a maximum subarray in a given matrix (Max Subarray).
10. Finding the Median node of a weighted graph (Median).
11. Finding the Radius of a weighted graph (Radius).
12. Computing the Betweenness Centrality of a given node in a weighted graph (BC).
13. Computing the Wiener Index of a weighted graph (Wiener Index).

Some of the equivalences above have been strengthened to preserve sparsity [Agarwal and Ramachandran \[2016\]](#) and [Lincoln, Vassilevska Williams, and R. R. Williams \[2018\]](#) and even the range of weights [Roditty and Vassilevska Williams \[2011\]](#). Beyond the above equivalences, there have been multiple APSP-hardness results. Computing the edit distance between two rooted ordered trees with nodes labeled from a fixed alphabet (Tree Edit Distance) [Bringmann, Gawrychowski, Mozes, and Weimann \[2017\]](#) is known to require cubic time if APSP does. An equivalence with APSP is an open problem. [Abboud and Vassilevska Williams \[2014\]](#) provided tight hardness for dynamic problems under the APSP Hypothesis. The main results are for Bipartite Maximum Weight Matching and  $s$ - $t$  Shortest Path, showing that the trivial dynamic algorithms are optimal, unless APSP can be solved faster. For instance, any algorithm that can maintain the distance in a weighted graph between a fixed source node  $s$  and a fixed target  $t$ , while supporting edge deletions, must either perform  $n^{3-o(1)}$  time preprocessing, or either the update or the query time must be  $n^{2-o(1)}$  ([Abboud and Vassilevska Williams \[ibid.\]](#) following [Roditty and Zwick \[2004\]](#)). [Henzinger et al. \[2017\]](#) give tight lower bounds under the APSP Hypothesis for *sensitivity* problems such as answering Graph Diameter or  $s$ - $t$  Shortest Path queries for any single edge failure. [Abboud and Dahlgaard \[2016\]](#) gave the first fine-grained lower bound for a problem in *planar* graphs: no algorithm for dynamic shortest paths or maximum weight bipartite matching in planar graphs can support both updates and queries in amortized  $O(n^{1/2-\varepsilon})$  time, for any  $\varepsilon > 0$ , unless the APSP Hypothesis fails.

The main technical hurdle in showing the equivalences and most hardness results above, overcome by [Vassilevska Williams and R. Williams \[2010\]](#), is in reducing APSP to the Negative Triangle Problem. Negative Triangle is a simple decision problem, and reducing it to the other problems above is doable, with sufficient gadgetry.

Below we will outline the reduction *from APSP to Negative Triangle*. It is a true fine-grained reduction — it produces many instances, reducing a function problem to a decision problem.

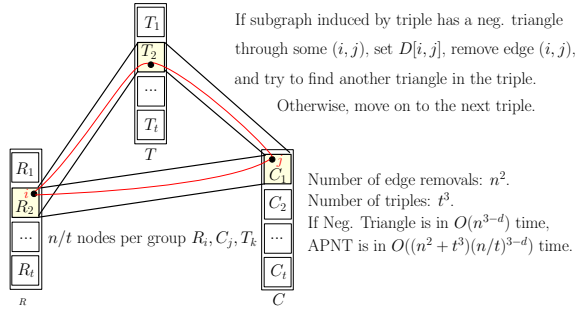


Figure 2

The first step is to formulate APSP as a problem involving triangles, the *All-Pairs Negative Triangles (APNT)* problem defined as follows: given a tripartite graph  $G$  with node partitions  $R, T, C$ , with arbitrary edges in  $R \times T, T \times C, C \times R$ , with integer edge weights  $w(\cdot)$ , for every  $i \in R, j \in C$ , determine whether there exists a  $t \in T$  so that  $w(i, t) + w(t, j) + w(j, i) < 0$ .

Reducing APSP to APNT (see [Vassilevska Williams and R. Williams \[2010\]](#)) is done by using a known equivalence [Fischer and Meyer \[1971\]](#) between APSP and the Distance Product problem of computing a product of two matrices over the  $(\min, +)$  semiring. Then Distance Product is solved by using calls to APNT to binary search for the entries in the output matrix.

Now, it suffices to reduce the All-Pairs Negative Triangles problem to just detecting a Negative Triangle. The reduction picks a parameter  $t = n^{2/3}$ . Then, it arbitrarily partitions  $R, T, C$  into  $t$  pieces each of size roughly  $n/t$ :  $(R_1, \dots, R_t), (T_1, \dots, T_t), (C_1, \dots, C_t)$ . Every negative triangle is in some triple  $(R_i, T_j, C_k)$ . We will create Negative Triangle instances for these triples as follows. See [Figure 2](#).

Create an  $n \times n$  all 0 matrix  $D$  that at the end will have  $D[i, j] = 1$  if and only if there is some  $\ell \in T$  so that  $i, \ell, j$  is a negative triangle.

Now, for every triple of parts  $(R_i, T_j, C_k)$  in turn, starting with  $(R_1, T_1, C_1)$ , while the subgraph induced by the triple  $(R_i, T_j, C_k)$  contains a negative triangle (this is a call to Negative Triangle), find one such negative triangle  $i \in R, \ell \in T, j \in C$  via self-reduction (see [Vassilevska Williams and R. Williams \[2010\]](#)). Set  $D[i, j] = 1$  and remove  $(i, j)$  from the entire graph; we do this so that there will be no more negative triangles containing  $(i, j)$  in any of the subsequent Negative Triangle calls. Thus, the number of calls that do find a negative triangle is  $\leq n^2$ .

The algorithm keeps calling Negative Triangle on a triple until the triple has no more negative triangles, and then moves on to the next triple. At the end it just returns  $D$ . The

number of calls to Negative Triangle that do not return a negative triangle are bounded by the number of triples which is  $t^3$ .

The number of calls to Negative Triangle is at most  $t^3 + n^2$ , and each call is on a graph with  $O(n/t)$  nodes. As  $t = n^{2/3}$ , we get  $O(n^2)$  calls to instances of size  $O(n^{1/3})$ , i.e. a subcubic fine-grained reduction.

## 6 Other hypotheses

Beyond the three main hardness hypotheses, there are several other ones that have come to the forefront of fine-grained complexity. Again, the model of computation is the word-RAM with  $O(\log n)$  bit words.

**Hitting set.** The *Hitting Set (HS)* problem is as follows: given two sets of vectors  $S, T \subseteq \{0, 1\}^d$  for  $d = \omega(\log n)$ , determine whether there is some  $s \in S$  such that  $s \cdot t \neq 0$  for all  $t \in T$ , in other words a vector in  $S$  that hits all vectors in  $T$ .

**Hypothesis 5** (HS Hypothesis [Abboud, Vassilevska Williams, and Wang \[2016\]](#)). *No randomized algorithm can solve HS on  $n$  vectors in  $\{0, 1\}^d$  in  $n^{2-\varepsilon} \text{poly}(d)$  time for  $\varepsilon > 0$ .*

To see why this new hypothesis is useful, consider the converse HS: decide whether  $\forall s \in S \exists t \in T$  such that  $s \cdot t = 0$ . This is OV with the first  $\exists$  quantifier replaced with  $\forall$ . This quantifier flip allows for different hardness results to be proven. For instance the Radius Problem asks, given a graph, whether there exists a vertex  $c$  such that for all other vertices  $v$ ,  $d(v, c)$  is most  $R$ . The converse asks whether  $\forall c \exists v$  such that  $d(v, c) > R$  which is the  $\forall \exists$  variant of Diameter, which is  $\exists \exists$ . While it was not hard to reduce the  $\exists \exists$  problem OV to Diameter, reducing it to the  $\exists \forall$  Radius seemed problematic. On the other hand, since HS has the  $\forall \exists$  structure, [Abboud, Vassilevska Williams, and Wang \[ibid.\]](#) are able to reduce it Radius, so that Radius on sparse graphs requires  $n^{2-o(1)}$  time under the HS Hypothesis. HS and its hypothesis are also studied in [Gao, Impagliazzo, Kolokolova, and R. R. Williams \[2017\]](#).

The HS Hypothesis implies the OV Hypothesis [Abboud, Vassilevska Williams, and Wang \[2016\]](#) and [Abboud, R. R. Williams, and Yu \[2015\]](#), but the reverse is not known to be true.

**Hypotheses on the complexity of  $k$ -Clique.** For a constant  $k \geq 3$ , the  $k$ -Clique problem is as follows: given a graph  $G = (V, E)$  on  $n$  vertices, does  $G$  contain  $k$  distinct vertices  $a_1, \dots, a_k$  so that for every  $i, j$ ,  $i \neq j$ ,  $(a_i, a_j) \in E$ ? Such a  $k$  node graph is called a  $k$ -clique.

The  $k$ -Clique problem can easily be solved in  $O(n^k)$  time by enumerating all  $k$ -tuples of vertices. A faster algorithm [Itai and Rodeh \[1978\]](#) and [Nešetřil and Poljak \[1985\]](#) reduces the problem to multiplying square matrices, giving an  $O(n^{\omega k/3}) \leq O(n^{0.8k})$  time algorithm when  $k$  is divisible by 3. Recall,  $\omega < 2.373$  is the exponent of square matrix multiplication [Vassilevska Williams \[2012\]](#), [Stothers \[2010\]](#), and [Le Gall \[2014\]](#). If  $k$  is not divisible by 3, the fastest known algorithm for  $k$ -Clique runs asymptotically in the time to multiply an  $n^{\lfloor k/3 \rfloor} \times n^{\lceil k/3 \rceil}$  matrix by an  $n^{\lceil k/3 \rceil} \times n^{k - \lfloor k/3 \rfloor - \lceil k/3 \rceil}$  matrix, which is no more than  $O(n^{2+\omega k/3})$ ; tighter bounds are known ([Le Gall \[2012\]](#), [Coppersmith \[1997\]](#), [Gall and Urrutia \[2018\]](#), and [Huang and Pan \[1998\]](#)).

**Hypothesis 6** ( $k$ -Clique Hypothesis). *No randomized algorithm can detect a  $k$ -Clique in an  $n$  node graph in  $O(n^{\frac{\omega k}{3} - \varepsilon})$  time for  $\varepsilon > 0$ .*

The Hypothesis is usually used for  $k$  divisible by 3. Also, since  $\omega \geq 2$  (one needs to output a matrix with  $n^2$  entries), the hypothesis asserts in particular that  $k$ -Clique requires  $n^{2k/3 - o(1)}$  time.

Two harder problems are the *Min-Weight  $k$ -Clique* and *Exact  $k$ -Clique* problems. In both problems, one is given a graph on  $n$  vertices and edge weights in  $\{-n^{100k}, \dots, n^{100k}\}$ . In the first, one seeks a  $k$ -Clique that minimizes the total sum of its edge weights. In the second, one seeks a  $k$ -Clique with weight sum of exactly 0. Neither of these problems are known to be solvable in  $O(n^{k-\varepsilon})$  time for any constant  $\varepsilon > 0$ .

**Hypothesis 7** (Min-Weight  $k$ -Clique Hypothesis). *The Min-Weight  $k$ -Clique problem on  $n$  node graphs with edge weights in  $\{-n^{100k}, \dots, n^{100k}\}$  requires (randomized)  $n^{k-o(1)}$  time.*

**Hypothesis 8** (Exact  $k$ -Clique Hypothesis). *The Exact  $k$ -Clique problem on  $n$  node graphs with edge weights in  $\{-n^{100k}, \dots, n^{100k}\}$  requires (randomized)  $n^{k-o(1)}$  time.*

It is known that the Min-Weight  $k$ -Clique Hypothesis implies the Exact  $k$ -Clique Hypothesis [Vassilevska and R. Williams \[2009\]](#). The version of Min-Weight  $k$ -Clique in which the weights are on the nodes, rather than on the edges, can be solved in the same time as the (unweighted)  $k$ -Clique problem [Czumaj and Lingas \[2007\]](#), [Vassilevska and R. Williams \[2009\]](#), and [Abboud, Lewi, and R. Williams \[2014\]](#), so that the Min-Weight  $k$ -Clique Hypothesis does not hold for node-weighted graphs.

Using results from [Abboud, Lewi, and R. Williams \[2014\]](#) and [R. Williams \[2005\]](#) and the known reduction from  $k$ -Clique to  $k$ -Dominating Set, one can show that Exact-Weight and Min-Weight  $k$ -Clique are  $(n^k, n^2)$ -reducible to 2-OV, and hence their hypotheses imply the OV Hypothesis [Abboud, Bringmann, Dell, and Nederlof \[2018\]](#).

Notably, the Min-Weight 3-Clique problem is equivalent to the Negative Triangle problem and hence also to APSP, under subcubic fine-grained reductions [Vassilevska Williams](#)

and R. Williams [2010]. Exact 3-Clique is just the Exact Triangle problem studied by Vassilevska Williams and R. Williams [2013]. Exact 3-Clique seems genuinely more difficult than Min-Weight 3-Clique. First, the latter problem can be solved in  $n^3 / \exp(\sqrt{\log n})$  time R. Williams [2014], whereas the fastest algorithm for Exact 3-Clique runs in  $n^3 (\log \log n)^2 / \log n$  time Jørgensen and Pettie [2014]. Second, as we mentioned in the section on 3-SUM, Exact 3-Clique requires  $n^{3-o(1)}$  under both the 3-SUM and the APSP Hypotheses, whereas Min-Weight 3-Clique is equivalent to APSP which is not known to be related to 3-SUM and could be potentially easier.

The following tight lower bounds under the  $k$ -Clique Hypothesis are known: Context Free Grammar Recognition for  $O(1)$  size grammars, RNA-Folding, and Language Edit Distance require  $n^{\omega-o(1)}$  time Abboud, Backurs, and Vassilevska Williams [2015a] and Chang [2016], and Tree-adjoining grammar parsing Bringmann and Wellnitz [2017] requires the unusual running time of  $n^{2\omega}$ , tight due to Rajasekaran and Yooseph [1998].

The following problems have tight conditional lower bounds under the Min-Weight  $k$ -Clique Hypothesis: all problems hard under the APSP Hypothesis, the Local Alignment Problem Abboud, Vassilevska Williams, and Weimann [2014], the Viterbi problem of finding the most likely path in a Hidden Markov Model (HMM) that results in a given sequence of observations Backurs and Tzamos [2017], the Maximum Weight Box problem that given weighted points (positive or negative) in  $d$  dimensions, asks to find the axis-aligned box which maximizes the total weight of the points it contains Backurs, Dikkala, and Tzamos [2016].

Recently, the  $k$ -Clique and Min-Weight  $k$ -Clique Hypotheses have been used to show hardness for graph problems for almost all sparsities. Recall that under SETH one could show that many problems in very sparse graphs (with a near-linear number of edges) are hard. On the other hand, the APSP Hypothesis implied hardness for problems in dense graphs, i.e. when the runtime is measured solely in terms of the number of vertices. However, neither of these hypotheses seem to address questions such as “Can APSP be solved in  $O(n^2 + m^{3/2})$  time?”. Such a runtime would be consistent with the APSP Hypothesis and with the fact that in sparse graphs one needs  $\Omega(n^2)$  time to write down the output.

For APSP and many other graph problems on  $m$  edges and  $n$  vertices, the best known running times are of the form  $\tilde{O}(mn)$ : APSP, Shortest Cycle, Replacement Paths, Radius, Wiener Index etc. There is no faster algorithm for any sparsity  $m$ . Lincoln et al. Lincoln, Vassilevska Williams, and R. R. Williams [2018] address this by showing that for any constant  $k \geq 1$ , if one assumes the Min-Weight  $2k + 1$ -Clique Hypothesis, then APSP, Shortest Cycle, Replacement Paths, Radius, Wiener Index etc. require  $mn^{1-o(1)}$  time in weighted graphs with  $m = \Theta(n^{1+1/k})$  edges. In other words, for an infinite number of sparsities,  $mn$  is the right answer. Under the  $k$ -Clique Hypothesis, Lincoln, Vassilevska Williams, and R. R. Williams [ibid.] provide weaker lower bounds for the same problems in unweighted graphs.



**Boolean Matrix Multiplication (BMM)..** The BMM problem is, given two  $n \times n$  matrices  $A$  and  $B$ , to compute the  $n \times n$  matrix  $C$  with  $C[i, j] = \bigvee_{k=1}^n (A[i, k] \wedge B[k, j])$  for all  $i, j$ . BMM can be solved using the known matrix multiplication algorithms over a field by embedding the Boolean semiring into the Rationals. Thus BMM on  $n \times n$  matrices is in  $O(n^{2.373})$  time [Vassilevska Williams \[2012\]](#) and [Le Gall \[2014\]](#). However, the theoretically fast algorithms for matrix multiplication are considered inefficient. The desire for more practical algorithms motivates the notion of “combinatorial” algorithms. This notion is not well-defined, however it roughly means that the runtime should have a small constant in the big-O, and that the algorithm is feasibly implementable.

There is a “*BMM hypothesis*” (in quotes as this is not well-defined) asserting that any combinatorial BMM algorithm requires  $n^{3-o(1)}$  time. This is supported by the lack of truly subcubic combinatorial BMM algorithms: the fastest is by [Yu \[2015\]](#) and runs in  $n^3(\log \log n)^{O(1)} / \log^4 n$  time. The first combinatorial BMM algorithm is the so called Four-Russians algorithm [Arlazarov, Dinic, Kronrod, and Faradzev \[1970\]](#), which was later improved by [Chan \[2015\]](#), [Bansal and R. Williams \[2009\]](#), and [Yu \[2015\]](#).

The BMM Hypothesis has been used to explain the lack of fast combinatorial algorithms for many problems: many dynamic problems [Abboud and Vassilevska Williams \[2014\]](#) and [Roditty and Zwick \[2004\]](#), Context Free Grammar Parsing [L. Lee \[2002\]](#),  $2k$ -Cycle in undirected graphs [Dahlgard, Knudsen, and Stöckel \[2017\]](#), etc. Also many fine-grained combinatorial equivalences to BMM are known (e.g. [Vassilevska Williams and R. Williams \[2010\]](#)).

**Online Matrix Vector Multiplication (OMV)..** The BMM Hypothesis is unsatisfactory due to the undefined combinatorial notion, and there has been some work to replace it with something else. [Henzinger et al. \[2015\]](#) define the Online Matrix Vector (OMV) hypothesis which makes the BMM hypothesis about an online version of the problem for which even non-combinatorial subcubic algorithms seem out of reach. The OMV problem is well-studied [R. Williams \[2007a\]](#), [Blleloch, Vassilevska, and R. Williams \[2008\]](#), and [Larsen and R. Williams \[2017\]](#): given an  $n \times n$  Boolean matrix, preprocess it so that future products with arbitrary query  $n \times 1$  vectors are efficient.

**Hypothesis 9 (OMV Hypothesis).** *Every (randomized) algorithm that can process a given  $n \times n$  Boolean matrix  $A$ , and then in an online way can compute the products  $Av_i$  for any  $n$  vectors  $v_1, \dots, v_n$ , must take total time  $n^{3-o(1)}$ .*

The best algorithm for OMV is by [Larsen and R. Williams \[2017\]](#) who show that the OMV problem (for  $n$  queries) can be solved in total time  $n^3 / \exp(\sqrt{\log n})$  via a reduction to the OV problem. Moreover, [Larsen and R. Williams \[ibid.\]](#) give a cell probe

algorithm that can solve the problem using  $O(n^{11/4}/\sqrt{\log n})$  probes, thus ruling out an unconditional lower bound for OMV using purely information theoretic techniques.

The OMV Hypothesis is particular suited to proving conditional lower bounds for dynamic problems. Such lower bounds are known for practically all dynamic problems for which there is a known BMM-based combinatorial lower bound [Henzinger, Krinninger, Nanongkai, and Saranurak \[2015\]](#), and many other problems (e.g. [Dahlgaard \[2016\]](#)).

**Nondeterministic Strong Exponential Time Hypothesis (NSETH).** Surprisingly, the fastest algorithm for CNF-SAT on formulas on  $n$  variables, even using *nondeterminism*, still runs in roughly  $2^n$  time. This motivated Carmosino et al. [Carmosino, Gao, Impagliazzo, Mihajlin, Paturi, and Schneider \[2016\]](#) to define the following.

**Hypothesis 10** (Nondeterministic Strong Exponential Time Hypothesis (NSETH)). *Refuting unsatisfiable  $k$ -CNF formulas on  $n$  variables requires nondeterministic  $2^{n-o(n)}$  time for unbounded  $k$ .*

It is worth noting that NSETH does not allow randomization. In early work, Carmosino et al. also proposed a Merlin-Arthur and Arthur-Merlin SETH that assert that no constant round probabilistic proof system can refute unsatisfiable  $k$ -CNF formulas in  $2^{n-\Omega(n)}$  time. Williams [R. R. Williams \[2016\]](#) shows that these hypotheses are false in a very strong way, exhibiting proof systems that prove that the number of satisfying assignments of any given  $o(n)$ -depth, bounded-fan-in circuit is a given value, using a proof of length  $2^{n/2}\text{poly}(n)$  that can be verified in  $2^{n/2}\text{poly}(n)$  time with high probability, using only  $O(n)$  random bits.

The fact that the AM and MA versions of NSETH are false casts doubt on the veracity of NSETH. Nevertheless, disproving NSETH seems challenging. Assuming NSETH, [Carmosino, Gao, Impagliazzo, Mihajlin, Paturi, and Schneider \[2016\]](#) prove that there can be no deterministic reduction from OV to 3-SUM or APSP. This is done by exhibiting fast nondeterministic algorithms for the latter two problems, whereas OV cannot have a non-trivial nondeterministic refutation algorithm, under NSETH, via Williams' [R. Williams \[2005\]](#) reduction from CNF-SAT to OV that we presented earlier.

**SETH for other Circuit Satisfiability Problems.** As we mentioned in the introduction, CNF places a restriction on the input of the more general SAT problem. When represented as a circuit, a  $k$ -CNF formula has depth two — it is an AND of ORs. Moreover, due to the Sparsification Lemma of [Impagliazzo and Paturi \[2001\]](#), SETH really concerns the satisfiability problem for depth two circuits of  $O(n)$  size. To make SETH more believable, we can instead consider the satisfiability of less restricted classes of inputs to Circuit SAT.

Consider a Boolean function  $f$  on  $n$  bit inputs for which we want to prove satisfiability. It is not hard to see that any algorithm whose only access to  $f$  is by querying the value of  $f$

on various inputs, must spend  $\Omega(2^n)$  time to check if there is an  $n$ -bit  $x$  for which  $f(x) = 1$ . A clever algorithm would do more than query the function. It would attempt to analyze  $f$  to decrease the runtime of SAT. How much power algorithms have to analyze  $f$  depends crucially on the representation of  $f$ . It is impossible for a black box representation, and it is quite trivial if  $f$  is given as a DNF formula (ORs of ANDs of literals).

For each class  $C$  of representations, we can define the corresponding  $C$ -SETH that states that SAT with a representation from  $C$  cannot be solved in  $O(2^{(1-\varepsilon)n})$  time for  $\varepsilon > 0$ .

$C$ -SETH for  $k$ -CNF Formulas as  $k$  grows is just SETH. NC-SETH on the other hand asserts that SAT of polynomial size, polylogarithmic depth circuits requires  $2^{n-o(n)}$  time. NC circuits are much more powerful than CNF Formulas. They can perform most linear algebraic operations, and they can implement cryptographic primitives like One Way Functions and Pseudorandom Generators, for which the ability to hide satisfiability is crucial.

$C$ -SETH was defined by Abboud et al. [Abboud, Hansen, Vassilevska Williams, and R. Williams \[2016\]](#) who gave fine-grained lower bounds for sequence alignment problems such as Edit Distance, Frechet Distance, LCS. For instance, these problems on  $n$  length sequences require  $n^{2-o(1)}$  time unless NC-SETH fails, thus replacing the prior SETH hardness results with NC-SETH hardness. The results of [Abboud, Hansen, Vassilevska Williams, and R. Williams \[ibid.\]](#) also imply that a truly subquadratic algorithm for any of these problems would imply novel circuit lower bounds for classes such as  $E^{NP}$ . More surprisingly, if these problems can be solved in  $O(n^2/\log^c n)$  time for all constants  $c$ , then  $NTIME[2^{O(n)}]$  does not have non-uniform polynomial-size log-depth circuits. Hence, shaving all polylogs over the textbook quadratic runtime would result in a major advance in complexity theory.

**Two Problems Harder than CNF-SAT, 3-SUM and APSP..** The search for more believable hypotheses than SETH, and the APSP and 3-SUM Hypotheses motivates the following more believable conjecture: “At least one of SETH, the APSP Hypothesis and the 3-SUM Hypothesis is true.”

To prove hardness under this conjecture, one would have to perform three reductions (from  $k$ -SAT, APSP and from 3-SUM) instead of just one; this can be cumbersome. It is also not apriori clear that any natural problems are hard under all three conjectures. Abboud et al. [Abboud, Vassilevska Williams, and Yu \[2015\]](#) define two simple combinatorial problems and reduce  $k$ -SAT, APSP and 3-SUM to them. The goal is then to use these as the basis of hardness.

The first problem is *Triangle Collection*: Given a graph  $G = (V, E)$  with node colors  $c : V \rightarrow \{1, \dots, n\}$ , decide whether there exist colors  $c_1, c_2, c_3 \in \{1, \dots, n\}$  such that

there are NO triangles  $u, v, w$  in  $G$  (i.e.  $(u, v), (v, w), (w, u) \in E$ ), such that  $c(u) = c_1, c(v) = c_2, c(w) = c_3$ .

The second problem is *Matching Triangles*: Given a graph  $G = (V, E)$  with node colors  $c : V \rightarrow \{1, \dots, n\}$  and an integer  $\Delta$ , decide whether there exist colors  $c_1, c_2, c_3 \in \{1, \dots, n\}$  such that there are at least  $\Delta$  triangles  $u, v, w$  in  $G$  (i.e.  $(u, v), (v, w), (w, u) \in E$ ), such that  $c(u) = c_1, c(v) = c_2, c(w) = c_3$ .

Abboud et al. [Abboud, Vassilevska Williams, and Yu \[ibid.\]](#) show that if either Triangle Collection or Matching Triangles on  $n$  node graphs admit an  $O(n^{3-\varepsilon})$  time algorithm for any  $\varepsilon > 0$ , then all three of SETH, the APSP Hypothesis and the 3-SUM Hypothesis are false. In fact, this is true for a restricted version Triangle Collection\* of Triangle Collection that is easier to work with, and [Abboud, Vassilevska Williams, and Yu \[ibid.\]](#) give conditional hardness under it for several dynamic graph problems under edge insertions and deletions such dynamic Max Flow, or maintaining the number of nodes reachable from a fixed source node. [Dahlgaard \[2016\]](#) gave conditional lower bounds for approximating the graph Diameter both statically and dynamically, under the Triangle Collection\* hypothesis. Hence many problems are known to be difficult under all three main hypotheses.

## 7 Further Applications of Fine-grained Complexity

The fine-grained approach has found applications in many other areas of TCS:

- **FPT in P.** Parameterized complexity strives to classify problems according to their time complexity as a function of multiple parameters of the input or output. This is a different way to classify problems on a finer scale. It is particularly interesting for NP-hard problems.

A problem is FPT with respect to a set of parameters if it can be solved in time  $f(k_1, \dots, k_t)\text{poly}(n)$  on inputs of size  $n$  and parameters set to  $k_1, \dots, k_t$ ; here  $f$  can be any computable function. FPT problems can be solved in polynomial time when the parameters are constant; this can often make NP-hard FPT problems tractable. Parameterized complexity has identified many problems that are FPT and has developed a theory to explain which problems are likely not to be FPT (see e.g. [Flum and Grohe \[2006\]](#) and [Cygan, Fomin, Kowalik, Lokshtanov, Marx, Pilipczuk, Pilipczuk, and Saurabh \[2015\]](#)).

Abboud et al. [Abboud, Vassilevska Williams, and Wang \[2016\]](#) consider a notion of FPT for polynomial time problems: Fixed Parameter Subquadratic (FPS)—parameterized problems that admit algorithms running in time  $f(k)n^{2-\varepsilon}$  for  $\varepsilon > 0$  on inputs of size  $n$  and parameter(s) set to  $k$ , for some computable function  $f$ . [Abboud,](#)

Vassilevska Williams, and Wang [ibid.] show that the Diameter and Radius problems in graphs with parameter treewidth are FPS. They also give conditional lower bounds on the function  $f$  for these problems. Their work was continued by Fomin et al. Fomin, Lokshtanov, Pilipczuk, Saurabh, and Wrochna [2017] who added fixed parameter results for other polynomial time problems. Notice that parameterized algorithms have long been used within Algorithms: e.g. for graph problems, runtimes are often measured in terms of both the number of edges and the number of vertices. Abboud et al. Abboud, Vassilevska Williams, and Wang [2016] are the first to give fine-grained conditional lower bounds for parameterized polynomial time solvable problems.

- **Unconditional CONGEST Lower Bounds.** Abboud et al. Abboud, Censor-Hillel, and Khoury [2016] consider the CONGEST model in distributed computing in which processors are  $n$  nodes in a graph, and computation proceeds in rounds in which every processor can send  $O(\log n)$  bits of information to all adjacent processors. Abboud, Censor-Hillel, and Khoury [ibid.] (see also Bringmann and Krinninger [2017]) show how to convert some conditional lower bounds based on the OV Hypothesis to *unconditional* lower bounds in the CONGEST model. For instance, they show that in the CONGEST model, any algorithm that can compute a  $3/2 - \varepsilon$  approximation to the diameter of the graph of a  $5/3 - \varepsilon$  approximation to the eccentricities for any  $\varepsilon > 0$  needs  $\Omega(n)$  rounds of communication.

The basic idea in Abboud, Censor-Hillel, and Khoury [2016] is that OV is equivalent to Set Disjointness which has an unconditional  $\Theta(n)$  lower bound in communication complexity. The proofs show that any Diameter or Eccentricities protocol that takes too few rounds is solving Set Disjointness with too little communication.

- **Fine-Grained Cryptography.** Two papers begin the study of creating cryptographic primitives from fine-grained assumptions. Degwekar et al. Degwekar, Vaikuntanathan, and Vasudevan [2016] develop cryptographic protocols secure against adversaries that are at most as powerful as low circuit classes within P such as  $NC_1$  — this is more fine-grained but does not address runtime. More recently, Ball, Rosen, Sabin, and Vasudevan [2017], provide several problems that are provably hard on average, under SETH or the 3-SUM or APSP Hypotheses. Then they use these problems to construct a Proof of Work scheme. They leave as an open problem to develop more cryptographic primitives, such as One Way Functions, from fine-grained assumptions.
- **Fine-Grained Time/Space Tradeoffs for Algorithms.** Besides considering the runtime as the main measure of complexity, one can also consider the space usage. Lincoln et al. Lincoln, Vassilevska Williams, Wang, and R. R. Williams [2016] study the time/space tradeoffs of 3-SUM, building on prior work by Wang [2014].

Besides developing new algorithms, [Lincoln, Vassilevska Williams, Wang, and R. R. Williams \[2016\]](#) show that the 3-SUM hypothesis is equivalent to the following hypothesis: *There is some  $\delta > 0$ , such that every algorithm that uses  $O(n^{0.5+\delta})$  space, needs  $n^{2-o(1)}$  time to solve 3-SUM.*

This makes the 3-SUM Hypothesis look even more plausible as it only applies to space bounded algorithms. Also, one might conceivably be able to prove it unconditionally: restricting the space usage has been sufficient to prove unconditional lower bounds for SAT, among other problems [R. R. Williams \[2008\]](#).

- **Fine-Grained Time/Space Tradeoffs for Data Structures.** Goldstein et al. [Goldstein, Kopelowitz, M. Lewenstein, and Porat \[2017\]](#) define various data structure variants of 3-SUM, BMM and Directed Reachability, formulate novel conjectures and show consequences for the time/space tradeoffs for various data structure problems.
- **Fine-Grained Complexity in the I/O Model.** Demaine et al. [E. D. Demaine, Lincoln, Liu, Lynch, and Vassilevska Williams \[2018\]](#) initiate the study of the I/O model from the perspective of fine-grained complexity. The paper proposes plausible I/O hardness hypotheses, and uses these, together with fine-grained I/O reductions, to show that many known I/O upper bounds are tight. For instance, the best known upper bound on the I/O complexity of LCS is tight under one of the assumptions. Finally, they prove an analogue of the Time Hierarchy Theorem in the I/O model.

Fine-grained complexity is a growing field and we hope that its ideas will spread to many other parts of TCS and beyond.

**Acknowledgments.** The author would like to thank Ryan Williams for useful comments and Erik Demaine and Timothy M. Chan for pointers for references on 3-SUM hard problems.

## References

Amir Abboud, Arturs Backurs, Karl Bringmann, and Marvin Künnemann (2017). “Fine-Grained Complexity of Analyzing Compressed Data: Quantifying Improvements over Decompress-and-Solve”. In: *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pp. 192–203 (cit. on p. 3481).

- Amir Abboud, Arturs Backurs, Thomas Dueholm Hansen, Virginia Vassilevska Williams, and Or Zamir (2016). “Subtree Isomorphism Revisited”. In: *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pp. 1256–1271 (cit. on p. 3481).
- Amir Abboud, Arturs Backurs, and Virginia Vassilevska Williams (2015a). “If the Current Clique Algorithms are Optimal, So is Valiant’s Parser”. In: *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pp. 98–117 (cit. on p. 3489).
- (2015b). “Tight Hardness Results for LCS and Other Sequence Similarity Measures”. In: *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pp. 59–78 (cit. on p. 3481).
- Amir Abboud, Karl Bringmann, Holger Dell, and Jesper Nederlof (2018). “More Consequences of Falsifying SETH and the Orthogonal Vectors Conjecture”. In: *Proceedings of the 50th ACM Symposium on Theory of Computing, STOC 2018, Los Angeles, California, USA, 25-29 June 2018*, to appear (cit. on p. 3488).
- Amir Abboud, Karl Bringmann, Danny Hermelin, and Dvir Shabtay (2017). “SETH-Based Lower Bounds for Subset Sum and Bicriteria Path”. *CoRR* abs/1704.04546 (cit. on p. 3481).
- Amir Abboud, Keren Censor-Hillel, and Seri Khoury (2016). “Near-Linear Lower Bounds for Distributed Distance Computations, Even in Sparse Networks”. In: *Distributed Computing - 30th International Symposium, DISC 2016, Paris, France, September 27-29, 2016. Proceedings*, pp. 29–42 (cit. on p. 3494).
- Amir Abboud and Søren Dahlgaard (2016). “Popular Conjectures as a Barrier for Dynamic Planar Graph Algorithms”. In: *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pp. 477–486 (cit. on p. 3485).
- Amir Abboud, Fabrizio Grandoni, and Virginia Vassilevska Williams (2015). “Subcubic Equivalences Between Graph Centrality Problems, APSP and Diameter”. In: *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pp. 1681–1697 (cit. on p. 3484).
- Amir Abboud, Thomas Dueholm Hansen, Virginia Vassilevska Williams, and Ryan Williams (2016). “Simulating branching programs with edit distance and friends: or: a polylog shaved is a lower bound made”. In: *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pp. 375–388 (cit. on p. 3492).
- Amir Abboud, Kevin Lewi, and Ryan Williams (2014). “Losing Weight by Gaining Edges”. In: *Algorithms - ESA 2014 - 22th Annual European Symposium, Wroclaw, Poland, September 8-10, 2014. Proceedings*, pp. 1–12 (cit. on p. 3488).



- Amir Abboud, Aviad Rubinfeld, and R. Ryan Williams (2017). “Distributed PCP Theorems for Hardness of Approximation in P”. In: *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pp. 25–36 (cit. on p. 3482).
- Amir Abboud and Virginia Vassilevska Williams (2014). “Popular Conjectures Imply Strong Lower Bounds for Dynamic Problems”. In: *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pp. 434–443 (cit. on pp. 3480, 3484, 3485, 3490).
- Amir Abboud, Virginia Vassilevska Williams, and Joshua R. Wang (2016). “Approximation and Fixed Parameter Subquadratic Algorithms for Radius and Diameter in Sparse Graphs”. In: *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pp. 377–391 (cit. on pp. 3480, 3487, 3493, 3494).
- Amir Abboud, Virginia Vassilevska Williams, and Oren Weimann (2014). “Consequences of Faster Alignment of Sequences”. In: *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*, pp. 39–51 (cit. on pp. 3480, 3484, 3489).
- Amir Abboud, Virginia Vassilevska Williams, and Huacheng Yu (2015). “Matching Triangles and Basing Hardness on an Extremely Popular Conjecture”. In: *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pp. 41–50 (cit. on pp. 3480, 3492, 3493).
- Amir Abboud, Richard Ryan Williams, and Huacheng Yu (2015). “More Applications of the Polynomial Method to Algorithm Design”. In: *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pp. 218–230 (cit. on pp. 3474, 3487).
- Manuel Abellanas, Ferran Hurtado, Christian Icking, Rolf Klein, Elmar Langetepe, Li-hong Ma, Belén Palop, and Vera Sacristán (2001). “Smallest Color-Spanning Objects”. In: *Algorithms - ESA 2001, 9th Annual European Symposium, Aarhus, Denmark, August 28-31, 2001, Proceedings*, pp. 278–289 (cit. on p. 3483).
- Udit Agarwal and Vijaya Ramachandran (2016). “Fine-Grained Complexity and Conditional Hardness for Sparse Graphs”. *CoRR* abs/1611.07008 (cit. on p. 3485).
- Parag Agrawal, Arvind Arasu, and Raghav Kaushik (2010). “On indexing error-tolerant set containment”. In: *Proceedings of the ACM SIGMOD International Conference on Management of Data, SIGMOD 2010, Indianapolis, Indiana, USA, June 6-10, 2010*, pp. 927–938 (cit. on p. 3474).
- Josh Alman, Timothy M. Chan, and R. Ryan Williams (2016). “Polynomial Representations of Threshold Functions and Algorithmic Applications”. In: *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pp. 467–476 (cit. on p. 3481).



- Josh Alman and Ryan Williams (2015). “Probabilistic Polynomials and Hamming Nearest Neighbors”. In: *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pp. 136–150 (cit. on p. 3480).
- Amihood Amir, Timothy M. Chan, Moshe Lewenstein, and Noa Lewenstein (2014). “On Hardness of Jumbled Indexing”. In: *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*, pp. 114–125 (cit. on p. 3484).
- Amihood Amir, Tsvi Kopelowitz, Avivit Levy, Seth Pettie, Ely Porat, and B. Riva Shalom (2016). “Mind the Gap: Essentially Optimal Algorithms for Online Dictionary Matching with One Gap”. In: *27th International Symposium on Algorithms and Computation, ISAAC 2016, December 12-14, 2016, Sydney, Australia*, 12:1–12:12 (cit. on p. 3484).
- Daniel Archambault, William S. Evans, and David G. Kirkpatrick (2005). “Computing the Set of all the Distant Horizons of a Terrain”. *Int. J. Comput. Geometry Appl.* 15.6, pp. 547–564 (cit. on p. 3483).
- Esther M. Arkin, Yi-Jen Chiang, Martin Held, Joseph S. B. Mitchell, Vera Sacristán, Steven Skiena, and Tae-Heng Yang (1998). “On Minimum-Area Hulls”. *Algorithmica* 21.1, pp. 119–136 (cit. on p. 3483).
- V. L. Arlazarov, E. A. Dinic, M. A. Kronrod, and I. A. Faradzev (1970). “On economical construction of the transitive closure of an oriented graph”. *Soviet Math. Dokl.* 11, pp. 1209–1210 (cit. on p. 3490).
- Boris Aronov and Sariel Har-Peled (2008). “On Approximating the Depth and Related Problems”. *SIAM J. Comput.* 38.3, pp. 899–921 (cit. on p. 3483).
- Arturs Backurs, Nishanth Dikkala, and Christos Tzamos (2016). “Tight Hardness Results for Maximum Weight Rectangles”. In: *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, 81:1–81:13 (cit. on pp. 3484, 3489).
- Arturs Backurs and Piotr Indyk (2015). “Edit Distance Cannot Be Computed in Strongly Subquadratic Time (unless SETH is false)”. In: *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pp. 51–58 (cit. on p. 3481).
- (2016). “Which Regular Expression Patterns Are Hard to Match?” In: *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pp. 457–466 (cit. on p. 3481).
- Arturs Backurs, Piotr Indyk, and Ludwig Schmidt (2017). “On the Fine-Grained Complexity of Empirical Risk Minimization: Kernel Methods and Neural Networks”. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pp. 4311–4321 (cit. on p. 3482).

- Arturs Backurs, Liam Roditty, Gilad Segal, Virginia Vassilevska Williams, and Nicole Wein (2018). “Towards Tight Approximation Bounds for Graph Diameter and Eccentricities”. In: *Proceedings of the 50th ACM Symposium on Theory of Computing, STOC 2018, Los Angeles, California, USA, 25-29 June 2018*, to appear (cit. on pp. [3478–3480](#)).
- Arturs Backurs and Christos Tzamos (2017). “Improving Viterbi is Hard: Better Runtimes Imply Faster Clique Algorithms”. In: *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pp. 311–321 (cit. on p. [3489](#)).
- Marshall Ball, Alon Rosen, Manuel Sabin, and Prashant Nalini Vasudevan (2017). “Average-case fine-grained hardness”. In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pp. 483–496 (cit. on p. [3494](#)).
- N. Bansal and R. Williams (2009). “Regularity Lemmas and Combinatorial Algorithms”. In: *Proc. FOCS*, pp. 745–754 (cit. on p. [3490](#)).
- I. Baran, E.D. Demaine, and M. Pătraşcu (2008). “Subquadratic Algorithms for 3SUM”. *Algorithmica* 50.4, pp. 584–596 (cit. on p. [3471](#)).
- Gill Barequet and Sarel Har-Peled (2001). “Polygon Containment and Translational Min-Hausdorff-Distance Between Segment Sets are 3SUM-Hard”. *Int. J. Comput. Geometry Appl.* 11.4, pp. 465–474 (cit. on p. [3483](#)).
- Mark de Berg, Marko de Groot, and Mark H. Overmars (1997). “Perfect Binary Space Partitions”. *Comput. Geom.* 7, pp. 81–91 (cit. on p. [3483](#)).
- P. Bille and M. Farach-Colton (2008). “Fast and Compact Regular Expression Matching”. *Theoretical Computer Science*, 409.3, pp. 486–496 (cit. on p. [3469](#)).
- Andreas Björklund, Rasmus Pagh, Virginia Vassilevska Williams, and Uri Zwick (2014). “Listing Triangles”. In: *Automata, Languages, and Programming - 41st International Colloquium, ICALP 2014, Copenhagen, Denmark, July 8-11, 2014, Proceedings, Part I*, pp. 223–234 (cit. on p. [3484](#)).
- G. Blelloch, V. Vassilevska, and R. Williams (2008). “A New Combinatorial Approach to Sparse Graph Problems”. In: *Proc. ICALP*, pp. 108–120 (cit. on p. [3490](#)).
- Prosenjit Bose, Marc J. van Kreveld, and Godfried T. Toussaint (1998). “Filling polyhedral molds”. *Computer-Aided Design* 30.4, pp. 245–254 (cit. on p. [3483](#)).
- Karl Bringmann (2014). “Why Walking the Dog Takes Time: Frechet Distance Has No Strongly Subquadratic Algorithms Unless SETH Fails”. In: *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pp. 661–670 (cit. on p. [3481](#)).
- Karl Bringmann, Pawel Gawrychowski, Shay Mozes, and Oren Weimann (2017). “[Tree Edit Distance Cannot be Computed in Strongly Subcubic Time \(unless APSP can\)](#)”. *CoRR*. arXiv: [1703.08940](#) (cit. on p. [3485](#)).

- Karl Bringmann, Allan Grønlund, and Kasper Green Larsen (2017). “A Dichotomy for Regular Expression Membership Testing”. In: *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pp. 307–318 (cit. on p. 3481).
- Karl Bringmann and Sebastian Krinninger (2017). “Brief Announcement: A Note on Hardness of Diameter Approximation”. In: *31st International Symposium on Distributed Computing, DISC 2017, October 16-20, 2017, Vienna, Austria*, 44:1–44:3 (cit. on p. 3494).
- Karl Bringmann and Marvin Künnemann (2015a). “Improved Approximation for Fréchet Distance on c-packed Curves Matching Conditional Lower Bounds”. In: *Algorithms and Computation - 26th International Symposium, ISAAC 2015, Nagoya, Japan, December 9-11, 2015, Proceedings*, pp. 517–528 (cit. on p. 3481).
- (2015b). “Quadratic Conditional Lower Bounds for String Problems and Dynamic Time Warping”. In: *IEEE 56th Annual Symposium on Foundations of Computer Science, FOCS 2015, Berkeley, CA, USA, 17-20 October, 2015*, pp. 79–97 (cit. on p. 3481).
- Karl Bringmann and Wolfgang Mulzer (2016). “Approximability of the discrete Fréchet distance”. *JoCG* 7.2, pp. 46–76 (cit. on p. 3481).
- Karl Bringmann and Philip Wellnitz (2017). “Clique-Based Lower Bounds for Parsing Tree-Adjoining Grammars”. In: *28th Annual Symposium on Combinatorial Pattern Matching, CPM 2017, July 4-6, 2017, Warsaw, Poland*, 12:1–12:14 (cit. on p. 3489).
- Samuel R. Buss and Ryan Williams (2012). “Limits on Alternation-Trading Proofs for Time-Space Lower Bounds”. In: *Proceedings of the 27th Conference on Computational Complexity, CCC 2012, Porto, Portugal, June 26-29, 2012*, pp. 181–191 (cit. on p. 3467).
- Massimo Cairo, Roberto Grossi, and Romeo Rizzi (2016). “New Bounds for Approximating Extremal Distances in Undirected Graphs”. In: *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pp. 363–376 (cit. on p. 3478).
- Marco L. Carmosino, Jiawei Gao, Russell Impagliazzo, Ivan Mihajlin, Ramamohan Paturi, and Stefan Schneider (2016). “Nondeterministic Extensions of the Strong Exponential Time Hypothesis and Consequences for Non-reducibility”. In: *Proceedings of the 2016 ACM Conference on Innovations in Theoretical Computer Science, Cambridge, MA, USA, January 14-16, 2016*, pp. 261–270 (cit. on p. 3491).
- Timothy M. Chan (2015). “Speeding up the Four Russians Algorithm by About One More Logarithmic Factor”. In: *Proceedings of the Twenty-Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2015, San Diego, CA, USA, January 4-6, 2015*, pp. 212–217 (cit. on p. 3490).
- (2018). “More Logarithmic-Factor Speedups for 3SUM, (median,+)-Convolution, and Some Geometric 3SUM-Hard Problems”. In: *Proceedings of SODA 2018*. to appear (cit. on p. 3471).

- Timothy M. Chan and Ryan Williams (2016). “Deterministic APSP, Orthogonal Vectors, and More: Quickly Derandomizing Razborov-Smolensky”. In: *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pp. 1246–1255 (cit. on p. 3474).
- Yi-Jun Chang (2016). “Hardness of RNA Folding Problem With Four Symbols”. In: *27th Annual Symposium on Combinatorial Pattern Matching, CPM 2016, June 27-29, 2016, Tel Aviv, Israel*, 13:1–13:12 (cit. on p. 3489).
- Krishnendu Chatterjee, Wolfgang Dvorák, Monika Henzinger, and Veronika Loitzenbauer (2016a). “Conditionally Optimal Algorithms for Generalized Büchi Games”. In: *41st International Symposium on Mathematical Foundations of Computer Science, MFCS 2016, August 22-26, 2016 - Kraków, Poland*, 25:1–25:15 (cit. on p. 3481).
- (2016b). “Model and Objective Separation with Conditional Lower Bounds: Disjunction is Harder than Conjunction”. In: *Proceedings of the 31st Annual ACM/IEEE Symposium on Logic in Computer Science, LICS '16, New York, NY, USA, July 5-8, 2016*, pp. 197–206 (cit. on p. 3481).
- Bernard Chazelle (1993). “Cutting Hyperplanes for Divide-and-Conquer”. *Discrete & Computational Geometry* 9, pp. 145–158 (cit. on p. 3474).
- Shiri Chechik, Daniel H. Larkin, Liam Roditty, Grant Schoenebeck, Robert Endre Tarjan, and Virginia Vassilevska Williams (2014). “Better Approximation Algorithms for the Graph Diameter”. In: *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pp. 1041–1052 (cit. on pp. 3476, 3478, 3480).
- Otfried Cheong, Alon Efrat, and Sarel Har-Peled (2007). “Finding a Guard that Sees Most and a Shop that Sells Most”. *Discrete & Computational Geometry* 37.4, pp. 545–563 (cit. on p. 3483).
- Paul Christiano, Jonathan A. Kelner, Aleksander Madry, Daniel A. Spielman, and Shang-Hua Teng (2011). “Electrical flows, laplacian systems, and faster approximation of maximum flow in undirected graphs”. In: *Proceedings of the 43rd ACM Symposium on Theory of Computing, STOC 2011, San Jose, CA, USA, 6-8 June 2011*, pp. 273–282 (cit. on p. 3466).
- Michael B. Cohen, Jonathan A. Kelner, John Peebles, Richard Peng, Anup B. Rao, Aaron Sidford, and Adrian Vladu (2017). “Almost-linear-time algorithms for Markov chains and new spectral primitives for directed graphs”. In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pp. 410–419 (cit. on p. 3466).
- Michael B. Cohen, Yin Tat Lee, Gary L. Miller, Jakub Pachocki, and Aaron Sidford (2016). “Geometric median in nearly linear time”. In: *Proceedings of the 48th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2016, Cambridge, MA, USA, June 18-21, 2016*, pp. 9–21 (cit. on p. 3466).

- Michael B. Cohen, Aleksander Madry, Piotr Sankowski, and Adrian Vladu (2017). “Negative-Weight Shortest Paths and Unit Capacity Minimum Cost Flow in  $\tilde{O}(m^{10/7} \log W)$  Time (Extended Abstract)”. In: *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pp. 752–771 (cit. on p. 3466).
- Michael B. Cohen, Aleksander Madry, Dimitris Tsipras, and Adrian Vladu (2017). “Matrix Scaling and Balancing via Box Constrained Newton’s Method and Interior Point Methods”. In: *58th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2017, Berkeley, CA, USA, October 15-17, 2017*, pp. 902–913 (cit. on p. 3466).
- D. Coppersmith (1997). “Rectangular matrix multiplication revisited”. *Journal of Complexity* 13, pp. 42–49 (cit. on p. 3488).
- M. Cygan, H. N. Gabow, and P. Sankowski (2012). “Algorithmic Applications of Baur-Strassen’s Theorem: Shortest Cycles, Diameter and Matchings”. In: *Proc. FOCS* (cit. on p. 3476).
- Marek Cygan, Holger Dell, Daniel Lokshtanov, Dániel Marx, Jesper Nederlof, Yoshio Okamoto, Ramamohan Paturi, Saket Saurabh, and Magnus Wahlström (May 2016). “On Problems As Hard As CNF-SAT”. *ACM Trans. Algorithms* 12.3, 41:1–41:24 (cit. on p. 3473).
- Marek Cygan, Fedor V. Fomin, Lukasz Kowalik, Daniel Lokshtanov, Dániel Marx, Marcin Pilipczuk, Michal Pilipczuk, and Saket Saurabh (2015). *Parameterized Algorithms*. Springer (cit. on pp. 3471, 3493).
- A. Czumaj and A. Lingas (2007). “Finding a Heaviest Triangle is not Harder than Matrix Multiplication”. In: *Proc. SODA*, pp. 986–994 (cit. on p. 3488).
- Søren Dahlgaard (2016). “On the Hardness of Partially Dynamic Graph Problems and Connections to Diameter”. In: *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, 48:1–48:14 (cit. on pp. 3480, 3491, 3493).
- Søren Dahlgaard, Mathias Bæk Tejs Knudsen, and Morten Stöckel (2017). “Finding even cycles faster via capped k-walks”. In: *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing, STOC 2017, Montreal, QC, Canada, June 19-23, 2017*, pp. 112–120 (cit. on p. 3490).
- Akshay Degwekar, Vinod Vaikuntanathan, and Prashant Nalini Vasudevan (2016). “Fine-Grained Cryptography”. In: *Advances in Cryptology - CRYPTO 2016 - 36th Annual International Cryptology Conference, Santa Barbara, CA, USA, August 14-18, 2016, Proceedings, Part III*, pp. 533–562 (cit. on p. 3494).
- Erik D. Demaine, Andrea Lincoln, Quanquan C. Liu, Jayson Lynch, and Virginia Vassilevska Williams (2018). “Fine-Grained I/O Complexity via Reductions: New lower

- bounds, faster algorithms, and a time hierarchy”. In: *Proc. ITCS*, to appear (cit. on p. 3495).
- Jeff Erickson (1995). “Lower Bounds for Linear Satisfiability Problems”. In: *Proceedings of the Sixth Annual ACM-SIAM Symposium on Discrete Algorithms, 22-24 January 1995. San Francisco, California*. Pp. 388–395 (cit. on p. 3471).
- (1999). “New Lower Bounds for Convex Hull Problems in Odd Dimensions”. *SIAM J. Comput.* 28.4, pp. 1198–1214 (cit. on p. 3483).
- Jeff Erickson, Sarel Har-Peled, and David M. Mount (2006). “On the Least Median Square Problem”. *Discrete & Computational Geometry* 36.4, pp. 593–607 (cit. on p. 3483).
- Jeff Erickson and Raimund Seidel (1995). “Better Lower Bounds on Detecting Affine and Spherical Degeneracies”. *Discrete & Computational Geometry* 13, pp. 41–57 (cit. on p. 3471).
- M. J. Fischer and A. R. Meyer (1971). “Boolean Matrix Multiplication and Transitive Closure”. In: *Proc. FOCS*, pp. 129–131 (cit. on p. 3486).
- J. Flum and M. Grohe (2006). *Parameterized Complexity Theory (Texts in Theoretical Computer Science. An EATCS Series)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc. (cit. on p. 3493).
- Fedor V. Fomin, Daniel Lokshtanov, Michal Pilipczuk, Saket Saurabh, and Marcin Wrochna (2017). “Fully polynomial-time parameterized computations for graphs and matrices of low treewidth”. In: *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pp. 1419–1432 (cit. on p. 3494).
- M.L. Fredman (1976). “New bounds on the complexity of the shortest path problem”. *SIAM Journal on Computing* 5, pp. 49–60 (cit. on p. 3472).
- A. Gajentaan and M. Overmars (1995). “On a class of  $O(n^2)$  problems in computational geometry”. *Computational Geometry* 5.3, pp. 165–185 (cit. on pp. 3471, 3482, 3483).
- Anka Gajentaan and Mark H. Overmars (2012). “On a class of  $O(n^2)$  problems in computational geometry”. *Comput. Geom.* 45.4, pp. 140–152 (cit. on pp. 3471, 3482, 3483).
- François Le Gall and Florent Urrutia (2018). “Improved Rectangular Matrix Multiplication using Powers of the Coppersmith-Winograd Tensor”. In: *Proc. SODA*, to appear (cit. on p. 3488).
- Jiawei Gao, Russell Impagliazzo, Antonina Kolokolova, and R. Ryan Williams (2017). “Completeness for First-Order Properties on Sparse Structures with Algorithmic Applications”. In: *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pp. 2162–2181 (cit. on pp. 3474, 3487).
- Ashish Goel and Pankaj Gupta (2010). “Small subset queries and bloom filters using ternary associative memories, with applications”. In: *SIGMETRICS 2010, Proceedings*



- of the 2010 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems, New York, New York, USA, 14-18 June 2010*, pp. 143–154 (cit. on p. [3474](#)).
- Isaac Goldstein, Tsvi Kopelowitz, Moshe Lewenstein, and Ely Porat (2016). “How Hard is it to Find (Honest) Witnesses?”. In: *24th Annual European Symposium on Algorithms, ESA 2016, August 22-24, 2016, Aarhus, Denmark*, 45:1–45:16 (cit. on p. [3484](#)).
- (2017). “Conditional Lower Bounds for Space/Time Tradeoffs”. In: *Algorithms and Data Structures - 15th International Symposium, WADS 2017, St. John’s, NL, Canada, July 31 - August 2, 2017, Proceedings*, pp. 421–436 (cit. on p. [3495](#)).
- Szymon Grabowski (2014). “New Tabulation and Sparse Dynamic Programming Based Techniques for Sequence Similarity Problems”. In: *Stringology*, pp. 202–211 (cit. on p. [3469](#)).
- J. Hartmanis and R. E. Stearns (1965). “On the computational complexity of algorithms”. *Transactions of the American Mathematical Society* 117, pp. 285–306 (cit. on p. [3466](#)).
- Monika Henzinger, Sebastian Krinninger, Danupon Nanongkai, and Thatchaphol Saranurak (2015). “Unifying and Strengthening Hardness for Dynamic Problems via the On-line Matrix-Vector Multiplication Conjecture”. In: *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing, STOC 2015, Portland, OR, USA, June 14-17, 2015*, pp. 21–30 (cit. on pp. [3490](#), [3491](#)).
- Monika Henzinger, Andrea Lincoln, Stefan Neumann, and Virginia Vassilevska Williams (2017). “Conditional Hardness for Sensitivity Problems”. In: *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, 26:1–26:31 (cit. on pp. [3480](#), [3485](#)).
- E. A. Hirsch (1998). “Two new upper bounds for SAT”. In: *Proc. SODA*, pp. 521–530 (cit. on p. [3467](#)).
- John E. Hopcroft and Robert Endre Tarjan (1974). “Efficient Planarity Testing”. *J. ACM* 21.4, pp. 549–568 (cit. on p. [3466](#)).
- X. Huang and V. Y. Pan (1998). “Fast rectangular matrix multiplication and applications”. *J. of Complexity* 14.2, pp. 257–299 (cit. on p. [3488](#)).
- R. Impagliazzo and R. Paturi (2001). “On the Complexity of k-SAT”. *J. Comput. Syst. Sci.* 62.2, pp. 367–375 (cit. on pp. [3470](#), [3475](#), [3491](#)).
- A. Itai and M. Rodeh (1978). “Finding a minimum circuit in a graph”. *SIAM J. Computing* 7.4, pp. 413–423 (cit. on p. [3488](#)).
- D. B. Johnson (1977). “Efficient algorithms for shortest paths in sparse networks”. *J. ACM* 24.1, pp. 1–13 (cit. on p. [3472](#)).
- Allan Grønlund Jørgensen and Seth Pettie (2014). “Threesomes, Degenerates, and Love Triangles”. In: *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pp. 621–630 (cit. on pp. [3471](#), [3489](#)).

- Daniel M. Kane, Shachar Lovett, and Shay Moran (2017). “Near-optimal linear decision trees for k-SUM and related problems”. *CoRR* abs/1705.01720 (cit. on p. 3471).
- Tsvi Kopelowitz, Seth Pettie, and Ely Porat (2016). “Higher Lower Bounds from the 3SUM Conjecture”. In: *Proceedings of the Twenty-Seventh Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2016, Arlington, VA, USA, January 10-12, 2016*, pp. 1272–1287 (cit. on p. 3484).
- Robert Krauthgamer and Ohad Trabelsi (2017). “Conditional Lower Bounds for All-Pairs Max-Flow”. In: *44th International Colloquium on Automata, Languages, and Programming, ICALP 2017, July 10-14, 2017, Warsaw, Poland*, 20:1–20:13 (cit. on p. 3480).
- Marvin Künnemann, Ramamohan Paturi, and Stefan Schneider (2017). “On the Fine-Grained Complexity of One-Dimensional Dynamic Programming”. In: *44th International Colloquium on Automata, Languages, and Programming, ICALP 2017, July 10-14, 2017, Warsaw, Poland*, 21:1–21:15 (cit. on p. 3482).
- Kasper Green Larsen and R. Ryan Williams (2017). “Faster Online Matrix-Vector Multiplication”. In: *Proceedings of the Twenty-Eighth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2017, Barcelona, Spain, Hotel Porta Fira, January 16-19*, pp. 2182–2189 (cit. on p. 3490).
- François Le Gall (2012). “Faster Algorithms for Rectangular Matrix Multiplication”. In: *53rd Annual IEEE Symposium on Foundations of Computer Science, FOCS 2012, New Brunswick, NJ, USA, October 20-23, 2012*, pp. 514–523 (cit. on p. 3488).
- (2014). “Powers of tensors and fast matrix multiplication”. In: *International Symposium on Symbolic and Algebraic Computation, ISSAC '14, Kobe, Japan, July 23-25, 2014*, pp. 296–303 (cit. on pp. 3476, 3488, 3490).
- Lillian Lee (2002). “Fast context-free grammar parsing requires fast boolean matrix multiplication”. *J. ACM* 49.1, pp. 1–15 (cit. on p. 3490).
- Yin Tat Lee and Aaron Sidford (2014). “Path Finding Methods for Linear Programming: Solving Linear Programs in  $\tilde{O}$ (vrank) Iterations and Faster Algorithms for Maximum Flow”. In: *55th IEEE Annual Symposium on Foundations of Computer Science, FOCS 2014, Philadelphia, PA, USA, October 18-21, 2014*, pp. 424–433 (cit. on p. 3466).
- Andrea Lincoln, Virginia Vassilevska Williams, Joshua R. Wang, and R. Ryan Williams (2016). “Deterministic Time-Space Trade-Offs for k-SUM”. In: *43rd International Colloquium on Automata, Languages, and Programming, ICALP 2016, July 11-15, 2016, Rome, Italy*, 58:1–58:14 (cit. on pp. 3494, 3495).
- Andrea Lincoln, Virginia Vassilevska Williams, and R. Ryan Williams (2018). “Tight Hardness for Shortest Cycles and Paths in Sparse Graphs”. In: *Proc. SODA*, to appear (cit. on pp. 3484, 3485, 3489).
- Aleksander Madry (2013). “Navigating Central Path with Electrical Flows: From Flows to Matchings, and Back”. In: *54th Annual IEEE Symposium on Foundations of Computer*



- Science, FOCS 2013, 26-29 October, 2013, Berkeley, CA, USA*, pp. 253–262 (cit. on p. 3466).
- Aleksander Madry (2016). “Computing Maximum Flow with Augmenting Electrical Flows”. In: *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS 2016, 9-11 October 2016, Hyatt Regency, New Brunswick, New Jersey, USA*, pp. 593–602 (cit. on p. 3466).
- William J. Masek and Michael S. Paterson (1980). “A faster algorithm computing string edit distances”. *Journal of Computer and System Sciences* 20.1, pp. 18–31 (cit. on pp. 3468, 3469).
- Jiří Matoušek (1993). “Range Searching with Efficient Hierarchical Cutting”. *Discrete & Computational Geometry* 10, pp. 157–182 (cit. on p. 3474).
- Sergey Melnik and Hector Garcia-Molina (2003). “Adaptive algorithms for set containment joins”. *ACM Trans. Database Syst.* 28, pp. 56–99 (cit. on p. 3474).
- Daniel Moeller, Ramamohan Paturi, and Stefan Schneider (2016). “Subquadratic Algorithms for Succinct Stable Matching”. In: *Computer Science - Theory and Applications - 11th International Computer Science Symposium in Russia, CSR 2016, St. Petersburg, Russia, June 9-13, 2016, Proceedings*, pp. 294–308 (cit. on p. 3481).
- B. Monien and E. Speckenmeyer (1985). “Solving satisfiability in less than  $2^n$  steps”. *Discrete Applied Mathematics* 10.3, pp. 287–295 (cit. on p. 3467).
- J. Nešetřil and S. Poljak (1985). “On the complexity of the subgraph problem”. *Commentationes Math. Universitatis Carolinae* 26.2, pp. 415–419 (cit. on p. 3488).
- C.H. Papadimitriou (1994). *Computational Complexity*. Addison-Wesley (cit. on p. 3466).
- M. Pătraşcu and R. Williams (2010). “On the Possibility of Faster SAT Algorithms”. In: *Proc. SODA*, pp. 1065–1075 (cit. on p. 3475).
- Mihai Pătraşcu (2010). “Towards polynomial lower bounds for dynamic problems”. In: *Proceedings of the 42nd ACM Symposium on Theory of Computing, STOC 2010, Cambridge, Massachusetts, USA, 5-8 June 2010*, pp. 603–610 (cit. on pp. 3483, 3484).
- R. Paturi, P. Pudlák, M. E. Saks, and F. Zane (2005). “An improved exponential-time algorithm for  $k$ -SAT”. *J. ACM* 52.3, pp. 337–364 (cit. on p. 3467).
- R. Paturi, P. Pudlák, and F. Zane (1999). “Satisfiability Coding Lemma”. *Chicago J. Theor. Comput. Sci.* 1999 (cit. on p. 3467).
- S. Pettie (2004). “A new approach to all-pairs shortest paths on real-weighted graphs”. *Theor. Comput. Sci.* 312.1, pp. 47–74 (cit. on p. 3476).
- Seth Pettie (2008). “All Pairs Shortest Paths in Sparse Graphs”. In: *Encyclopedia of Algorithms* (cit. on p. 3476).
- Seth Pettie and Vijaya Ramachandran (2005). “A Shortest Path Algorithm for Real-Weighted Undirected Graphs”. *SIAM J. Comput.* 34.6, pp. 1398–1431 (cit. on p. 3476).
- Sanguthevar Rajasekaran and Shibu Yooseph (1998). “TAL Recognition in  $O(M(n^2))$  Time”. *J. Comput. Syst. Sci.* 56.1, pp. 83–89 (cit. on p. 3489).

- Karthikeyan Ramasamy, Jignesh M. Patel, Jeffrey F. Naughton, and Raghav Kaushik (2000). “Set Containment Joins: The Good, The Bad and The Ugly”. In: *VLDB 2000, Proceedings of 26th International Conference on Very Large Data Bases, September 10-14, 2000, Cairo, Egypt*, pp. 351–362 (cit. on p. 3474).
- Liam Roditty and Virginia Vassilevska Williams (2011). “Minimum Weight Cycles and Triangles: Equivalences and Algorithms”. In: *IEEE 52nd Annual Symposium on Foundations of Computer Science, FOCS 2011, Palm Springs, CA, USA, October 22-25, 2011*, pp. 180–189 (cit. on p. 3485).
- (2013). “Fast approximation algorithms for the diameter and radius of sparse graphs”. In: *Proceedings of the 45th annual ACM symposium on Symposium on theory of computing*. STOC ’13, pp. 515–524 (cit. on pp. 3476, 3478, 3480).
- Liam Roditty and Uri Zwick (2004). “On Dynamic Shortest Paths Problems”. In: *Algorithms - ESA 2004, 12th Annual European Symposium, Bergen, Norway, September 14-17, 2004, Proceedings*, pp. 580–591 (cit. on pp. 3472, 3485, 3490).
- I. Schiermeyer (1992). “Solving 3-Satisfiability in Less Than  $1.579^n$  Steps”. In: *CSL*, pp. 379–394 (cit. on p. 3467).
- U. Schöning (1999). “A probabilistic algorithm for  $k$ -SAT and constraint satisfaction problems”. In: *Proc. FOCS*, pp. 410–414 (cit. on p. 3467).
- Michael A. Soss, Jeff Erickson, and Mark H. Overmars (2003). “Preprocessing chains for fast dihedral rotations is hard or even impossible”. *Comput. Geom.* 26.3, pp. 235–246 (cit. on p. 3483).
- Daniel A. Spielman and Shang-Hua Teng (2004). “Nearly-linear time algorithms for graph partitioning, graph sparsification, and solving linear systems”. In: *Proceedings of the 36th Annual ACM Symposium on Theory of Computing, Chicago, IL, USA, June 13-16, 2004*, pp. 81–90 (cit. on p. 3466).
- (2014). “Nearly Linear Time Algorithms for Preconditioning and Solving Symmetric, Diagonally Dominant Linear Systems”. *SIAM J. Matrix Analysis Applications* 35.3, pp. 835–885 (cit. on p. 3466).
- A. Stothers (2010). “On the Complexity of Matrix Multiplication”. *Ph.D. Thesis, U. Edinburgh* (cit. on p. 3488).
- Robert Endre Tarjan (1971). “Depth-First Search and Linear Graph Algorithms (Working Paper)”. In: *12th Annual Symposium on Switching and Automata Theory, East Lansing, Michigan, USA, October 13-15, 1971*, pp. 114–121 (cit. on p. 3466).
- (1972). “Depth-First Search and Linear Graph Algorithms”. *SIAM J. Comput.* 1.2, pp. 146–160 (cit. on p. 3466).
  - (1974). “Testing Graph Connectivity”. In: *Proceedings of the 6th Annual ACM Symposium on Theory of Computing, April 30 - May 2, 1974, Seattle, Washington, USA*, pp. 185–193 (cit. on p. 3466).

- P. van Emde Boas (1990). “Machine models and simulations”. In: *The Handbook of Theoretical Computer Science, vol. I: Algorithms and Complexity*. Ed. by J. van Leeuwen. Cambridge, Massachusetts: MIT Press. Chap. 1, pp. 1–61 (cit. on p. 3468).
- V. Vassilevska Williams and R. Williams (2010). “Subcubic equivalences between path, matrix and triangle problems”. In: *Proc. FOCS*, pp. 645–654 (cit. on pp. 3472, 3484–3486, 3488, 3490).
- (2018). “Subcubic equivalences between path, matrix and triangle problems”. *Journal of the ACM*. to appear (cit. on p. 3484).
- Virginia Vassilevska Williams (2012). “Multiplying matrices faster than Coppersmith-Winograd”. In: *Proceedings of the 44th Symposium on Theory of Computing Conference, STOC 2012, New York, NY, USA, May 19 - 22, 2012*, pp. 887–898 (cit. on pp. 3476, 3488, 3490).
- Virginia Vassilevska Williams and Ryan Williams (2013). “Finding, Minimizing, and Counting Weighted Subgraphs”. *SIAM J. Comput.* 42.3, pp. 831–854 (cit. on pp. 3483, 3484, 3489).
- V. Vassilevska and R. Williams (2009). “Finding, minimizing, and counting weighted subgraphs”. In: *Proc. STOC*, pp. 455–464 (cit. on pp. 3483, 3488).
- Robert A. Wagner and Michael J. Fischer (Jan. 1974). “The String-to-String Correction Problem”. *J. ACM* 21.1, pp. 168–173 (cit. on p. 3468).
- Joshua R. Wang (2014). “Space-Efficient Randomized Algorithms for K-SUM”. In: *Algorithms - ESA 2014 - 22th Annual European Symposium, Wroclaw, Poland, September 8-10, 2014. Proceedings*, pp. 810–829 (cit. on p. 3494).
- R. Williams (2007a). “Matrix-vector multiplication in sub-quadratic time (some preprocessing required)”. In: *Proc. SODA*, pp. 995–1001 (cit. on p. 3490).
- R. Ryan Williams (2008). “Time-Space Tradeoffs for Counting NP Solutions Modulo Integers”. *Computational Complexity* 17.2, pp. 179–219 (cit. on pp. 3467, 3495).
- Richard Ryan Williams (2016). “Strong ETH Breaks With Merlin and Arthur: Short Non-Interactive Proofs of Batch Evaluation”. In: *31st Conference on Computational Complexity, CCC 2016, May 29 to June 1, 2016, Tokyo, Japan*, 2:1–2:17 (cit. on p. 3491).
- Ryan Williams (2005). “A new algorithm for optimal 2-constraint satisfaction and its implications”. *Theor. Comput. Sci.* 348.2-3, pp. 357–365 (cit. on pp. 3473, 3475, 3488, 3491).
- (2007b). “Algorithms and Resource Requirements for Fundamental Problems”. CMU-CS-07-147. PhD thesis. Pittsburgh, PA, USA: Carnegie Mellon University (cit. on p. 3475).
- (2014). “Faster all-pairs shortest paths via circuit complexity”. In: *Symposium on Theory of Computing, STOC 2014, New York, NY, USA, May 31 - June 03, 2014*, pp. 664–673 (cit. on pp. 3472, 3489).

- (2018). “On the Complexity of Furthest, Closest, and Orthogonality Problems in Low Dimensions”. In: *Proc. SODA 2018*. to appear (cit. on p. [3482](#)).
- Ryan Williams and Huacheng Yu (2014). “Finding orthogonal vectors in discrete structures”. In: *Proceedings of the Twenty-Fifth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA 2014, Portland, Oregon, USA, January 5-7, 2014*, pp. 1867–1877 (cit. on p. [3474](#)).
- Huacheng Yu (2015). “An Improved Combinatorial Algorithm for Boolean Matrix Multiplication”. In: *Automata, Languages, and Programming - 42nd International Colloquium, ICALP 2015, Kyoto, Japan, July 6-10, 2015, Proceedings, Part I*, pp. 1094–1105 (cit. on p. [3490](#)).

Received 2017-12-07.

VIRGINIA VASSILEVSKA WILLIAMS  
MASSACHUSETTS INSTITUTE OF TECHNOLOGY, EECS AND CSAIL  
[virgi@mit.edu](mailto:virgi@mit.edu)



# CONVECTION-DIFFUSION-REACTION AND TRANSPORT-FLOW PROBLEMS MOTIVATED BY MODELS OF SEDIMENTATION: SOME RECENT ADVANCES

RAIMUND BÜRGER, JULIO CAREAGA, STEFAN DIEHL,  
CAMILO MEJÍAS AND RICARDO RUIZ BAIER

## Abstract

The sedimentation of a suspension is a unit operation widely used in mineral processing, chemical engineering, wastewater treatment, and other industrial applications. Mathematical models that describe these processes and may be employed for simulation, design and control are usually given as nonlinear, time-dependent partial differential equations that in one space dimension include strongly degenerate convection-diffusion-reaction equations with discontinuous coefficients, and in two or more dimensions, coupled flow-transport problems. These models incorporate non-standard properties that have motivated original research in applied mathematics and numerical analysis. This contribution summarizes recent advances, and presents original numerical results, for three different topics of research: a novel method of flux identification for a scalar conservation law from observation of curved shock trajectories that can be observed in sedimentation in a cone; a new description of continuous sedimentation with reactions including transport and reactions of biological components; and the numerical solution of a multi-dimensional sedimentation-consolidation system by an augmented mixed-primal method, including an a posteriori error estimation.

## 1 Introduction

**1.1 Scope.** The sedimentation of small particles dispersed in a viscous fluid under the influence of a (mostly gravitational) body force is a process of theoretical and practical

---

R.B. is supported by Conicyt (Chile) through Fondecyt project 1170473; BASAL project PFB03 CMM, U. de Chile and  $CI^2MA$ , U. de Concepción; and CRHIAM, project CONICYT/FONDAP/15130015. C.M. is supported by Conicyt scholarship. R.R.B. is supported by Engineering and Physical Sciences Research Council (EPSRC) through the grant EP/R00207X/1.

*MSC2010:* primary 35L65; secondary 76T20, 35Q35, 65M60, 65M08.

interest that appears as a controlled unit operation in mineral processing, wastewater treatment, the pulp-and-paper and chemical industry, medicine, volcanology, and other areas where a suspension must be separated into a clarified liquid and concentrated sediment. The authors are involved in the development and the mathematical and numerical analysis of models that describe these processes and may be employed for simulation and control in industrial applications. This contribution provides a survey of some recent advances in this area, which is related to nonlinear, time-dependent partial differential equations (PDEs).

**1.2 Two-phase flow models of sedimentation.** Sedimentation models for these applications should predict the behaviour of a given unit on relatively large temporal and spatial scales, while microscopical information such as the position of a given particle is of little interest. These considerations justify representing the liquid and the solid particles as superimposed continuous phases, namely a liquid phase and one or several solid phases. Since gravity acts in one dimension and computational resources for simulations are limited, spatially one-dimensional models are common. The continuous sedimentation of a suspension subject to applied feed and bulk flows, hindered settling and sediment compressibility can be modelled by a nonlinear, strongly degenerate parabolic PDE for the solids concentration  $\phi = \phi(z, t)$  as a function of depth  $z$  and time  $t$  (Bürger, Karlsen, and Towers [2005]). This PDE is based on the solid and liquid mass balances, and its coefficients depend discontinuously on  $z$ .

To introduce the two-phase flow setting, we let  $\phi$  denote the total solids volume fraction and  $\mathbf{v}_s$  and  $\mathbf{v}_f$  the solids and fluid phase velocity, respectively. Moreover,  $\mathbf{v}_r := \mathbf{v}_s - \mathbf{v}_f$  and  $\mathbf{q} := \phi \mathbf{v}_s + (1 - \phi) \mathbf{v}_f$  are the solid-fluid relative velocity (or drift velocity) and the volume average velocity of the mixture, respectively. Then the conservation of mass equations for the solid and the mixture can be written as

$$(1-1) \quad \partial_t \phi + \nabla \cdot (\phi \mathbf{q} + \phi(1 - \phi) \mathbf{v}_r) = 0, \quad \nabla \cdot \mathbf{q} = 0.$$

A constitutive assumption is introduced to specify  $\mathbf{v}_r$  (see below). In one space dimension, the model (1-1) is closed with  $q$  (i.e.,  $\mathbf{q}$  in one dimension) given by feed input as a function of  $t$  and by operating input and output flows as a piecewise constant function of  $z$ , while in two or three space dimensions, additional equations such as the Navier-Stokes equations need to be solved for the components of  $\mathbf{q}$ . In one space dimension, the simplest complete model is based on the kinematic assumption Kynch [1952] that  $v_r$  is a given function of  $\phi$ , or equivalently, that the hindered settling function  $v_{hs}(\phi) = (1 - \phi)v_r(\phi)$  is given. Then the evolution of  $\phi$  in a column is given by the scalar conservation law

$$(1-2) \quad \partial_t \phi - \partial_x f(\phi) = 0, \quad 0 < x < 1,$$

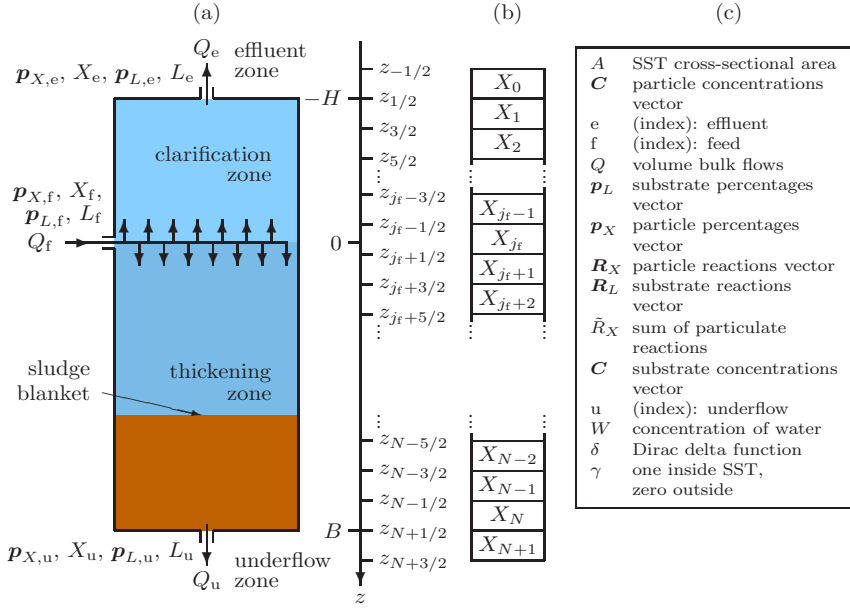


Figure 1: (a) An ideal secondary settling tank (SST) with variables of the feed inlet, effluent and underflow indexed with  $f$ ,  $e$  and  $u$ , respectively (Bürger, Diehl, and Mejías [n.d.]). The sludge blanket (concentration discontinuity) separates the hindered settling and compression regions. (b) Subdivision into computational cells. (c) Nomenclature.

with the nonlinear batch flux density function (Kynch [ibid.])

$$(1-3) \quad f(\phi) = \phi v_{hs}(\phi).$$

Here,  $x$  denotes height,  $x = 0$  is the bottom of the column, and  $x = 1$  the meniscus of the suspension. The initial and boundary conditions are  $\phi(x, 0) = \phi_0$  for  $x \in (0, 1)$ , and  $\phi(0^+, t) = 1$  and  $\phi(1^-, t) = 0$  for  $t > 0$ . If  $f$  has exactly one inflection point, this problem has three different qualitative solutions, depending on the value of  $\phi_0$  (see Bürger and Diehl [2013]). Recent references to the background of (1-2), (1-3) include Betancourt, Bürger, Ruiz-Baier, Torres, and Vega [2014] and Diehl [2012].

**1.3 A model PDE with rough coefficients.** Continuous sedimentation is the process where gravity settling occurs in a large tank which is continuously fed with a suspension and from which a clarified liquid at the top and a thickened slurry at the bottom are



withdrawn. For a tank with constant cross-sectional area this process can in one spatial dimension be modelled by the following PDE:

$$(1-4) \quad \partial_t \phi + \partial_z F(\phi, z, t) = \partial_z (\gamma(z) \partial_z D(\phi)) + s(t) \delta(z).$$

Here, the total flux function  $F(\phi, z, t) = q(z, t)\phi + \gamma(z)f(\phi)$  contains the piecewise constant bulk velocity  $q(\cdot, t)$ , which has a discontinuity at the feed inlet depth  $z = 0$ . The source term is the product of the suspension feed flux  $s(t)$  and the delta distribution  $\delta(z)$ . The characteristic function  $\gamma$  equals 1 inside the tank and 0 outside. Hence,  $F(\phi, \cdot, t)$  has three discontinuities, namely at  $z = 0$  and at the bottom ( $z = B$ ) and top ( $z = -H$ ) of the SST (Figure 1). The batch flux density function is given by (1-3) where  $v_{hs}$  can be given by the Richardson-Zaki expression

$$(1-5) \quad v_{hs}(\phi) = v_0(1 - \phi)^{n_{RZ}}, \quad n_{RZ} \geq 2,$$

by the Vesilind expression  $v_{hs}(\phi) = v_0 \exp(-r_V \phi)$ ,  $r_V > 0$ , or its correction

$$(1-6) \quad v_{hs}(\phi) = v_0 (\exp(-r_V \phi) - \exp(-r_V \phi_{\max})), \quad r_V > 0,$$

or the formula (Diehl [2015])

$$(1-7) \quad v_{hs}(\phi) = v_0 / (1 + (\phi / \bar{\phi})^r), \quad \bar{\phi}, r > 0,$$

where  $v_0 > 0$  is a constant that in (1-5) and (1-7) denotes the settling velocity of single particle in unbounded fluid, and  $\phi_{\max}$  in (1-6) denotes a maximum solids concentration (see Diehl [ibid.] for references). Moreover, sediment compressibility is modeled by the degenerating diffusion term that involves the integrated diffusion coefficient

$$(1-8) \quad D(\phi) = \int_0^\phi \frac{\rho_X v_{hs}(s) \sigma'_e(s)}{g(\rho_X - \rho_L)} ds,$$

where  $\rho_X$  and  $\rho_L$  denote the constant solid and fluid mass densities and  $\sigma'_e$  is the derivative of the so-called effective solid stress function  $\sigma_e = \sigma_e(\phi)$  that satisfies

$$(1-9) \quad \sigma'_e(\phi) = \frac{d\sigma_e(\phi)}{d\phi} = \begin{cases} = 0 & \text{for } \phi \leq \phi_c, \\ > 0 & \text{for } \phi > \phi_c, \end{cases}$$

where  $\phi_c$  denotes a critical concentration above which solid particles are assumed to form a porous network capable of supporting solid stress.

The well-posedness of the model described herein was established and numerical schemes were developed in Bürger, Karlsen, and Towers [2005]. It has meanwhile been extended in various directions, including reactive settling (Bürger, Careaga, Diehl, Mejías, Nopens,

Torfs, and Vanrolleghem [2016] and Bürger, Diehl, and Mejias [n.d.]; see Section 3). Its usefulness for practical simulations (Bürger, Diehl, and Nopens [2011]), however, depends critically on that one can reliably identify the material specific model functions  $f$  and  $\sigma_e$  for the given material. The function  $f$  is usually identified via a batch settling experiment in a cylindrical vessel, but as we show in Section 2, this can be done more efficiently by a settling test in a cone.

**1.4 A multi-dimensional model of sedimentation.** In Section 4 we turn to the description of sedimentation processes in a multidimensional setting. We assume that the viscous fluid is incompressible so its mass and momentum balances are governed by the Navier-Stokes equations with variable viscosity, and the mass balance of the solid phase is described by a nonlinear advection-diffusion equation. Consequently, while in one space dimension one needs to solve only one scalar PDE such as (1-4) for the solids volume fraction  $\phi$ , in several space dimensions we are faced with a system of PDEs that form coupled transport-flow problem for the computation of  $\phi$ , the velocity field  $\mathbf{q}$ , and a pressure  $p$ .

The mathematical difficulties associated with such a problem include highly nonlinear (and typically degenerate) advection and diffusion terms, strong interaction of the  $\mathbf{q}$  and  $\phi$  fields via the Cauchy stress tensor and the forcing term, nonlinear structure of the overall coupled flow-transport problem, saddle-point structure of the flow problem, and non-homogeneous and mixed boundary conditions. These complications affect the solvability analysis of the model, the construction of numerical schemes, and the derivation of stability results and error bounds.

We are also interested in the construction of accurate, robust and reliable methods for the discretization of the model equations, and special emphasis is placed in primal-mixed finite element formulations, meaning that at both continuous and discrete levels, the flow equations possess a saddle-point structure involving the Cauchy stress as additional unknown, whereas the formulation of the advection-diffusion equation is written exclusively in terms of the primal variable, in this case  $\phi$ . Such a structure yields stress approximations without postprocessing them from a low-order discrete velocity (which may lead to insufficiently reliable approximations). In Section 4 we review some recent developments on these lines.

## 2 Flux identification via curved shock trajectories

**2.1 Model of sedimentation in a vessel with varying cross-sectional area.** The batch settling of a suspension of initial concentration  $\phi_0$  in a vessel that occupies the height interval  $x \in [0, 1]$  and that at height  $x$  has the cross-sectional area  $A(x)$  can be described

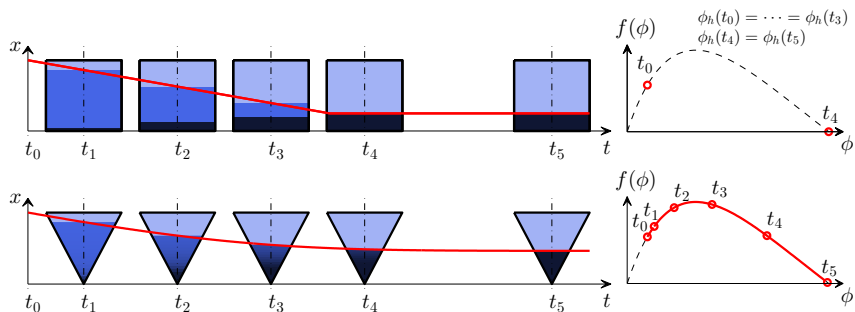


Figure 2: Schematic of settling of a suspension in a cylinder (top) and in a cone (bottom).

by the initial-boundary value problem

$$(2-1) \quad \begin{aligned} \partial_t(A(x)\phi) - \partial_x(A(x)f(\phi)) &= 0, \quad 0 < x < 1, \quad t > 0; \\ \phi(x, 0) &= \phi_0, \quad 0 < x < 1; \quad \phi(0^+, t) = \phi_{\max} = 1, \quad \phi(1^-, t) = 0, \quad t > 0, \end{aligned}$$

where we assume that  $0 \leq f \in C^2$  such that  $f(0) = f(1) = 0$ , with a single maximum at  $\hat{\phi}$  and an inflexion point  $\phi_{\text{infl}} \in (\hat{\phi}, 1]$ , such that  $f''(\phi) < 0$  for  $\phi < \phi_{\text{infl}}$  and  $f''(\phi) > 0$  for  $\phi > \phi_{\text{infl}}$ . Furthermore, we assume that  $A(x)$  is invertible with  $A'(x) \geq 0$ . Specifically, we assume that

$$(2-2) \quad A(x) = \left( \frac{p + qx}{p + q} \right)^{1/q} \quad \text{for } 0 \leq x \leq 1$$

for constants  $p \geq 0$  and  $q \geq 0$  ( $p^2 + q^2 \neq 0$ ). Of particular interest is the case  $p = 0$  and  $q = 1/2$  that corresponds to a full cone, while  $p > 0$  and  $q = 1/2$  refers to a truncated cone. Cones are widely used for routine tests in sanitary engineering, where they are known as “Imhoff cones” (Bürger, Careaga, Diehl, Merckel, and Zambrano [n.d.]). The recent contribution by Bürger, Careaga, and Diehl [2017] related to (2-1) is the construction of explicit solutions to this problem. The basic difficulty associated with (2-1) is that characteristic curves and iso-concentration lines do not coincide. Furthermore, our solution handles functions  $f$  that have one inflection point, while the solution to (2-1) by Anestis [1981] was reduced to  $f(\phi) = \phi(1 - \phi)$ .

The practical interest in solving (2-1) for settling in a cone is illustrated in Figure 2: it turns out that in the conical case, the concentration  $\phi$  beneath the suspension-supernate interface gradually increases, so that the velocity of descent of that interface decreases, while

in the cylindrical case that concentration and velocity are constant. As a consequence, that velocity of descent depends on a whole interval of  $\phi$ -values and corresponding flux values  $f(\phi)$ . It is therefore possible to reconstruct the function  $\phi \mapsto f(\phi)$  on a whole interval, which may be as large as  $(\phi_0, \phi_{\max}]$ , where  $\phi_0$  is the initial concentration, from a single batch test, while the cylindrical case permits only to obtain one point  $(\phi_0, f(\phi_0))$  in addition to  $(\phi_{\max}, f(\phi_{\max}))$ , so a separate test has to be performed for each initial concentration.

**2.2 Solution of the initial-boundary value problem.** The reconstruction is achieved through the exact solution of (2-1) by the method of characteristics wherever  $\phi$  is smooth, combined with the solution of the ordinary differential equations for the suspension height  $h$  as a function of time  $t$ . The method of characteristics (see Holden and Risebro [2015]), applied to the PDE in (2-1) written in quasilinear form  $\partial_t \phi - f'(\phi) \partial_x \phi = (A'(x)/A(x)) f(\phi)$ , yields that we may choose  $t$  as a parameter along characteristics, and that for a non-characteristic initial curve  $(x, t, \phi) = (\xi, \tau, \varphi)$ , the quantities  $x = X(t)$  and  $\phi = \Phi(t)$  satisfy the characteristic equations

$$\begin{aligned} X'(t) &= -f'(\Phi), & t > \tau; & & \Phi'(t) &= (A'(X)/A(X)) f(\Phi), & t > \tau; \\ X(\tau) &= \xi, & & & \Phi(\tau) &= \varphi, \end{aligned}$$

from which we already read off that  $A' > 0$  implies  $\Phi' > 0$ , i.e. the concentration increases along characteristics. For  $A$  given by (2-2) we get the characteristic system

$$(2-3) \quad \frac{t - \tau}{p + qx} = f(q) \int_{\varphi}^{\Phi} \frac{d\Phi}{f(\Phi)^{1+q}}, \quad \frac{f(\phi)}{f(\varphi)} = \left( \frac{p + q\xi}{p + qx} \right)^{1/q}.$$

For  $\varphi = \phi_0$  specified at initial time  $\tau = 0$ , the first equation in (2-3) yields

$$(2-4) \quad \psi(x, t) := \frac{t}{p + qx} = f(\phi)^q \int_{\phi_0}^{\phi} \frac{d\Phi}{f(\Phi)^{1+q}} =: Q(\phi).$$

Thus, the solution  $\phi = \phi(x, t)$  for small times is implicitly given by the relation

$$(2-5) \quad \psi(x, t) = Q(\phi),$$

where  $Q$  is invertible in closed form only in exceptional cases. However, (2-5) informs that the curves of constancy of  $\psi$  in an  $x$  versus  $t$  plot are those of  $\phi$ , and for a (truncated) cone ( $q = 1/2$ ), these are straight lines that intersect at  $x = -p/q$ .

The integral in (2-4) cannot be evaluated in closed form in general, but this is possible for the following case treated in Anestis [1981]:

$$(2-6) \quad f(\phi) = \phi(1 - \phi/\phi_{\max}), \quad q = 1/2.$$

Here we emphasize that our treatment (Bürger, Careaga, and Diehl [2017]) is based on integrals with respect to  $\phi$ , while that of Anestis [1981] is based on integrating over values of  $f$ . This is the key insight that allowed us to handle flux functions having an inflection point.

Of course it is well known that the projected characteristics  $t \mapsto x(t)$  for a quasi-linear first-order PDE may intersect after finite time and give rise to discontinuities. If  $\phi^+(t) \neq \phi^-(t)$  are solution values adjacent to a curve  $t \mapsto x_d(t)$ , then these must satisfy the Rankine-Hugoniot condition

$$(2-7) \quad -x'_d = S(\phi^-, \phi^+) := (f(\phi^+) - f(\phi^-))/(\phi^+ - \phi^-)$$

and the entropy jump condition

$$(2-8) \quad S(u, \phi^-) \geq S(\phi^+, \phi^-) \text{ for all } u \text{ between } \phi^+ \text{ and } \phi^-.$$

**Definition 2.1.** *A function  $\phi$  is an entropy solution of (2-1) if  $\phi$  is a  $C^1$  solution of (2-1) everywhere with the exception of a finite number of curves  $x_d(t) \in C^1$  of discontinuities. At each jump,  $\phi^\pm := \phi(x_d(t)^\pm, t)$  satisfy (2-7) and (2-8).*

Our approach is based on piecing together solutions  $\phi = \phi(x, t)$  in smooth regions, where these are defined by (2-3), along with trajectories of discontinuities that satisfy (2-7) and (2-8). The entropy solution defined here is also the unique entropy solution in the sense of Kružkov-type entropy inequalities (Holden and Risebro [2015]). Such a solution may be used to provide exact reference solutions to test numerical schemes.

We illustrate in Figure 3 the construction for the case (2-6), for which the integral in (2-4) is available in closed form and  $Q$  is invertible, as considered in Anestis [1981]. The characteristics are upwards-bent curves, and the straight lines  $\psi = \text{const.}$  intersect at  $x = -p/q = -1/9$ . These lines carry  $\phi$ -values ranging from  $\phi_0 = 0.35$  to  $\phi_{\max} = 0.66$ . The characteristic area is enclosed by two convex curves that separate the suspension from the clear liquid region ( $\phi = 0$ ) and the sediment ( $\phi = \phi_{\max}$ ) from the suspension, and which intersect at some time to form a stationary solution.

The construction of an entropy solution for a function  $f$  having an inflection point is more involved; see Bürger, Careaga, and Diehl [2017, n.d.] for full details. We here only provide those preliminaries that permit stating the final results in self-contained form.

To classify the generic cases that may arise for a function  $f$  with exactly one inflection point  $\phi_{\text{infl}}$ , we introduce the operations  $\phi \mapsto \phi^*$  and  $\phi \mapsto \phi^{**}$ :

$$\begin{aligned} \phi^* &:= \sup \{u > \phi : S(\phi, u) \leq S(\phi, v) \ \forall v \in (\phi, u]\} \quad \text{for } \phi \in [0, \phi_{\text{infl}}], \\ \phi^{**} &:= \inf \{u < \phi : u^* = \phi\} \quad \text{for } \phi \in [\phi_{\text{infl}}, \phi_{\max}]. \end{aligned}$$

The generic cases are then those of a low (L), medium (M), and high (H) value of  $\phi_0$  in terms of comparisons with  $\phi_{\text{infl}}$  and  $\phi_{\max}^{**}$ , see Figure 4.



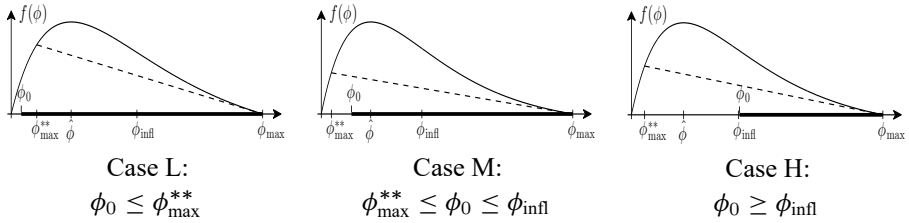


Figure 4: Generic cases of a low (L), medium (M), and high (H) value of  $\phi_0$ . The thick lines show the intervals of possible identification of the flux.

- (iv)  $R_{\text{IIa}} = \emptyset$  if  $\phi_{\text{infl}} \leq \phi_0 < \phi_{\max}$  (Case H) or if  $P(\phi_{\text{infl}}) \leq 0$  and  $\phi_G < \phi_0 < \phi_{\text{infl}}$ . Otherwise,  $\phi > \phi_{\text{infl}}$  in  $R_{\text{IIa}}$ , and strictly concave characteristics emanate tangentially from  $b(t)$  for  $t_1 \leq t \leq t_2$ .
- (v)  $R_{\text{IIb}} = \emptyset$  if  $\phi_0 \leq \phi_{\max}^{**}$  (Case L). Otherwise  $R_{\text{IIb}}$  is filled with concave characteristics emanating from  $(x, t) = (0, 0)$  with initial values in  $(\phi_0^*, \phi_{\max})$  in Case M, and in  $(\phi_0, \phi_{\max})$  in Case H.

Note that [Theorem 2.1](#) does not cover the case of a full cone, that is,  $q = 1/2$  and  $p = 0$ . In fact, it is not entirely straightforward to take the limit  $p \rightarrow 0^+$  in the proof of [Bürger, Careaga, and Diehl \[2017\]](#) since a singularity arises at  $(x, t) = (0, 0)$  even if no singularity is created for  $p > 0$ . For the identification problem, the case  $p = 0$  is of interest since full cones are common laboratory equipment, and more importantly, for the following reason. The conversion of the curve  $(t, h(t))$  into a portion of the flux, that is, into pairs  $(\phi, f(\phi))$  on a certain  $\phi$ -interval is possible for  $0 \leq t \leq t_{2.5}$  [Bürger, Careaga, and Diehl \[2017, n.d.\]](#). However, the time  $t_{2.5}$ , that is the moment of merger of  $b(t)$  and  $h(t)$ , may be hard to be detect. Fortunately, for  $p = 0$  it turns out that  $t_{2.5} = t_3$  (under some mild conditions), and therefore the entire curve  $h(t)$  may be used for all times for flux identification. The following theorem is proved in [Bürger, Careaga, and Diehl \[n.d.\]](#).

**Theorem 2.2.** Assume that  $A$  is given by (2-2) with  $p = 0$  and  $q > 0$ . The entropy solution  $\phi = \phi(x, t)$  of (2-1) is piecewise smooth and satisfies (i) and (ii) of [Theorem 2.1](#). If  $f'(\phi_{\max}) < 0$ , then  $t_3 < \infty$  and  $\phi \equiv \phi_{\max}$  in  $R_{\text{III}}$ , which is bounded by the upper shock curve  $x = h(t)$  and the line  $x = \ell(t) := -f'(\phi_{\max})t$ . If  $f'(\phi_{\max}) = 0$ , then  $R_{\text{III}} = \emptyset$ . Furthermore, we define  $P(\phi) := \frac{\mathcal{Q}'(\phi)}{qf(\phi)^{q-1}}$ , and have the following.

- (i) Independently of  $\phi_0$ : If  $P(\phi_{\text{infl}}) > 0$ , then the solution is continuous in  $0 \leq x \leq h(t)$ ,  $t > 0$ , without a bottom discontinuity  $b(t)$ . (See [Figure 6](#).)
- (ii) If  $\phi_0 \leq \phi_{\text{infl}}$  (Cases L and M) and  $P(\phi_{\text{infl}}) \leq 0$ , then the solution has both discontinuities, where  $b(t)$  is a straight line originating from the bottom, having the

constant  $\phi = \phi_G$  just above it, where  $G(\phi_G) = 0$  and we define the function  $G(\varphi) := S(\varphi, \varphi^-) + \frac{1}{qQ(\varphi)}$ .

**2.3 Solution of the inverse problem and curved trajectories.** Let us now come back to the inverse problem. We assume that  $A(x)$  is given by (2-2) with  $p, q \geq 0$ , that the initial concentration  $\phi_0$  is given, and that the flux is unknown but has the following properties:  $f \in C^2$  is a nonnegative function with  $f(0) = f(1) = 0$ , one maximum  $\hat{\phi}$  and one inflection point  $\phi_{\text{infl}} \in (\hat{\phi}, 1]$  such that  $f''(\phi) < 0$  for  $\phi < \phi_{\text{infl}}$  and  $f''(\phi) > 0$  for  $\phi > \phi_{\text{infl}}$ . Then the inverse problem can be formulated as follows (see Figure 2):

(IP) *Given the interface trajectory  $[t_{\text{start}}, t_{\text{end}}] \ni t \mapsto h(t)$ , find the portion of  $\phi \mapsto f(\phi)$  corresponding to the interval of adjacent  $\phi$ -values.*

The idea to solve (IP) is based on the representation of the explicit solution according to Theorems 2.1 and 2.2. In Bürger, Careaga, and Diehl [ibid.] the solution of (IP) is given as a parametric explicit formula for the flux. If  $h(t)$  is not provided in closed algebraic form, for instance if only pointwise experimental data are available, then a suitable decreasing and convex approximation can be generated by solving a constrained least-squares approximation (quadratic programming) problem; see Bürger, Careaga, and Diehl [n.d.] and Bürger and Diehl [2013].

To elucidate a relation between curved shock trajectories and the functional form of the nonlinear flux, let us consider for the moment the cylindrical case  $A \equiv \text{const.}$ , for which the identification problem was handled in Bürger and Diehl [2013]. Then, the upper discontinuity  $x = h(t)$  is initially a straight line; see Figure 2. For a medium large initial value  $\phi_0 \in (\phi_{\text{max}}^*, \phi_{\text{infl}})$ , a rarefaction fan emerges from  $(x, t) = (0, 0)$ . After this wave has met the upper discontinuity  $h(t)$  at  $t = t_{\text{start}}$ , the latter becomes convex for some  $t \in [t_{\text{start}}, t_{\text{end}}]$ . Kynch [1952] presented a graphical procedure for obtaining  $f$  in the interval  $[\phi_0^*, \phi_{\text{max}}]$  (the ‘tail’). Diehl [2007] showed that Kynch’s graphical procedure can be written by representation formulas; namely the tail of  $f$  can be expressed as a function of the curved discontinuity  $h$  and its derivative  $h'$ . This is a solution of the inverse problem of obtaining (the tail of) the flux function  $f$  given the solution of (2-1) with  $A \equiv \text{const.}$ . It is interesting to note that Kunik [1993] presented a representation formula for the global solution of (2-1) with  $A \equiv \text{const.}$  for a monotone initial value function  $\phi(x, 0) = \phi_{\text{init}}(x)$ ,  $0 \leq x \leq 1$ . In the special batch-sedimentation case where  $\phi_{\text{init}} \equiv \phi_0$ , Kunik’s formulas relate the curved discontinuity  $h$  as a function of the flux function  $f$  in precisely the same way as Diehl’s formulas relate  $f$  as a function of  $h$ . To elucidate this symmetry, we denote the concentration just below the curved discontinuity by

$$(2-9) \quad \phi_h(t) := \phi(h(t)^-, t) \quad \text{for } t_{\text{start}} \leq t \leq t_{\text{end}},$$



where  $\phi_h$  is an increasing  $C^1$  function that maps  $[t_{\text{start}}, t_{\text{end}}]$  to  $[\phi_0^*, \phi_{\text{max}}]$ . In the rest of this section, we restrict  $h$  and  $f$  to these respective intervals. Evaluating the formula  $x/t = -f'(\phi(x, t))$ , which describes the slope of characteristics within the rarefaction wave, and inserting (2-9) we obtain

$$(2-10) \quad h(t)/t = -f'(\phi_h(t)) \quad \text{for } t_{\text{start}} \leq t \leq t_{\text{end}}.$$

On the other hand, the jump condition (2-7) for  $x = h(t)$  implies that

$$(2-11) \quad -h'(t) = f(\phi_h(t))/\phi_h(t) \quad \text{for } t_{\text{start}} \leq t \leq t_{\text{end}}.$$

Note that replacing  $h$  by  $f$  and  $t$  by  $\phi_h$  in any of the formulas (2-10) and (2-11), the other is obtained. In fact, defining  $\eta(t) := h(t) - th'(t)$  and  $\tilde{\Phi}(\phi) := f(\phi) - \phi f'(\phi)$ , we obtain the following dual representation formulas [Bürger and Diehl \[2013\]](#):

$$(2-12) \quad (\phi, f(\phi)) = (H\phi_0/\eta(t))(1, -h'(t)) \quad \text{for } t_{\text{start}} \leq t \leq t_{\text{end}},$$

$$(2-13) \quad (t, h(t)) = (H\phi_0/\tilde{\Phi}(\phi))(1, -f'(\phi)) \quad \text{for } H\phi_0/\eta(t_{\text{start}}) = \phi_0^* \leq \phi \leq \phi_{\text{max}},$$

where (2-12) was derived by [Diehl \[2007\]](#) and (2-13) by [Kunik \[1993\]](#). Both  $f$  and  $h$  are decreasing, strictly convex and  $C^2$  functions (on the intervals of interest). Since both  $\eta$  and  $\tilde{\Phi}$  are invertible, explicit representation formulas may be obtained:

$$f(\phi) = -\phi h'(\eta^{-1}(H\phi_0/\phi)) \quad \text{for } \phi_0^* \leq \phi \leq \phi_{\text{max}},$$

$$h(t) = -tf'(\tilde{\Phi}^{-1}(H\phi_0/t)) \quad \text{for } t_{\text{start}} \leq t \leq t_{\text{end}}.$$

**2.4 A numerical example.** We are currently applying the new method of flux identification to synthetic and experimental data ([Bürger, Careaga, Diehl, Merckel, and Zambrano \[n.d.\]](#)). We show in [Figure 1](#) the numerical solution to a problem of flux recognition. The flux function  $f(\phi)$  defined by (1-3) and (1-6) with  $r_V = 5$  was used to produce the upper discontinuity by solving the corresponding jump condition ODE numerically. From the ODE solution, discrete data points were obtained and used to fit a piecewise cubic polynomial function  $h(t)$ . This function is then used in the explicit parametric formula (see [Bürger, Careaga, and Diehl \[n.d.\]](#)) for the flux. With sufficiently many data points, containing hardly any noise, many subintervals can be used and a portion of the flux identified accurately.

### 3 Reactive settling

**3.1 Introduction.** Models of continuously operated settling tanks form a topic for well-posedness and numerical analysis even in one space dimension due to the spatially discontinuous coefficients of the underlying strongly degenerate parabolic, nonlinear model

PDE (1-4). Such a model was recently extended (Bürger, Careaga, Diehl, Mejías, Nopens, Torfs, and Vanrolleghem [2016] and Bürger, Diehl, and Mejías [n.d.]) to multi-component particles that react with several soluble constituents of the liquid phase. The fundamental balance equations contain the mass percentages of the components of both phases. The equations are reformulated in Bürger, Diehl, and Mejías [n.d.] as a system of nonlinear PDEs that can be solved by an explicit numerical difference scheme. The scheme itself is not described in this contribution since space is limited. It combines a difference scheme for conservation laws with discontinuous flux, similar to that of Bürger, Karlsen, and Towers [2005], with numerical percentage propagation for multi-component flows (Diehl [1997]).

**3.2 Mathematical model.** The main variables are explained in Figure 1. The unknowns are  $X$ ,  $L$ ,  $\mathbf{p}_X$  and  $\mathbf{p}_L$  as functions of  $z$  and  $t$ . The solid and fluid densities,  $\rho_X$  and  $\rho_L$ , are assumed constant. The model keeps track of  $k_X$  particulate and  $k_L$  liquid components ( $k_L - 1$  substrates and water), whose concentrations are collected in vectors  $\mathbf{C}$  and  $\mathbf{S}$  along with  $W$ , or equivalently, percentage vectors  $\mathbf{p}_X$  and  $\mathbf{p}_L$ :

$$\mathbf{C} = \mathbf{p}_X X = \begin{pmatrix} p_X^{(1)} \\ \vdots \\ p_X^{(k_X)} \end{pmatrix} X, \quad \mathbf{p}_L L = \begin{pmatrix} p_L^{(1)} \\ \vdots \\ p_L^{(k_L)} \end{pmatrix} L = \begin{pmatrix} \mathbf{S} \\ W \end{pmatrix} = \begin{pmatrix} S^{(1)} \\ \vdots \\ S^{(k_L-1)} \\ W \end{pmatrix},$$

where  $p_X^{(1)} + \dots + p_X^{(k_X)} = 1$  and  $p_L^{(1)} + \dots + p_L^{(k_L)} = 1$ . The governing system of equations can be formulated as follows:

(3-1)

$$\partial_t X + \partial_z F_X = \delta(z) \frac{X_f Q_f}{A} + \gamma(z) \tilde{R}_X(X), \quad F_X := Xq + \gamma(z)(f(X) - \partial_z D(X)),$$

$$\partial_t (\mathbf{p}_X X) + \partial_z (\mathbf{p}_X X) = \delta(z) \frac{\mathbf{p}_{X,f} X_f Q_f}{A} + \gamma(z) \mathbf{R}_X,$$

$$L = \rho_L (1 - X/\rho_X),$$

$$\partial_t (\bar{\mathbf{p}}_L L) + \partial_z (\bar{\mathbf{p}}_L L) = \delta(z) \frac{\bar{\mathbf{p}}_{L,f} X_f Q_f}{A} + \gamma(z) \bar{\mathbf{R}}_L, \quad F_L := \rho_L \left( q - \frac{F_X}{\rho_X} \right),$$

$$p_L^{(k_L)} = 1 - (p_L^{(1)} + \dots + p_L^{(k_L-1)})$$

for  $z \in \mathbb{R}$  and  $t > 0$ , along with suitable initial conditions. The convective flux function  $F_X$  contains the spatially discontinuous bulk velocity  $q(z, t)$ , the hindered-settling flux function  $f$  given by (1-3) and the sediment compressibility function  $D$  by (1-8). Moreover,  $\bar{\mathbf{p}}_L = \bar{\mathbf{p}}_L(z, t)$  is a vector of components of the liquid phase formed by the

first  $k_L - 1$  components of  $\mathbf{p}_L$ . The reaction term vectors are denoted by  $\mathbf{R}_X$  and  $\bar{\mathbf{R}}_L$ , and lastly  $\tilde{\mathbf{R}}_X$  is the sum of all components of the vector  $\mathbf{R}_X$ .

The model (3-1) may include a full biokinetic Activated Sludge Model (ASMx; see [Henze, Grady, Gujer, Marais, and Matsuo \[1987\]](#)) at every depth  $z$  within  $\mathbf{R}_X$  and  $\bar{\mathbf{R}}_L$ , and is based on the idea that hindered and compressive settling depend on the total particulate concentration (flocculated biomass)  $X$  modelled by the first equation. The particular formulation (3-1) has two advantages. Firstly, for a numerical method with explicit time stepping such as the one advanced in [Bürger, Diehl, and Mejías \[n.d.\]](#), the new value of  $X$  is obtained by solving the first equation in (3-1) only. Then  $\mathbf{p}_X$  is updated by the second equation of (3-1), etc. Secondly, this formulation yields the invariant region property of the numerical scheme (see [Bürger, Diehl, and Mejías \[ibid., Theorem 4.1\]](#)), which states that the solution stays in

$$\tilde{\Omega} := \{ \mathbf{u} \in \mathbb{R}^{k_X + k_L + 2} : 0 \leq \mathbf{p}_X, \mathbf{p}_L \leq 1, 0 \leq X \leq X_{\max}, \\ \rho_L - rX_{\max} \leq L \leq \rho_L, p_X^{(1)} + \dots + p_X^{(k_X)} = 1, p_L^{(1)} + \dots + p_L^{(k_L)} = 1 \}$$

(vectors in inequalities should be interpreted component-wise), provided that the spatial meshwidth and the time step satisfy a suitable CFL condition.

We have no proof that an exact solution of system (3-1) stays in  $\tilde{\Omega}$  if the initial datum does since the well-posedness (existence and uniqueness) analysis of the model is not yet concluded, and a suitable concept of a (discontinuous) exact solution is not yet established. However, it is reasonable to expect that an exact solution of (3-1) should also assume values within  $\tilde{\Omega}$ . To support this conjecture, we mention first that the invariant region property proved in [Bürger, Diehl, and Mejías \[ibid.\]](#) holds uniformly for approximate solutions, and therefore will hold for any limit to which the scheme converges as discretization parameters tend to zero. This standard argument has been used for related models in [Bürger, Karlsen, Risebro, and Towers \[2004\]](#), [Bürger, Karlsen, and Towers \[2005\]](#), and [Karlsen, Risebro, and Towers \[2002\]](#). With the properties of the reaction term here, namely that  $\tilde{\mathbf{R}}_X = 0$  if  $X = 0$  or  $X = X_{\max}$ , the invariance property of the numerical scheme follows by a monotonicity argument ([Bürger, Diehl, and Mejías \[n.d., Lemma 4.3\]](#)). The convergence of that scheme with a reaction term being a function of  $X$  only (and utilizing that it is zero for  $X = 0$  or  $X = X_{\max}$ ) can be established by modifying the proof in [Bürger, Karlsen, and Towers \[2005\]](#).

**3.3 Numerical example.** To specify the function  $f$  given by (1-3), we utilize (1-7) with volume fraction  $\phi$  replaced by the equivalent local density  $X$  and the parameters  $\bar{X} = 3.87 \text{ kg m}^{-3}$  and  $r = 3.58$ . The function  $D$  that describes sediment compressibility is specified by (1-8), where we choose  $\sigma_e = 0$  for  $X < X_c$  and  $\sigma_e(X) = \alpha(X - X_c)$  for

$X > X_c$  with  $\alpha = 0.2 \text{ m}^2 \text{ s}^{-2}$  and  $X_c = 5 \text{ kg m}^{-3}$ . The velocity  $q$  is defined in terms of the given bulk flows as

$$q(z, t) = \frac{1}{A} \cdot \begin{cases} Q_e(t) = Q_f(t) - Q_u(t) & \text{for } z < 0, \\ Q_u(t) & \text{for } z > 0, \end{cases} \quad \text{where } A = 400 \text{ m}^2.$$

We use a reduced biological model of denitrification, distinguishing  $k_X = 2$  particulate components with concentrations  $X_{\text{OHO}}$  (ordinary heterotrophic organisms) and  $X_U$  (undegradable organics), and  $k_L = 4$  liquid components, namely the substrates  $S_{\text{NO}_3}$  (nitrate),  $S_S$  (readily biodegradable substrate) and  $S_{\text{N}_2}$  (nitrogen), and water, such that  $\mathbf{p}_X X = \mathbf{C} = (X_{\text{OHO}}, X_U)^T$  and  $\mathbf{S} = (S_{\text{NO}_3}, S_S, S_{\text{N}_2})^T$ . The reaction terms are then given by

$$\mathbf{R}_L = X_{\text{OHO}} \begin{pmatrix} -\frac{1-Y}{2.86Y} \mu(\mathbf{S}) \\ (1-f_p)b - \frac{1}{Y} \mu(\mathbf{S}) \\ \frac{1-Y}{2.86Y} \mu(\mathbf{S}) \\ 0 \end{pmatrix}, \quad \begin{aligned} \mathbf{R}_X &= X_{\text{OHO}} \begin{pmatrix} \mu(\mathbf{S}) - b \\ f_p b \end{pmatrix}, \\ \mu(\mathbf{S}) &:= \mu_{\max} \frac{S_{\text{NO}_3}}{K_{\text{NO}_3} + S_{\text{NO}_3}} \frac{S_S}{K_S + S_S}, \end{aligned}$$

where  $\mu(\mathbf{S})$  is the so-called growth rate function. (Values of constants are given in the caption of [Figure 9](#).) The resulting summed reaction terms are

$$\tilde{R}_X = (\mu(\mathbf{S}) - (1-f_p)b) X_{\text{OHO}}, \quad \tilde{R}_L = \left( (1-f_p)b - \frac{\mu(\mathbf{S})}{Y} \right) X_{\text{OHO}}.$$

We choose the volumetric flows  $Q_f$  and  $Q_u$  and the feed concentration  $X_f$  as piecewise constant functions of  $t$  (see [Figure 8](#)), and let  $\mathbf{p}_{X,f}$  and  $\mathbf{p}_{L,f}$  be constant.

The whole simulation is shown in [Figure 9](#). The initial steady state is kept during two hours of the simulation. There is a sludge blanket, i.e., a discontinuity from a low concentration up to  $X = X_c$ . At  $t = 4$  h, the step change of control functions causes a rapidly rising sludge blanket that nearly reaches the top of the SST around  $t = 5.8$  h, when the control variables are changed again. The fast reactions imply that the soluble  $\text{NO}_3$  is quickly converted to  $\text{N}_2$  in regions where the bacteria OHO are present, which is below the sludge blanket.

## 4 A multi-dimensional sedimentation model

**4.1 Coupled transport-flow problem.** Consider an incompressible mixture occupying the domain  $\Omega \subset \mathbb{R}^d$ ,  $d = 2$  or  $d = 3$ , and that the velocities  $\mathbf{q}$  and  $\mathbf{v}_r$  are as defined in

**Section 1.2.** Following [Bürger, Wendland, and Concha \[2000\]](#) and discarding quadratic terms for the filtration velocity, we may recast the governing equations as follows (cf. [Ruiz-Baier and Lunati \[2016\]](#)):

$$\begin{aligned}
 & \operatorname{div} \mathbf{q} = 0, \\
 & \partial_t \phi + \operatorname{div}(\phi \mathbf{q} - b(\phi) \mathbf{k}) = \operatorname{div}(\kappa(\phi) \nabla \phi), \\
 (4-1) \quad & \partial_t \mathbf{q} + \mathbf{q} \cdot \nabla \mathbf{q} - \frac{1}{\rho} \operatorname{div}(\mu(\phi) \boldsymbol{\varepsilon}(\mathbf{q}) - p \mathbf{I}) = \mathcal{Q}(\phi)(\partial_t \mathbf{v}_r + \mathbf{q} \cdot \nabla \mathbf{v}_r) \\
 & \quad + \mathcal{Q}(\phi) \mathbf{v}_r \cdot \nabla \mathbf{q} + g \mathbf{k},
 \end{aligned}$$

where  $\rho = \phi \rho_X + (1 - \phi) \rho_L$  is the local density of the mixture,  $\mathcal{Q}(\phi) = \rho^{-1}(\rho_X - \rho_L)\phi(1 - \phi)$ , and  $b(\phi)$  is the Kynch batch flux density function, i.e.,  $b(\phi) = f(\phi)$  in the notation of Sections 1.2 and 1.3, where we assume that this function is given by (1-3), (1-5) with  $n_{\text{RZ}} = 0$ . The coefficient functions  $\kappa(\phi) := (\mathrm{d}D(\phi)/\mathrm{d}\phi)/\rho_X$  (see (1-8)) and  $\mu(\phi) := (1 - \phi)^{-3}$  account for compressibility of the sediment and mixture viscosity, respectively.

The primal unknowns are the volume average flow velocity of the mixture  $\mathbf{q}$ , the solids concentration  $\phi$ , and the pressure field  $p$ . Next we proceed to recast (4-1) in mixed form, also making the assumption that the flow regime is laminar: Find the Cauchy fluid pseudo-stress  $\boldsymbol{\sigma}$ , the velocity  $\mathbf{q}$ , and the volume fraction  $\phi$  satisfying

$$\begin{aligned}
 (4-2) \quad & (\mu(\phi))^{-1} \boldsymbol{\sigma}^{\mathrm{d}} = \nabla \mathbf{q}, \quad \partial_t \mathbf{q} - \operatorname{div} \boldsymbol{\sigma} = \mathbf{f} \phi, \quad \operatorname{div} \mathbf{q} = 0 \quad \text{in } \Omega, \\
 & \tilde{\boldsymbol{\sigma}} = \vartheta(\phi) \nabla \phi - \phi \mathbf{q} + b(\phi) \mathbf{k}, \quad \partial_t \phi - \operatorname{div} \tilde{\boldsymbol{\sigma}} = g \quad \text{in } \Omega.
 \end{aligned}$$

This system is supplemented with the following boundary conditions:

$$(4-3) \quad \mathbf{q} = \mathbf{q}_D, \quad \phi = \phi_D \quad \text{on } \Gamma_D; \quad \boldsymbol{\sigma} \mathbf{v} = \mathbf{0}, \quad \tilde{\boldsymbol{\sigma}} \cdot \mathbf{v} = 0 \quad \text{on } \Gamma_N$$

along with the initial data  $\mathbf{q}(0) = \mathbf{q}_0$ ,  $s(0) = s_0$  in  $\Omega \times \{0\}$ . Here  $(\cdot)^{\mathrm{d}}$  denotes the deviatoric operator,  $\mathbf{k}$  is a vector pointing in the direction of gravity and  $\mathbf{f} \in \mathbf{L}^\infty(\Omega)$ ,  $\mathbf{q}_D \in \mathbf{H}^{1/2}(\Gamma_D)$ ,  $g \in L^2(\Omega)$  are given functions.

Even if problems with the ingredients mentioned above have successfully been simulated numerically by many techniques (see e.g. [Betancourt, Bürger, Ruiz-Baier, Torres, and Vega \[2014\]](#), [Khalili, Basu, Pietrzyk, and Jørgensen \[1999\]](#), [Ekama, Barnard, Günthert, Krebs, McCorquodale, Parker, and Wahlberg \[1997\]](#), and [Rao, Mondy, and Altobelli \[2007\]](#)), the study of mathematical properties of (4-1) and the rigorous analysis of discretizations is still an open problem in the general case. The parabolic regularization approach has been exploited in [Bürger, Liu, and Wendland \[2001\]](#) to address the well-posedness of (4-1) for a large fluid viscosity. Its formulation in terms of Stokes

flow and the steady coupling to compression effects has been recently studied in [Alvarez, Gatica, and Ruiz-Baier \[2015\]](#). That contribution assumes that the nonlinear diffusivity depends on the concentration gradient, which is also done, for instance, for reacting non-Newtonian fluids [Buliček and Pustějovská \[2014\]](#). More general viscosity and diffusivity functions were analyzed in [Alvarez, Gatica, and Ruiz-Baier \[2016b\]](#), but still assuming non-degeneracy of the diffusion term. Models of sedimentation-consolidation are similar in structure to Boussinesq- and Oldroyd-type models, for which several mixed formulations have been analyzed (see [Colmenares, Gatica, and Oyarzúa \[2016\]](#), [Farhloul and Zine \[2011\]](#), and [Cox, Lee, and Szurley \[2007\]](#) and references cited in these papers). Augmentation of the formulation, as done in [Alvarez, Gatica, and Ruiz-Baier \[2015, 2016b\]](#), simplifies the analysis of continuous and discrete problems associated to (4.2)–(4.3).

**4.2 Finite volume element schemes.** The dominance of convection in the diffusive transport equation in (4.1) suggests the use of finite volume (FV)-based discretizations. In turn, finite element (FE) formulations are more suitable for error analysis by energy arguments and for setting up mixed formulations. Finite-volume-element (FVE) schemes retain properties of both FV and FE methods. Their construction hinges on defining fluxes across element boundaries defined on a dual partition of the domain (see [Bank and Rose \[1987\]](#) for details and [Quarteroni and Ruiz-Baier \[2011\]](#), [Kumar and Ruiz-Baier \[2015\]](#), and [Wen, He, and Yang \[2013\]](#) for recent applications in incompressible flows). Variants of FVE schemes have been employed for reactive flows ([Ewing, Lazarov, and Lin \[2000\]](#)), variable viscosity flows ([Calgare, Creusé, and Goudon \[2008\]](#)), sedimentation equations in axisymmetric form and including mild (pointwise) degeneracy ([Bürger, Ruiz-Baier, and Torres \[2012\]](#)), incorporating convective terms and using a conforming approximation in primal form ([Ruiz-Baier and Torres \[2015\]](#)), defining discontinuous discretizations for velocity-pressure and concentration ([Bürger, Kumar, and Ruiz-Baier \[2015\]](#)), also in the case of porous materials ([Bürger, Kumar, Kenettinkara, and Ruiz-Baier \[2016\]](#)).

**4.3 A posteriori error estimation.** Mesh adaptivity guided by a posteriori error estimates has a considerable potential in sedimentation-consolidation problems. Exploiting intrinsic differences in spatio-temporal scales, adaptive methods have been developed for the 1D case ([Bürger, Ruiz, Schneider, and Sepúlveda \[2008\]](#)) using multiresolution techniques, whereas the a posteriori error analysis for general coupled viscous flow-transport problems has only been addressed in [Alvarez, Gatica, and Ruiz-Baier \[2016b\]](#), [Braack and Richter \[2007\]](#), and [Larson, Söderlund, and Bengzon \[2008\]](#), and [Alvarez, Gatica, and Ruiz-Baier \[2017\]](#) in a specific application to sedimentation processes in porous media. In [Alvarez, Gatica, and Ruiz-Baier \[ibid.\]](#) efficient and reliable residual-based a posteriori error estimators for augmented mixed–primal FE schemes for stationary versions

of (4-2)–(4-3) are proposed, and a generalization to the transient case can be defined as described below.

Given an element of the FE mesh  $K \in \mathcal{T}_h$ , we denote by  $\mathcal{E}_h(K)$  the set of its edges not sharing any boundary segments, and let  $\mathcal{E}_h^{\Gamma_D}(K)$  denote the set of edges of  $K$  lying on the boundary  $\Gamma_D$ . The unit normal vector on each edge is  $\mathbf{v}_e := (v_1, v_2)^T$ , and let  $\mathbf{s}_e := (-v_2, v_1)^T$  be the corresponding fixed unit tangential vector along  $e$ . We let  $\llbracket \mathbf{v} \cdot \mathbf{v}_e \rrbracket$  be the corresponding jump across  $e$ . Then we define the approximate flux vector as  $\tilde{\boldsymbol{\sigma}}_h := \vartheta(\phi_h) \nabla \phi_h - \phi_h \mathbf{q}_h - b(\phi_h) \mathbf{k}$  and define an element-wise local error indicator associated to a semidiscretization of (4-2)–(4-3) as follows:

$$\begin{aligned} \theta_K^2 := & \|\mathbf{f} \phi_h - (\partial_t \mathbf{q}_h - \mathbf{div} \boldsymbol{\sigma}_h)\|_{0,K}^2 + \|\nabla \mathbf{q}_h - (\mu(\phi_h))^{-1} \boldsymbol{\sigma}_h^d\|_{0,K}^2 \\ & + h_K^2 \|g - (\partial_t \phi_h - \mathbf{div} \tilde{\boldsymbol{\sigma}}_h)\|_{0,K}^2 + h_K^2 \|\mathbf{curl}((\mu(\phi_h))^{-1} \boldsymbol{\sigma}_h^d)\|_{0,K}^2 \\ & + \sum_{e \in \mathcal{E}(K)} h_e \left( \|\llbracket (\mu(\phi_h))^{-1} \boldsymbol{\sigma}_h^d \mathbf{s}_e \rrbracket\|_{0,e}^2 + \|\llbracket \tilde{\boldsymbol{\sigma}}_h \cdot \mathbf{v}_e \rrbracket\|_{0,e}^2 \right) + \sum_{e \in \mathcal{E}^{\Gamma_D}(K)} \|\mathbf{q}_D - \mathbf{q}_h\|_{0,e}^2 \\ & + \sum_{e \in \mathcal{E}^{\Gamma_N}(K)} h_e \|\tilde{\boldsymbol{\sigma}}_h \cdot \mathbf{v}_e\|_{0,e}^2 + \sum_{e \in \mathcal{E}^{\Gamma_D}(K)} h_e \left\| \frac{d\mathbf{q}_D}{ds_e} - (\mu(\phi_h))^{-1} \boldsymbol{\sigma}_h^d \mathbf{s}_e \right\|_{0,e}^2. \end{aligned}$$

A global residual error estimator can then be defined as  $\boldsymbol{\theta} := \{\sum_{K \in \mathcal{T}_h} \theta_K^2\}^{1/2}$ , which has resemblance to the first residual-based indicator proposed in Alvarez, Gatica, and Ruiz-Baier [2017], and which has been shown to be efficient and reliable.

**4.4 Numerical example.** Let us consider a zeolite suspension in a secondary clarifier unit, where domain configuration and dimensions are taken from the Eindhoven WWTP (see Figure 10), and whose geometry is precisely described in Bürger, Kumar, and Ruiz-Baier [2015]. A numerical simulation using axisymmetric discontinuous FVE schemes for primal formulations has been developed in Bürger, Kumar, and Ruiz-Baier [ibid.]. We use the model parameters of that study, but here stating the set of equations in mixed form (4-2) and employ a lowest-order mixed-primal scheme as the one proposed in Alvarez, Gatica, and Ruiz-Baier [2016b]. A backward Euler method is used for the time discretization setting a fixed timestep of  $\Delta t = 5$  s and the system is evolved until  $t_{\text{final}} = 12000$  s. The device features a feed inlet  $\Gamma_{\text{in}}$  and a peripheral overflow annular region  $\Gamma_{\text{ofl}}$ . A suspension is injected through  $\Gamma_{\text{in}}$  with constant velocity  $\mathbf{q}_{\text{in}} = (0, 0.17)^T$  and having a concentration of  $\phi = 0.08$ . On  $\Gamma_{\text{out}}$  we set  $\mathbf{q}_{\text{out}} = (0, -1.5e^6)^T$  and on  $\Gamma_{\text{ofl}}$  we impose zero normal Cauchy stresses; and on the remainder of  $\partial\Omega$  we prescribe  $\mathbf{q} = \mathbf{0}$  and no-flux conditions for  $\phi$ .

The remaining parameters are chosen as  $\sigma'_c(\phi) = (\sigma_0 \alpha / \phi_c^\alpha) \phi^{\alpha-1}$ ,  $\sigma_0 = 0.22$  Pa,  $\alpha = 5$ ,  $\beta = 2.5$ ,  $\rho_L = 998.2$  kg/m<sup>3</sup>,  $\rho_X = 1750$  kg/m<sup>3</sup>,  $\phi_c = 0.014$ ,  $\tilde{\phi}_{\text{max}} = 0.95$ ,  $v_\infty =$

$0.0028935 \text{ m/s}$ ,  $g = 9.8 \text{ m/s}^2$ , and  $D_0 = 0.0028935 \text{ m}^2/\text{s}$ . The physical bounds for the concentration imply that the stabilisation parameters needed for the augmented mixed-primal FE method take the values  $\kappa_1 = 0.256$  and  $\kappa_2 = 0.25$ .

We implement an adaptive mesh refinement strategy according to the a posteriori error indicator  $\theta$ , which we invoke at the end of each time step. The marking-refining algorithm is based on the equi-distribution of the error indicators in such a way that the diameter of each new element (contained in a generic element  $K$  on the initial coarse mesh) is proportional to the initial diameter times the ratio  $\bar{\theta}_h/\eta_K$ , where  $\bar{\theta}_h$  is the mean value of  $\theta$  over the initial mesh (Verfürth [1996]). On each time step we then solve the coupled set of nonlinear equations using a fixed point method, stopping the Picard iterations when a residual tolerance of  $1\text{e-}6$  is attained. Inside each fixed-point step we solve the discretized mixed Stokes equations with a preconditioned BiCGStab method, and a nested Newton solver is employed for the nonlinear transport equation using the same value for the residual tolerance as stopping criterion and the same solver for the corresponding linear systems.

Figure 11 (top rows) presents snapshots of the numerically computed concentration profiles on a surface line integration visualization of the velocity field. We observe velocity patterns avoiding the skirt baffle and the accumulation of sediment on the bottom of the tank. The sequence of refined meshes indicates that the a posteriori error estimator identifies the zones of high concentration gradients and marked flow features. A cluster of elements is formed near these particular zones.

**4.5 Ongoing extensions.** The theory exposed above still does not cover the analysis of flow coupled to degenerate elliptic or parabolic equations, that is when the diffusivity vanishes for all concentrations below a critical value  $\phi_c$ , invalidating the fundamental assumptions of strong ellipticity and monotonicity that permits the derivation of solvability and stability of continuous and discrete problems. Then the classical tools employed in the continuous analysis as well as in the construction and analysis of the associated numerical method (Alvarez, Gatica, and Ruiz-Baier [2015, 2016b,a], Bürger, Kumar, and Ruiz-Baier [2015], Bürger, Kumar, Kenettinkara, and Ruiz-Baier [2016], Bürger, Ruiz-Baier, and Torres [2012], and Ruiz-Baier and Torres [2015]), need to be extended. Part of such a theoretical formalism has been around for many years in the context of hyperbolic conservation laws (cf. Andreianov, Karlsen, and Risebro [2011] and Berres, Bürger, Karlsen, and Tory [2003] and the references therein), but has not yet been exploited in multidimensional models of sedimentation. These developments will need to encompass entropy solutions, low-regularity finite element discretizations, discontinuous FVE, and non-conforming methods. It is also left to investigate the performance of a posteriori error indicators developed for FVE schemes applied to (4-1), where sample preliminary studies include the case of convection-reaction-diffusion (Lazarov and Tomov [2002]).



## References

- Mario Alvarez, Gabriel N. Gatica, and Ricardo Ruiz-Baier (2015). “An augmented mixed-primal finite element method for a coupled flow-transport problem”. *ESAIM Math. Model. Numer. Anal.* 49.5, pp. 1399–1427. MR: [3423229](#) (cit. on pp. [3523](#), [3525](#)).
- (2016a). “A mixed-primal finite element approximation of a sedimentation-consolidation system”. *Math. Models Methods Appl. Sci.* 26.5, pp. 867–900. MR: [3464424](#) (cit. on p. [3525](#)).
  - (2016b). “A posteriori error analysis for a viscous flow-transport problem”. *ESAIM Math. Model. Numer. Anal.* 50.6, pp. 1789–1816. MR: [3580122](#) (cit. on pp. [3523](#)–[3525](#)).
  - (2017). “A posteriori error estimation for an augmented mixed-primal method applied to sedimentation-consolidation systems”. *CI<sup>2</sup>MA preprint* (cit. on pp. [3523](#), [3524](#)).
- Boris Andreianov, Kenneth Hvistendahl Karlsen, and Nils Henrik Risebro (2011). “A theory of  $L^1$ -dissipative solvers for scalar conservation laws with discontinuous flux”. *Arch. Ration. Mech. Anal.* 201.1, pp. 27–86. MR: [2807133](#) (cit. on p. [3525](#)).
- Georg Anestis (1981). “Eine eindimensionale Theorie der Sedimentation in Absetzbehältern veränderlichen Querschnitts und in Zentrifugen”. PhD thesis. TU Vienna, Austria (cit. on pp. [3512](#)–[3514](#)).
- Randolph E. Bank and Donald J. Rose (1987). “Some error estimates for the box method”. *SIAM J. Numer. Anal.* 24.4, pp. 777–787. MR: [899703](#) (cit. on p. [3523](#)).
- Stefan Berres, Raimund Bürger, Kenneth H. Karlsen, and Elmer M. Tory (2003). “Strongly degenerate parabolic-hyperbolic systems modeling polydisperse sedimentation with compression”. *SIAM J. Appl. Math.* 64.1, pp. 41–80. MR: [2029124](#) (cit. on p. [3525](#)).
- Fernando Betancourt, Raimund Bürger, Ricardo Ruiz-Baier, Héctor Torres, and Carlos A. Vega (2014). “On numerical methods for hyperbolic conservation laws and related equations modelling sedimentation of solid-liquid suspensions”. In: *Hyperbolic conservation laws and related analysis with applications*. Vol. 49. Springer Proc. Math. Stat. Springer, Heidelberg, pp. 23–68. MR: [3111126](#) (cit. on pp. [3509](#), [3522](#)).
- Malte Braack and Thomas Richter (2007). “Solving multidimensional reactive flow problems with adaptive finite elements”. In: *Reactive flows, diffusion and transport*. Springer, Berlin, pp. 93–112. MR: [2275759](#) (cit. on p. [3523](#)).
- Miroslav Bulíček and Petra Pustějovská (2014). “Existence analysis for a model describing flow of an incompressible chemically reacting non-Newtonian fluid”. *SIAM J. Math. Anal.* 46.5, pp. 3223–3240. MR: [3262601](#) (cit. on p. [3523](#)).
- Raimund Bürger, Julio Careaga, and Stefan Diehl (n.d.). “Flux identification of scalar conservation laws from sedimentation in a cone”. To appear in *IMA J. Appl. Math.* (cit. on pp. [3514](#), [3516](#)–[3518](#)).

- (2017). “Entropy solutions of a scalar conservation law modeling sedimentation in vessels with varying cross-sectional area”. *SIAM J. Appl. Math.* 77.2, pp. 789–811. MR: [3640636](#) (cit. on pp. [3512](#), [3514–3516](#)).
- Raimund Bürger, Julio Careaga, Stefan Diehl, Camilo Mejías, Ingmar Nopens, Elena Torfs, and Peter A Vanrolleghem (2016). “Simulations of reactive settling of activated sludge with a reduced biokinetic model”. *Computers & Chemical Engineering* 92, pp. 216–229 (cit. on pp. [3510](#), [3519](#), [3534](#)).
- Raimund Bürger, Julio Careaga, Stefan Diehl, Ryan Merckel, and Jesús Zambrano (n.d.). *Estimating the hindered-settling flux function from a batch test in a cone*. Submitted (cit. on pp. [3512](#), [3518](#)).
- Raimund Bürger and Stefan Diehl (2013). “Convexity-preserving flux identification for scalar conservation laws modelling sedimentation”. *Inverse Problems* 29.4, pp. 045008, 30. MR: [3042084](#) (cit. on pp. [3509](#), [3517](#), [3518](#)).
- Raimund Bürger, Stefan Diehl, and Camilo Mejías (n.d.). *A difference scheme for a de-generating convection-diffusion-reaction system modelling continuous sedimentation* (cit. on pp. [3509–3511](#), [3519](#), [3520](#)).
- Raimund Bürger, Stefan Diehl, and Ingmar Nopens (2011). “A consistent modelling methodology for secondary settling tanks in wastewater treatment”. *Water Research* 45.6, pp. 2247–2260 (cit. on p. [3511](#)).
- Raimund Bürger, Kenneth H. Karlsen, and John D. Towers (2005). “A model of continuous sedimentation of flocculated suspensions in clarifier-thickener units”. *SIAM J. Appl. Math.* 65.3, pp. 882–940. MR: [2136036](#) (cit. on pp. [3508](#), [3510](#), [3519](#), [3520](#)).
- Raimund Bürger, Kenneth H. Karlsen, Nils Henrik Risebro, and John D. Towers (2004). “Well-posedness in  $BV_t$  and convergence of a difference scheme for continuous sedimentation in ideal clarifier-thickener units”. *Numer. Math.* 97.1, pp. 25–65. MR: [2045458](#) (cit. on p. [3520](#)).
- Raimund Bürger, Sarvesh Kumar, Sudarshan Kumar Kenettinkara, and Ricardo Ruiz-Baier (2016). “Discontinuous approximation of viscous two-phase flow in heterogeneous porous media”. *J. Comput. Phys.* 321, pp. 126–150. MR: [3527561](#) (cit. on pp. [3523](#), [3525](#)).
- Raimund Bürger, Sarvesh Kumar, and Ricardo Ruiz-Baier (2015). “Discontinuous finite volume element discretization for coupled flow-transport problems arising in models of sedimentation”. *J. Comput. Phys.* 299, pp. 446–471. MR: [3384736](#) (cit. on pp. [3523–3525](#)).
- Raimund Bürger, Chun Liu, and Wolfgang L. Wendland (2001). “Existence and stability for mathematical models of sedimentation-consolidation processes in several space dimensions”. *J. Math. Anal. Appl.* 264.2, pp. 288–310. MR: [1876734](#) (cit. on p. [3522](#)).

- Raimund Bürger, Ricardo Ruiz-Baier, and Héctor Torres (2012). “A stabilized finite volume element formulation for sedimentation-consolidation processes”. *SIAM J. Sci. Comput.* 34.3, B265–B289. MR: [2970279](#) (cit. on pp. [3523](#), [3525](#)).
- Raimund Bürger, Ricardo Ruiz, Kai Schneider, and Mauricio Sepúlveda (2008). “Fully adaptive multiresolution schemes for strongly degenerate parabolic equations in one space dimension”. *M2AN Math. Model. Numer. Anal.* 42.4, pp. 535–563. MR: [2437773](#) (cit. on p. [3523](#)).
- Raimund Bürger, Wolfgang L. Wendland, and Fernando Concha (2000). “Model equations for gravitational sedimentation-consolidation processes”. *ZAMM Z. Angew. Math. Mech.* 80.2, pp. 79–92. MR: [1742180](#) (cit. on p. [3522](#)).
- Caterina Calgari, Emmanuel Creusé, and Thierry Goudon (2008). “An hybrid finite volume-finite element method for variable density incompressible flows”. *J. Comput. Phys.* 227.9, pp. 4671–4696. MR: [2406553](#) (cit. on p. [3523](#)).
- Eligio Colmenares, Gabriel N. Gatica, and Ricardo Oyarzúa (2016). “Analysis of an augmented mixed-primal formulation for the stationary Boussinesq problem”. *Numer. Methods Partial Differential Equations* 32.2, pp. 445–478. MR: [3454217](#) (cit. on p. [3523](#)).
- Christopher Cox, Hyesuk Lee, and David Szurley (2007). “Finite element approximation of the non-isothermal Stokes-Oldroyd equations”. *Int. J. Numer. Anal. Model.* 4.3-4, pp. 425–440. MR: [2344050](#) (cit. on p. [3523](#)).
- Stefan Diehl (1997). “Continuous sedimentation of multi-component particles”. *Math. Methods Appl. Sci.* 20.15, pp. 1345–1364. MR: [1474212](#) (cit. on p. [3519](#)).
- (2007). “Estimation of the batch-settling flux function for an ideal suspension from only two experiments”. *Chemical Engineering Science* 62.17, pp. 4589–4601 (cit. on pp. [3517](#), [3518](#)).
  - (2012). “Shock-wave behaviour of sedimentation in wastewater treatment: a rich problem”. In: *Analysis for science, engineering and beyond*. Vol. 6. Springer Proc. Math. Springer, Heidelberg, pp. 175–214. MR: [3288029](#) (cit. on p. [3509](#)).
  - (2015). “Numerical identification of constitutive functions in scalar nonlinear convection-diffusion equations with application to batch sedimentation”. *Appl. Numer. Math.* 95, pp. 154–172. MR: [3349692](#) (cit. on p. [3510](#)).
- George A. Ekama, James L. Barnard, F. Wolfgang Günthert, Peter Krebs, J. Alex McCorquodale, Denny S. Parker, and Eric J. Wahlberg (1997). *Secondary Settling Tanks-Theory, Modeling, Design and Operation*. Tech. rep. International Association on Water Quality, London (cit. on p. [3522](#)).
- Richard Ewing, Raytcho Lazarov, and Yanping Lin (2000). “Finite volume element approximations of nonlocal reactive flows in porous media”. *Numer. Methods Partial Differential Equations* 16.3, pp. 285–311. MR: [1752414](#) (cit. on p. [3523](#)).

- M. Farhloul and A. Zine (2011). “[A dual mixed formulation for non-isothermal Oldroyd-Stokes problem](#)”. *Math. Model. Nat. Phenom.* 6.5, pp. 130–156. MR: [2825226](#) (cit. on p. [3523](#)).
- Mogens Henze, C.P. Leslie Grady, Willi Gujer, Gerrit v.R. Marais, and Tomonori Matsuo (1987). Tech. rep. International Association on Water Quality, London (cit. on pp. [3520](#), [3534](#)).
- Helge Holden and Nils Henrik Risebro (2015). *Front tracking for hyperbolic conservation laws*. Second. Vol. 152. Applied Mathematical Sciences. Springer, Heidelberg, pp. xiv+515. MR: [3443431](#) (cit. on pp. [3513](#), [3514](#)).
- K. H. Karlsen, N. H. Risebro, and J. D. Towers (2002). “[Upwind difference approximations for degenerate parabolic convection-diffusion equations with a discontinuous coefficient](#)”. *IMA J. Numer. Anal.* 22.4, pp. 623–664. MR: [1937244](#) (cit. on p. [3520](#)).
- Arzhang Khalili, A.J. Basu, Uwe Pietrzyk, and Bo Barker Jørgensen (1999). “Advective transport through permeable sediments: a new numerical and experimental approach”. *Acta Mechanica* 132.1-4, pp. 221–227 (cit. on p. [3522](#)).
- Sarvesh Kumar and Ricardo Ruiz-Baier (2015). “[Equal order discontinuous finite volume element methods for the Stokes problem](#)”. *J. Sci. Comput.* 65.3, pp. 956–978. MR: [3417268](#) (cit. on p. [3523](#)).
- Matthias Kunik (1993). “[A solution formula for a nonconvex scalar hyperbolic conservation law with monotone initial data](#)”. *Math. Methods Appl. Sci.* 16.12, pp. 895–902. MR: [1247889](#) (cit. on pp. [3517](#), [3518](#)).
- George J Kynch (1952). “A theory of sedimentation”. *Transactions of the Faraday society* 48, pp. 166–176 (cit. on pp. [3508](#), [3509](#), [3517](#)).
- Mats G. Larson, Robert Söderlund, and Fredrik Bengzon (2008). “[Adaptive finite element approximation of coupled flow and transport problems with applications in heat transfer](#)”. *Internat. J. Numer. Methods Fluids* 57.9, pp. 1397–1420. MR: [2435098](#) (cit. on p. [3523](#)).
- Raytcho Lazarov and Stanimire Tomov (2002). “[A posteriori error estimates for finite volume element approximations of convection-diffusion-reaction equations](#)”. *Comput. Geosci.* 6.3-4. Locally conservative numerical methods for flow in porous media, pp. 483–503. MR: [1956027](#) (cit. on p. [3525](#)).
- Alfio Quarteroni and Ricardo Ruiz-Baier (2011). “[Analysis of a finite volume element method for the Stokes problem](#)”. *Numer. Math.* 118.4, pp. 737–764. MR: [2822498](#) (cit. on p. [3523](#)).
- Rekha R Rao, Lisa A Mondy, and Stephen A Altobelli (2007). “Instabilities during batch sedimentation in geometries containing obstacles: A numerical and experimental study”. *International Journal for Numerical Methods in Fluids* 55.8, pp. 723–735 (cit. on p. [3522](#)).

- Ricardo Ruiz-Baier and Ivan Lunati (2016). “Mixed finite element–discontinuous finite volume element discretization of a general class of multicontinuum models”. *J. Comput. Phys.* 322, pp. 666–688. MR: [3534882](#) (cit. on p. [3522](#)).
- Ricardo Ruiz-Baier and Héctor Torres (2015). “Numerical solution of a multidimensional sedimentation problem using finite volume–element methods”. *Appl. Numer. Math.* 95, pp. 280–291. MR: [3349700](#) (cit. on pp. [3523](#), [3525](#)).
- Rüdiger Verfürth (1996). *A review of a posteriori error estimation and adaptive mesh-refinement techniques*. Wiley-Teubner (Chichester) (cit. on p. [3525](#)).
- Juan Wen, Yinnian He, and Jianhong Yang (2013). “Multiscale enrichment of a finite volume element method for the stationary Navier-Stokes problem”. *Int. J. Comput. Math.* 90.9, pp. 1938–1957. MR: [3171872](#) (cit. on p. [3523](#)).

Received 2017-11-30.

RAIMUND BÜRGER  
CI<sup>2</sup>MA AND DEPARTAMENTO DE INGENIERÍA MATEMÁTICA  
UNIVERSIDAD DE CONCEPCIÓN  
CASILLA 160-C, CONCEPCIÓN  
CHILE  
[rburger@ing-mat.udec.cl](mailto:rburger@ing-mat.udec.cl)

JULIO CAREAGA  
CENTRE FOR MATHEMATICAL SCIENCES  
LUND UNIVERSITY  
P.O. Box 118  
S-221 00 LUND  
SWEDEN  
[julio.careaga@math.lth.se](mailto:julio.careaga@math.lth.se)

STEFAN DIEHL  
CENTRE FOR MATHEMATICAL SCIENCES  
LUND UNIVERSITY  
P.O. Box 118  
S-221 00 LUND  
SWEDEN  
[stefan.diehl@math.lth.se](mailto:stefan.diehl@math.lth.se)

CAMILO MEJÍAS  
CI<sup>2</sup>MA AND DEPARTAMENTO DE INGENIERÍA MATEMÁTICA  
UNIVERSIDAD DE CONCEPCIÓN  
CASILLA 160-C, CONCEPCIÓN  
CHILE  
[cmejias@ing-mat.udec.cl](mailto:cmejias@ing-mat.udec.cl)

RICARDO RUIZ BAIER

MATHEMATICAL INSTITUTE  
OXFORD UNIVERSITY  
ANDREW WILES BUILDING  
WOODSTOCK ROAD  
OX2 6GG OXFORD  
UK  
[ruizbaier@maths.ox.ac.uk](mailto:ruizbaier@maths.ox.ac.uk)

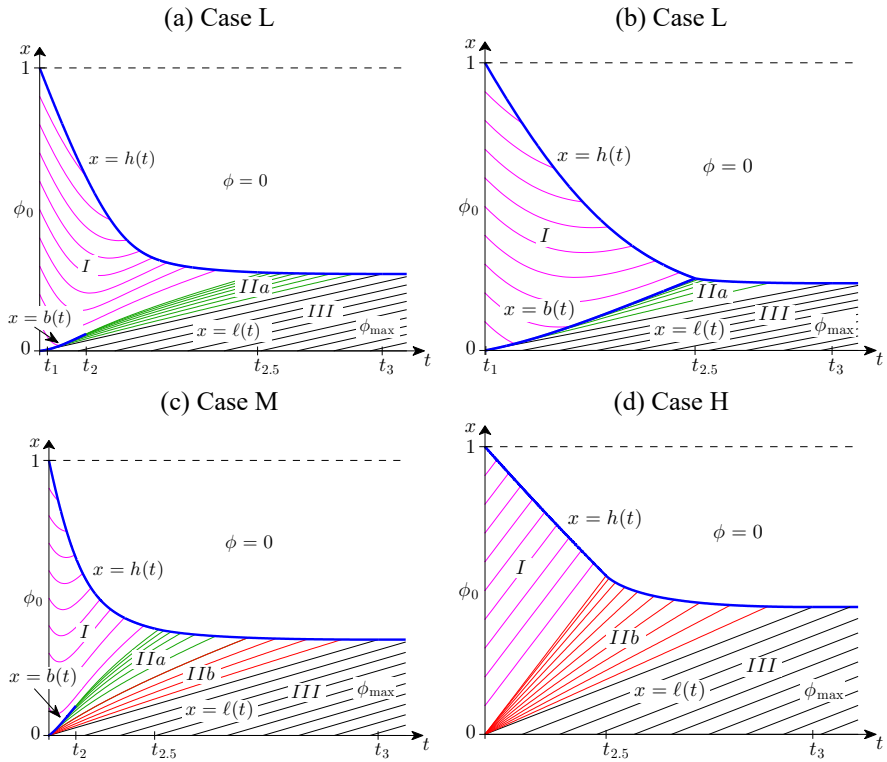


Figure 5: Solutions of (2-1) in a truncated cone ( $q = 1/2$ ,  $p > 0$ ) with  $f$  given by (1-6): (a) Case L,  $r_V = 4$ ,  $\phi_0 = 0.04$ ,  $p = 1/18$ ; (b) Case L,  $r_V = 4$ ,  $\phi_0 = 0.1$ ,  $p = 1/3$ ; (c) Case M,  $r_V = 5$ ,  $\phi_0 = 0.12$ ,  $p = 1/6$ ; (d) Case H,  $r_V = 4.7$ ,  $\phi_0 = 0.43$ ,  $p = 9.5$ . The solid blue curves are discontinuities.

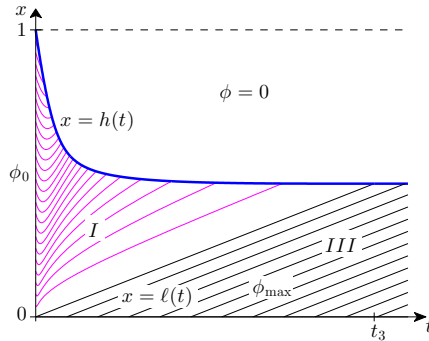


Figure 6: Solution corresponding to item (i) of [Theorem 2.2](#).

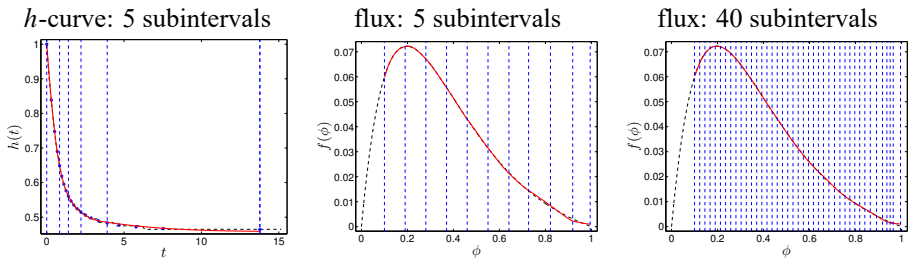


Figure 7: Flux identification via settling in a cone with  $\phi_0 = 0.1$  from synthetic data of the discontinuity  $x = h(t)$ . The number of subintervals is that of cubic polynomials used for the  $h$ -curve. The true flux is shown in dashed and the identified fluxes in solid red.

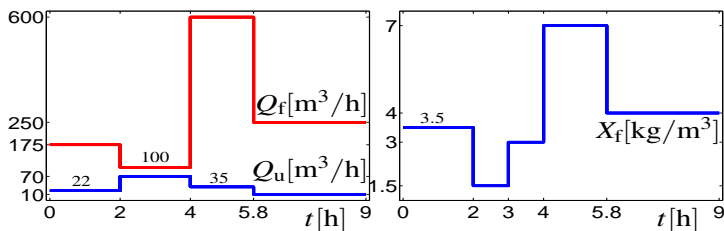


Figure 8: Piecewise constant functions  $Q_f$  and  $Q_u$  (feed and underflow volume rates) and  $X_f$  (solids feed concentration) for the numerical example of reactive settling ([Figure 9](#)).



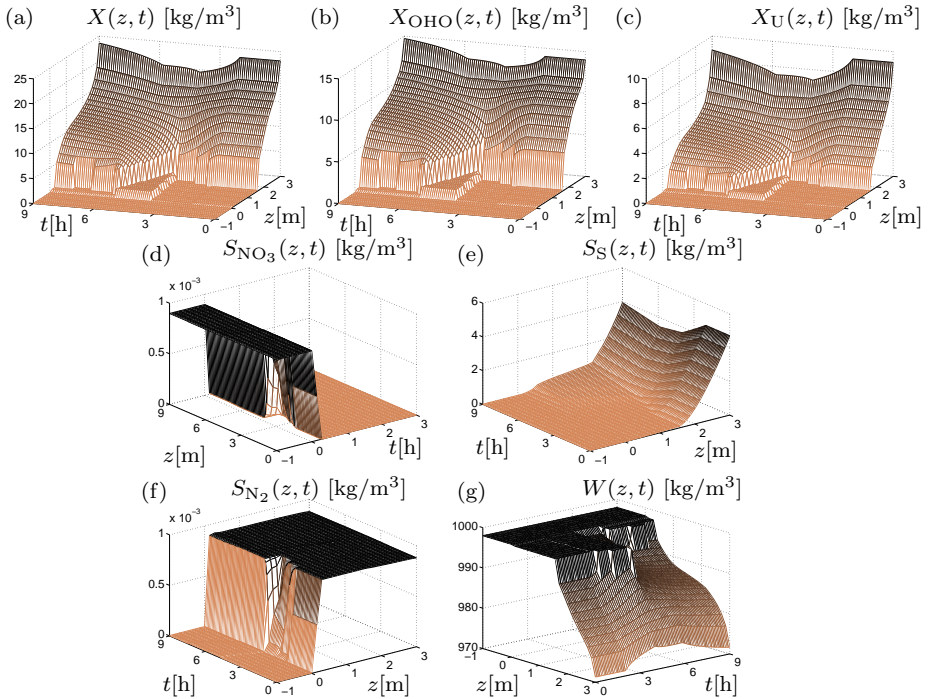


Figure 9: Simulation of reactive settling (denitrification) in an SST under variations of  $Q_u$ ,  $Q_f$  and  $X_f$  (see Figure 8). Constants are standard in ASM1 (Henze, Grady, Gujer, Marais, and Matsuo [1987]) or arise in a strongly reduced model (Bürger, Careaga, Diehl, Mejias, Nopens, Torfs, and Vanrolleghem [2016]):  $b = 6.94 \times 10^{-6} \text{ s}^{-1}$ ,  $f_p = 0.2$ ,  $K_{\text{NO}_3} = 5.0 \times 10^{-4} \text{ kg m}^{-3}$ ,  $X_{\text{max}} = 30 \text{ kg m}^{-3}$ , (the maximum solids concentration),  $\mu_{\text{max}} = 5.56 \times 10^{-5} \text{ s}^{-1}$ ,  $v_0 = 1.76 \times 10^{-3} \text{ m s}^{-1}$ ,  $\rho_X = 1050 \text{ kg m}^{-3}$ ,  $\rho_L = 998 \text{ kg m}^{-3}$ ,  $g = 9.8 \text{ m s}^{-2}$  (acceleration of gravity) and  $Y = 0.67$  (yield factor).

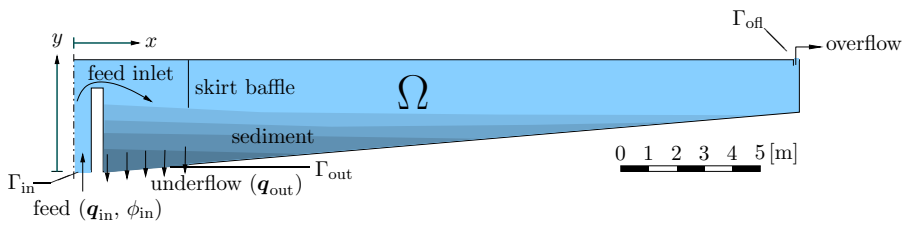


Figure 10: Settling tank from Eindhoven WWTP.

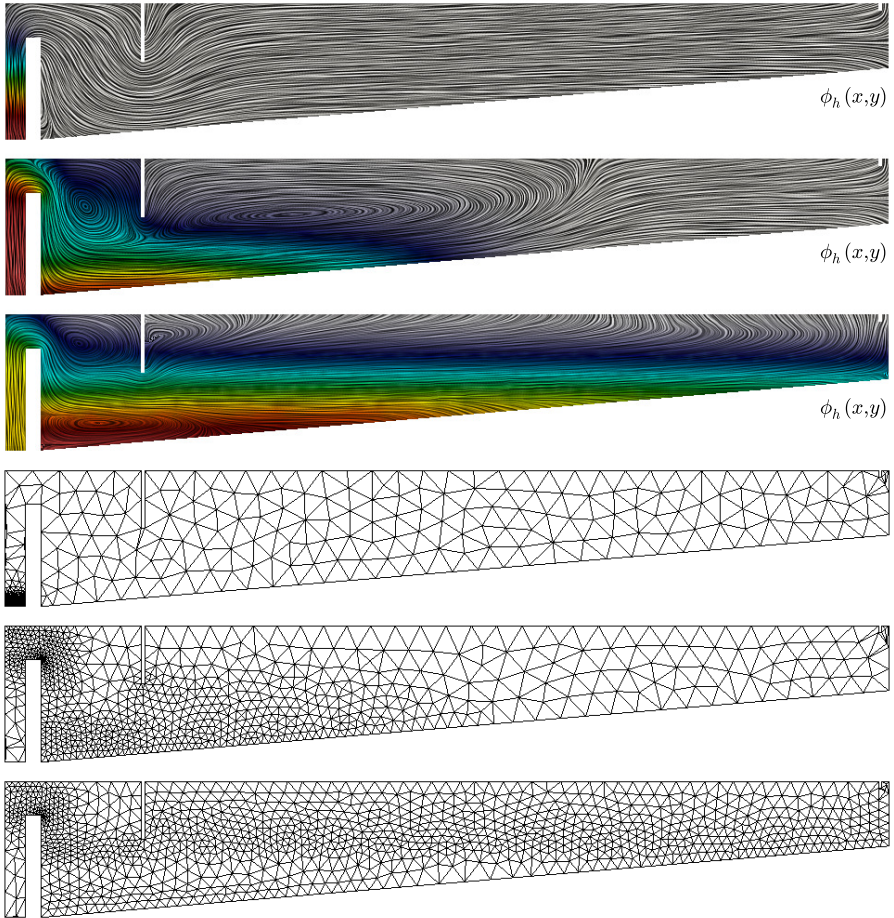


Figure 11: Example: Mixed-primal FE approximation of  $\phi$  at  $t = 200$  s,  $t = 4000$  s, and  $t = 12000$  s, and corresponding adapted meshes refined using the a posteriori error estimator  $\theta$ .

# A REVIEW ON HIGH ORDER WELL-BALANCED PATH-CONSERVATIVE FINITE VOLUME SCHEMES FOR GEOPHYSICAL FLOWS

MANUEL J. CASTRO, MARC DE LA ASUNCIÓN, ENRIQUE D. FERNÁNDEZ NIETO,  
JOSÉ M. GALLARDO, JOSÉ M. GONZÁLEZ VIDA, JORGE MACÍAS,  
TOMÁS MORALES, SERGIO ORTEGA AND CARLOS PARÉS

## Abstract

In this work a general strategy to design high order well-balanced schemes for hyperbolic system with nonconservative products and/or source terms is reviewed. We briefly recall the theory of Dal Maso-LeFloch-Murat to define weak solutions of nonconservative systems and how it has been used to establish the notion of path-conservative schemes. Next, a family of high order finite volume methods combining a reconstruction operator and a first order path-conservative scheme is described. Then, the well-balanced property of the proposed methods is analyzed. Finally, some challenging examples on tsunami modeling are shown.

## 1 Introduction

In the last few years, numerous publications have been devoted to the development of high order finite volume schemes for PDE systems of the form

$$(1) \quad \frac{\partial U}{\partial t} + \frac{\partial F_1}{\partial x}(U) + \frac{\partial F_2}{\partial y}(U) + B_1(U) \frac{\partial U}{\partial x} + B_2(U) \frac{\partial U}{\partial y} = S_1(U) \frac{\partial H}{\partial x} + S_2(U) \frac{\partial H}{\partial y},$$

where the unknown  $U(\mathbf{x}, t)$  is defined in  $D \times (0, T)$ ,  $D$  being a domain of  $\mathbb{R}^2$ , and takes values on an open convex subset  $\Omega$  of  $\mathbb{R}^N$ ;  $F_i$ ,  $i = 1, 2$  are two regular functions from

---

Manuel J. Castro wants to thank all collaborators such as M.L. Muñoz Ruiz, J. A. López, A. Pardo, C. Sánchez, C. Escalante and E. Guerrero (Univ. Málaga), J. M. Mantas (Univ. Granada), T. Chacón and G. Narbora (Univ. Sevilla), A. Marquina (Univ. Valencia), J. A. García and A. Ferreiro (Univ. Coruña), E. F. Toro, M. Dumbser and E. Gaburro (Univ. Trento), G. Russo (Univ. Catania), among others, with whom I have worked in recent years.

*MSC2010:* primary 65M08; secondary 65Y05, 65Y10, 86A05.

*Keywords:* high order finite volume methods, well-balanced methods, path-conservative schemes, FVM Riemann solvers, geophysical flows.

$\Omega$  to  $\mathbb{R}^N$ ;  $B_i$ ,  $i = 1, 2$  are two regular matrix-valued function from  $\Omega$  to  $\mathfrak{M}_{N \times N}(\mathbb{R})$ ;  $S_i$ ,  $i = 1, 2$  are two functions from  $\Omega$  to  $\mathbb{R}^N$ ; and finally  $H(\mathbf{x})$  is a known function from  $D$  to  $\mathbb{R}$ . See for example [Castro, Fernández-Nieto, Ferreiro, García-Rodríguez, and Parés \[2009\]](#), [Castro, Gallardo, and Parés \[2006\]](#), [Castro, Gallardo, López-García, and Parés \[2008a\]](#), [Dumbser, Enaux, and E. F. Toro \[2008\]](#), [Gallardo, Ortega, de la Asunción, and Mantas \[2011\]](#), [Gallardo, Parés, and Castro \[2007\]](#), [Lukáčová-Medvid'ová, Noelle, and Kraft \[2007\]](#), [Noelle, Pankratz, Puppo, and Natvig \[2006\]](#), [Noelle, Xing, and Shu \[2007\]](#), [Russo and Khe \[2009\]](#), [Xing and Shu \[2006\]](#), [Xing \[2017\]](#) among others, and [Castro, Morales de Luna, and Parés \[2017\]](#) for a review.

System (1) includes as particular cases: systems of conservation laws ( $B_i = 0$ ,  $S_i = 0$ ,  $i = 1, 2$ ); systems of conservation laws with source term or balance laws ( $B_i = 0$ ,  $i = 1, 2$ ); and coupled systems of conservation laws.

In particular, many interesting problem related to geophysical flows may be written in the form (1): shallow water systems (one layer or multi-layer systems) that govern the flow of homogeneous stratified fluid (see [E. F. Toro \[2001\]](#), [Audusse, Bristeau, Perthame, and Sainte-Marie \[2011\]](#), [Fernández-Nieto, Koné, and Chacón Rebollo \[2014\]](#)), Shallow-water Exner systems that are commonly used to model the evolution of a sediment layer submerged on a shallow-flow (see [Exner \[1925\]](#), [Grass \[1981\]](#), [Castro Díaz, Fernández-Nieto, and Ferreiro \[2008\]](#), [Fernández-Nieto, Lucas, Morales de Luna, and Cordier \[2014\]](#)), turbidity current models useful to simulate the hyperpycnal plume that is created when a river with a high concentration of suspended sediment flows into the sea (see [Bradford and Katopodes \[1999\]](#), [Morales de Luna, Castro Díaz, Parés Madroñal, and Fernández Nieto \[2009\]](#), [Morales de Luna, Fernández Nieto, and Castro Díaz \[2017\]](#), Ripa model or two-mode shallow-water system for modeling ocean currents (see [Ripa \[1993\]](#), [Khouider, Majda, and Stechmann \[2008\]](#), and [Castro Díaz, Cheng, Chertock, and Kurganov \[2014\]](#)). Systems with similar characteristics also appear in other fluid models as two-phase flows.

In this paper, we summarize our main contributions to the design of high-order and well-balanced finite volume solvers for system (1). Notice first that (1) can be rewritten in the form

$$(2) \quad W_t + \mathcal{Q}_1(W)W_x + \mathcal{Q}_2(W)W_y = 0,$$

by considering  $W = [U, H]^T$  and

$$\mathcal{Q}_i(W) = \begin{pmatrix} A_i(U) & -S_i(U) \\ 0 & 0 \end{pmatrix}, \quad i = 1, 2,$$

being  $A_i(U) = J_i(U) + B_i(U)$  where  $J_i(U) = \frac{\partial F_i}{\partial U}(U)$ ,  $i = 1, 2$  denote the Jacobians of  $F_i$ ,  $i = 1, 2$ . We also assume that (2) is strictly hyperbolic, i.e. for all  $W \in \tilde{\Omega} = \Omega \times \mathbb{R}$

and  $\forall \boldsymbol{\eta} = (\eta_x, \eta_y) \in \mathbb{S}^1$ , where  $\mathbb{S}^1 \subset \mathbb{R}^2$  denotes the unit sphere, the matrix

$$\mathcal{Q}(W, \boldsymbol{\eta}) = \mathcal{Q}_1(W)\eta_x + \mathcal{Q}_2(W)\eta_y$$

has  $M = N + 1$  real and distinct eigenvalues

$$\lambda_1(W, \boldsymbol{\eta}) < \cdots < \lambda_M(W, \boldsymbol{\eta})$$

and  $\mathcal{Q}(W, \boldsymbol{\eta})$  is thus diagonalizable.

The nonconservative products  $\mathcal{Q}_1(W)W_x$  and  $\mathcal{Q}_2(W)W_y$  do not make sense as distributions if  $W$  is discontinuous. However, the theory developed by Dal Maso, LeFloch and Murat in [Dal Maso, Lefloch, and Murat \[1995\]](#) allows to give a rigorous definition of nonconservative products as bounded measures provided that a family of Lipschitz continuous paths  $\Phi: [0, 1] \times \tilde{\Omega} \times \tilde{\Omega} \times \mathbb{S}^1 \rightarrow \tilde{\Omega}$  is prescribed. This family must satisfy certain natural regularity conditions, in particular:

1.  $\Phi(0; W_L, W_R, \boldsymbol{\eta}) = W_L$  and  $\Phi(1; W_L, W_R, \boldsymbol{\eta}) = W_R$ , for any  $W_L, W_R \in \tilde{\Omega}$ ,  $\boldsymbol{\eta} \in \mathbb{S}^1$ .
2.  $\Phi(s; W_L, W_R, \boldsymbol{\eta}) = \Phi(1-s; W_R, W_L, -\boldsymbol{\eta})$ , for any  $W_L, W_R \in \tilde{\Omega}$ ,  $s \in [0, 1]$ ,  $\boldsymbol{\eta} \in \mathbb{S}^1$ .

The choice of this family of paths should be based on the physics of the problem: for instance, it should be based on the viscous profiles corresponding to a regularized system in which some of the neglected terms (e.g. the viscous terms) are taken into account. Unfortunately, the explicit calculations of viscous profiles for a regularization of (2) is in general a difficult task. Some hints of how paths can be chosen is discussed in [Castro, Morales de Luna, and Parés \[2017\]](#). An alternative is to choose the ‘canonical’ path given by the family of segments:

$$(3) \quad \Phi(s; W_L, W_R, \boldsymbol{\eta}) = W_L + s(W_R - W_L),$$

that corresponds to the definition of nonconservative products proposed by Volpert (see [Volpert \[1967\]](#)). As shown in [Castro Díaz, Fernández-Nieto, Morales de Luna, Narbona-Reina, and Parés \[2013\]](#), this family is a sensible choice as it provides third order approximation of the correct jump conditions in the phase plane.

Suppose that a family of paths  $\Phi$  in  $\tilde{\Omega}$  has been chosen. Then a piecewise regular function  $W$  is a *weak solution* of (2) if and only if the two following conditions are satisfied:

- (i)  $W$  is a classical solution where it is smooth.
- (ii) At every point of a discontinuity  $W$  satisfies the jump condition

$$(4) \quad \int_0^1 (\sigma \mathbb{I} - \mathcal{Q}(\Phi(s; W^-, W^+, \boldsymbol{\eta}), \boldsymbol{\eta})) \frac{\partial \Phi}{\partial s}(s; W^-, W^+, \boldsymbol{\eta}) ds = 0,$$

where  $\mathbb{I}$  is the identity matrix;  $\sigma$ , the speed of propagation of the discontinuity;  $\boldsymbol{\eta}$  a unit vector normal to the discontinuity at the considered point; and  $W^-$ ,  $W^+$ , the lateral limits of the solution at the discontinuity.

As in conservative systems, together with the definition of weak solutions, a notion of entropy has to be chosen. We will assume here that the system can be endowed with an entropy pair  $(\mathcal{H}, \mathbf{G})$  i.e. a pair of regular functions  $\mathcal{H} : \tilde{\Omega} \rightarrow \mathbb{R}$ ,  $\mathcal{H}$  convex and  $\mathbf{G} = (G_1, G_2) : \Omega \rightarrow \mathbb{R}^2$  such that:

$$\nabla G_i(W) = \nabla \mathcal{H}(W) \cdot \mathcal{R}_i(W), \quad \forall W \in \Omega, \quad i = 1, 2.$$

Then, a weak solution is said to be an *entropy solution* if it satisfies the inequality

$$\partial_t \mathcal{H}(W) + \partial_x G_1(W) + \partial_y G_2(W) \leq 0,$$

in the sense of distributions.

**Acknowledgments.** Manuel J. Castro wants to thank all collaborators such as J. A. López, A. Pardo, C. Sánchez, C. Escalante and E. Guerrero (Univ. Málaga), J. M. Mantas (Univ. Granada), T. Chacón and G. Narbora (Univ. Sevilla), A. Marquina (Univ. Valencia), J. A. García and A. Ferreiro (Univ. Coruña), E. F. Toro, M. Dumbser and E. Gaburro (Univ. Trento), G. Russo (Univ. Catania), G. Puppo (Politecnico Torino), among others, with whom I have worked in recent years.

## 2 High-order finite volume schemes

To discretize (2) the computational domain  $D$  is decomposed into subsets with a simple geometry, called cells or finite volumes:  $V_i \subset \mathbb{R}^2$ . It is assumed that the cells are closed convex polygons whose intersections are either empty, a complete edge, or a vertex. Denote by  $\mathcal{T}$  the mesh, i.e., the set of cells, and by  $N_{\mathcal{T}}$  the number of cells.

Given a finite volume  $V_i$ ,  $|V_i|$  will represent its area;  $N_i \in \mathbb{R}^2$  its center;  $\mathfrak{N}_i$  the set of indexes  $j$  such that  $V_j$  is a neighbor of  $V_i$ ;  $E_{ij}$  the common edge of two neighboring cells  $V_i$  and  $V_j$ , and  $|E_{ij}|$  its length;  $\boldsymbol{\eta}_{ij} = (\eta_{ij,x}, \eta_{ij,y})$  the normal unit vector at the edge  $E_{ij}$  pointing towards the cell  $V_j$ ;  $\Delta x$  is the maximum of the diameters of the cells and  $W_i^n$  the constant approximation to the average of the solution in the cell  $V_i$  at time  $t^n$  provided by the numerical scheme:

$$W_i^n \cong \frac{1}{|V_i|} \int_{V_i} W(\mathbf{x}, t) d\mathbf{x}.$$

Following [Castro, Fernández-Nieto, Ferreiro, García-Rodríguez, and Parés \[2009\]](#), we are first going to describe a procedure to construct high order finite volume schemes for system (2). Let us, first recall the procedure for systems of conservation laws

$$(5) \quad W_t + F_1(W)_x + F_2(W)_y = 0.$$

High order methods based on the reconstruction of states can be built for (5) combining a first order conservative scheme with a consistent numerical flux function  $\mathfrak{F}(W_l, W_r, \eta)$  together with a reconstruction operator of order  $p$ . We will assume that the reconstructions are calculated as follows: given a family  $\{W_i\}_{i=1}^{N_{\mathcal{T}}}$  of cell values, first an approximation function is constructed at every cell  $V_i$ , based on the values at some of the cells close to  $V_i$ :

$$P_i(\mathbf{x}) = P_i(\mathbf{x}; \{W_j\}_{j \in \mathcal{B}_i}),$$

for some set of indexes  $\mathcal{B}_i$  (the stencil). If, for instance, the reconstruction only depends on the neighbor cells of  $V_i$ , then  $\mathcal{B}_i = \mathfrak{N}_i \cup \{i\}$ . These approximation functions are calculated usually by means of an interpolation or approximation procedure. Once these functions have been constructed, the reconstruction at  $\gamma \in E_{ij}$  are defined as follows:

$$(6) \quad W_{ij}^-(\gamma) = \lim_{\mathbf{x} \rightarrow \gamma} P_i(\mathbf{x}), \quad W_{ij}^+(\gamma) = \lim_{\mathbf{x} \rightarrow \gamma} P_j(\mathbf{x}).$$

As usual, the reconstruction operator must satisfy the following properties:

(P1) It is conservative, i.e. the following equality holds for any cell  $V_i$ :

$$(7) \quad W_i = \frac{1}{|V_i|} \int_{V_i} P_i(\mathbf{x}) d\mathbf{x}.$$

(P2) If the operator is applied to the cell averages  $\{W_i\}$  for some smooth function  $W(\mathbf{x})$ , then

$$W_{ij}^{\pm}(\gamma) = W(\gamma) + O(\Delta \mathbf{x}^p), \quad \forall \gamma \in E_{ij},$$

and

$$W_{ij}^+(\gamma) - W_{ij}^-(\gamma) = O(\Delta \mathbf{x}^{p+1}), \quad \forall \gamma \in E_{ij}.$$

In the literature one can find many examples of reconstruction operators that satisfy (P1) and (P2): ENO, WENO, CWENO, hyperbolic reconstructions, among others (see [Harten, Engquist, Osher, and Chakravarthy \[1987\]](#), [Marquina \[1994\]](#), [Shu \[1998\]](#), [Shu and Osher \[1989\]](#), [Dumbser, Balsara, E. F. Toro, and Munz \[2008\]](#), [Dumbser and Käser \[2007\]](#), and [Dumbser, Käser, Titarev, and E. F. Toro \[2007\]](#), [Gallardo, Ortega, de la Asunción, and Mantas \[2011\]](#), [Cravero and Semplice \[2016\]](#)).



Once the first order method and the reconstruction operator have been chosen, the method of lines can be used to develop high order methods for (5): the idea is to discretize only in space, what leads to a system of ODE that can be solved using TVD Runge-Kutta methods introduced in [Gottlieb and Shu \[1998\]](#) and [Shu and Osher \[1988\]](#). Other time discretization can be considered, as ADER schemes developed by Toro and Dumbser (see [Titarev and E. F. Toro \[2005\]](#) and [Dumbser, Castro, Parés, and E. F. Toro \[2009\]](#)).

Let  $\overline{W}_i(t)$  denote the cell average of a regular solution  $W$  of (5) over the cell  $V_i$  at time  $t$ :

$$\overline{W}_i(t) = \frac{1}{|V_i|} \int_{V_i} W(\mathbf{x}, t) d\mathbf{x}.$$

Integrating (5) over the cell  $V_i$ , the following equation can be easily obtained for the cell averages:

$$(8) \quad \overline{W}'_i(t) = -\frac{1}{|V_i|} \left( \sum_{j \in \mathfrak{n}_i} \int_{E_{ij}} F_{\eta_{ij}}(W(\gamma, t)) d\gamma \right)$$

where  $F_{\eta}(\cdot) = F_1(\cdot)\eta_x + F_2(\cdot)\eta_y$ . The first order method and the reconstructions are now used to approach the values of the fluxes at the edges:

$$(9) \quad W'_i(t) = -\frac{1}{|V_i|} \left( \sum_{j \in \mathfrak{n}_i} \int_{E_{ij}} \mathfrak{F}(W_{ij}^-(\gamma, t), W_{ij}^+(\gamma, t), \eta_{ij}) d\gamma \right),$$

$W_i(t)$  being the approximation to  $\overline{W}_i(t)$  provided by the scheme and  $W_{ij}^{\pm}(\gamma, t)$  the reconstruction at  $\gamma \in E_{ij}$  corresponding to the family  $\{W_i(t)\}_{i=1}^{N_T}$ . It can be shown that (9) is an approximation of order  $p$  of (8).

In practice, the integral terms in (9) are approached by means of a numerical quadrature formula of order  $\bar{r} \geq p$  at least.

Let us now generalize the semi-discrete method (9) to the nonconservative system (2). We will assume that the reconstruction operators satisfy (P1)-(P2) and the following properties:

(P3) It is of order  $q$  in the interior of the cells, i.e. if the operator is applied to a sequence  $\{W_i\}$  for some smooth function  $W(\mathbf{x})$ , then:

$$(10) \quad P_i(\mathbf{x}) = W(\mathbf{x}) + O(\Delta \mathbf{x}^q), \quad \forall \mathbf{x} \in \text{int}(V_i).$$

(P4) Under the assumption of the previous property, the gradient of  $P_i$  provides an approximation of order  $m$  of the gradient of  $W$ :

$$(11) \quad \nabla P_i(\mathbf{x}) = \nabla W(\mathbf{x}) + O(\Delta \mathbf{x}^m), \quad \forall \mathbf{x} \in \text{int}(V_i).$$

**Remark 1.** Notice that, in general,  $m \leq q \leq p$ . If, for instance, the approximation functions are polynomials of degree  $p$  obtained by interpolating the cell values on a fixed stencil, then  $m = p-1$  and  $q = p$ . In the case of WENO-like reconstructions (see [Shu and Osher \[1988\]](#)), the approximation functions are obtained as a weighted combination of interpolation polynomials whose accuracy is greater on the boundary than at the interior of the cell: in this case  $q < p$ . An interesting alternative of WENO reconstruction operator for which  $q = p$  is given by CWENO reconstruction (see [Cravero and Semplice \[2016\]](#)).

Let us denote by  $P_i^t$  the approximation functions defined using the cell averages  $W_i(t)$ , i.e.

$$P_i^t(\mathbf{x}) = P_i(\mathbf{x}; \{W_j(t)\}_{j \in \mathfrak{B}_i}).$$

$W_{ij}^-(\gamma, t)$  (resp.  $W_{ij}^+(\gamma, t)$ ) is then defined by

$$(12) \quad W_{ij}^-(\gamma, t) = \lim_{\mathbf{x} \rightarrow \gamma} P_i^t(\mathbf{x}), \quad W_{ij}^+(\gamma, t) = \lim_{\mathbf{x} \rightarrow \gamma} P_j^t(\mathbf{x}).$$

Note that (9) can be rewritten as follows using the divergence theorem:

$$(13) \quad \begin{aligned} W_i'(t) = & -\frac{1}{|V_i|} \sum_{j \in \mathfrak{N}_i} \int_{E_{ij}} \mathfrak{D}_{ij}^-(W_{ij}^-(\gamma, t), W_{ij}^+(\gamma, t), \eta_{ij}) d\gamma \\ & - \frac{1}{|V_i|} \int_{V_i} \left( J_1(P_i^t(\mathbf{x})) \frac{\partial P_i^t}{\partial x_1}(\mathbf{x}) + J_2(P_i^t(\mathbf{x})) \frac{\partial P_i^t}{\partial x_2}(\mathbf{x}) \right) d\mathbf{x}, \end{aligned}$$

where

$$(14) \quad \mathfrak{D}_{ij}^-(W_{ij}^-(\gamma, t), W_{ij}^-(\gamma, t), \eta_{ij}) = \mathfrak{F}(W_{ij}^-(\gamma, t), W_{ij}^+(\gamma, t), \eta_{ij}) - F_{\eta_{ij}}(W_{ij}^-(\gamma, t)).$$

According to Parés [Parés \[2006\]](#), (14) is a first order path-conservative scheme that is naturally defined from a standard conservative flux. This expression can be extended to nonconservative systems by replacing  $J_i$  by  $\mathfrak{Q}_i$ ,  $i = 1, 2$  and  $\mathfrak{D}_{ij}^-$  by the fluctuations of a path-conservative numerical method:

$$(15) \quad \begin{aligned} W_i'(t) = & -\frac{1}{|V_i|} \left[ \sum_{j \in \mathfrak{N}_i} \int_{E_{ij}} \mathfrak{D}_{\Phi}^-(W_{ij}^-(\gamma, t), W_{ij}^+(\gamma, t), \eta_{ij}) d\gamma \right. \\ & \left. + \int_{V_i} \left( \mathfrak{Q}_1(P_i^t(\mathbf{x})) \frac{\partial P_i^t}{\partial x}(\mathbf{x}) + \mathfrak{Q}_2(P_i^t(\mathbf{x})) \frac{\partial P_i^t}{\partial y}(\mathbf{x}) \right) d\mathbf{x} \right], \end{aligned}$$

where  $\mathfrak{D}_{\Phi}^{-}(W_{ij}^{-}(\gamma, t), W_{ij}^{-}(\gamma, t), \boldsymbol{\eta}_{ij})$  is a path-conservative scheme for system (2), that is  $\mathfrak{D}_{\Phi}^{-}(W_l, W_r, \boldsymbol{\eta})$  is a regular function from  $\tilde{\Omega} \times \tilde{\Omega} \times \mathbb{S}^1$  to  $\mathbb{R}^M$ ,  $M = N + 1$  satisfying

$$(16) \quad \mathfrak{D}_{\Phi}^{-}(W, W, \boldsymbol{\eta}) = 0 \quad \forall W \in \tilde{\Omega}, \quad \forall \boldsymbol{\eta} \in \mathbb{S}^1$$

and

$$(17) \quad \mathfrak{D}_{\Phi}^{-}(W_l, W_r, \boldsymbol{\eta}) + \mathfrak{D}_{\Phi}^{+}(W_l, W_r, \boldsymbol{\eta}) = \int_0^1 \mathfrak{Q}(\Phi(s; W_l, W_r, \boldsymbol{\eta}), \boldsymbol{\eta}) \frac{\partial \Phi}{\partial s}(s; W_l, W_r, \boldsymbol{\eta}) ds,$$

where  $\mathfrak{D}_{\Phi}^{+}(W_l, W_r, \boldsymbol{\eta}) = \mathfrak{D}_{\Phi}^{-}(W_r, W_l, -\boldsymbol{\eta})$  and  $\Phi$  is the chosen family of paths.

Note that the cell averages of a smooth solution of (2),  $\overline{W}_i(t)$ , satisfy:

$$(18) \quad \overline{W}_i'(t) = -\frac{1}{|V_i|} \int_{V_i} (\mathfrak{Q}_1(W(\mathbf{x}))W_x(\mathbf{x}) + \mathfrak{Q}_2(W(\mathbf{x}))W_y(\mathbf{x})) d\mathbf{x}.$$

Thus, (15) is expected to be an accurate approximation of (18). This fact is stated in the following result (see [Castro, Fernández-Nieto, Ferreiro, García-Rodríguez, and Parés \[2009\]](#)):

**Theorem 1.** *Let us assume that  $\mathfrak{Q}_1$  and  $\mathfrak{Q}_2$  are of class  $\mathcal{C}^2$  with bounded derivatives and  $\mathfrak{D}_{\Phi}^{-}(\cdot, \cdot, \boldsymbol{\eta}_{ij})$  is bounded for all  $i, j$ . Let us also suppose that the reconstruction operator satisfies the hypothesis (P1)-(P4). Then (15) is an approximation of order at least  $\alpha = \min(p, q, m)$  to the system (18) in the following sense:*

$$(19) \quad \begin{aligned} & \frac{1}{|V_i|} \sum_{j \in \mathbf{n}_i} \left[ \int_{E_{ij}} \left( \mathfrak{D}_{\Phi}^{-}(W_{ij}^{-}(\gamma, t), W_{ij}^{+}(\gamma, t), \boldsymbol{\eta}_{ij}) \right) d\gamma \right. \\ & \left. + \int_{V_i} \left( \mathfrak{Q}_1(P_i^t(\mathbf{x})) \frac{\partial P_i^t}{\partial x}(\mathbf{x}) + \mathfrak{Q}_2(P_i^t(\mathbf{x})) \frac{\partial P_i^t}{\partial y}(\mathbf{x}) \right) d\mathbf{x} \right] \\ & = \frac{1}{|V_i|} \sum_{j \in \mathbf{n}_i} \int_{V_i} (\mathfrak{Q}_1(W(\mathbf{x}, t))W_x(\mathbf{x}, t) + \mathfrak{Q}_2(W(\mathbf{x}, t))W_y(\mathbf{x}, t)) d\mathbf{x} + O(\Delta \mathbf{x}^{\alpha}), \end{aligned}$$

for every solution  $W$  smooth enough, being  $W_{ij}^{\pm}(\gamma, t)$  the associated reconstructions and  $P_i^t$  the approximation functions corresponding to the family

$$\overline{W}_i(t) = \frac{1}{|V_i|} \int_{V_i} W(\mathbf{x}, t) d\mathbf{x}.$$

**Remark 2.** According to [Remark 1](#) the expected order of the numerical scheme is  $m$ . Nevertheless, this theoretical result is rather pessimistic: in practice order  $q$  is often achieved.

Now, taking into account the relation between systems (2) and (1), it is possible to rewrite (15) as follows:

$$\begin{aligned}
 U'_i(t) = & -\frac{1}{|V_i|} \sum_{j \in \mathfrak{N}_i} \int_{E_{ij}} D_{\Phi}^-(U_{ij}^-(\gamma, t), U_{ij}^+(\gamma, t), H_{ij}^-(\gamma), H_{ij}^+(\gamma), \eta_{ij}) d\gamma \\
 & -\frac{1}{|V_i|} \sum_{j \in \mathfrak{N}_i} \int_{E_{ij}} F_{\eta_{ij}}(U_{ij}^-(\gamma, t)) d\gamma \\
 (20) \quad & -\frac{1}{|V_i|} \int_{V_i} B_1(P_i^{U,t}(\mathbf{x})) \frac{\partial P_i^{U,t}}{\partial x}(\mathbf{x}) + B_2(P_i^{U,t}(\mathbf{x})) \frac{\partial P_i^{U,t}}{\partial y}(\mathbf{x}) d\mathbf{x} \\
 & +\frac{1}{|V_i|} \int_{V_i} S_1(P_i^{U,t}(\mathbf{x})) \frac{\partial P_i^H}{\partial x}(\mathbf{x}) + S_2(P_i^{U,t}(\mathbf{x})) \frac{\partial P_i^H}{\partial y}(\mathbf{x}) d\mathbf{x}
 \end{aligned}$$

where  $P_i^{U,t}$  is the reconstruction approximation function at time  $t$  of  $U_i(t)$  at cell  $V_i$  defined using the stencil  $\mathfrak{B}_i$ :

$$P_i^{U,t}(\mathbf{x}) = P_i(\mathbf{x}; \{U_j(t)\}_{j \in \mathfrak{B}_i}),$$

and  $P_i^H$  is the reconstruction approximation function of  $H$ . The functions  $U_{ij}^{\pm}(\gamma, t)$  are given by

$$U_{ij}^-(\gamma, t) = \lim_{\mathbf{x} \rightarrow \gamma} P_i^{U,t}(\mathbf{x}), \quad U_{ij}^+(\gamma, t) = \lim_{\mathbf{x} \rightarrow \gamma} P_j^{U,t}(\mathbf{x}),$$

and  $H_{ij}^{\pm}(\gamma)$  are given by

$$H_{ij}^-(\gamma) = \lim_{\mathbf{x} \rightarrow \gamma} P_i^H(\mathbf{x}), \quad H_{ij}^+(\gamma) = \lim_{\mathbf{x} \rightarrow \gamma} P_j^H(\mathbf{x}).$$

In practice, the integral terms in (20) must be approximated numerically using a high order quadrature formula, whose order is related to the one of the reconstruction operator (see [Castro, Fernández-Nieto, Ferreiro, García-Rodríguez, and Parés \[ibid.\]](#) for more details).

In order to properly define a numerical scheme,  $D_{\Phi}^-(U_l, U_r, H_l, H_r, \eta)$  should be prescribed. In the next section we briefly describe a general procedure to define  $D_{\Phi}^-(U_l, U_r, H_l, H_r, \eta)$ .

Finally, let us remark that a well-known problem related to the design of numerical schemes for non-conservative systems is the analysis of the convergence towards the weak-solutions, here we refer to [Castro, Morales de Luna, and Parés \[2017\]](#) for a review on this particular subject.

**2.1 FVM path-conservative schemes.** In this section we briefly present a procedure to define a family of first order path-conservative schemes named as FVM methods (see [Castro Díaz and Fernández-Nieto \[2012\]](#) and [Castro, Gallardo, and Marquina \[2014\]](#) and [Castro, Morales de Luna, and Parés \[2017\]](#) for a review). FVM reads from *functional viscosity matrix* since as it will be shown in the next paragraphs, they are characterized by a numerical viscosity matrix that results from a functional evaluation of a Roe matrix for system (2) or, more generally, of the matrix  $\mathcal{Q}_\eta = \mathcal{Q}_1\eta_x + \mathcal{Q}_2\eta_y$  evaluated in a given intermediate state.

Given a family of paths  $\Phi$ , a *Roe linearization* of system (2) is a function

$$\mathcal{Q}_\Phi: \tilde{\Omega} \times \tilde{\Omega} \times S^1 \rightarrow \mathfrak{M}_M(\mathbb{R})$$

satisfying the following properties for each  $W_l, W_r \in \tilde{\Omega}$  and  $\eta \in S^1$ :

1.  $\mathcal{Q}_\Phi(W_l, W_r, \eta)$  has  $M$  distinct real eigenvalues

$$\lambda_1(W_l, W_r, \eta) < \lambda_2(W_l, W_r, \eta) < \dots < \lambda_M(W_l, W_r, \eta).$$

2.  $\mathcal{Q}_\Phi(W, W, \eta) = \mathcal{Q}(W, \eta)$ .

3.  $\mathcal{Q}_\Phi(W_l, W_r, \eta) \cdot (W_r - W_l) =$

$$(21) \quad \int_0^1 \mathcal{Q}(\Phi(s; W_l, W_r, \eta), \eta) \frac{\partial \Phi}{\partial s}(s; W_l, W_r, \eta) ds.$$

Note that in the particular case in which  $\mathcal{Q}_k(W)$ ,  $k = 1, 2$ , are the Jacobian matrices of smooth flux functions  $F_k(W)$ , property (21) does not depend on the family of paths and reduces to the usual Roe property:

$$(22) \quad \mathcal{Q}_\Phi(W_l, W_r, \eta) \cdot (W_r - W_l) = F_\eta(W_r) - F_\eta(W_l)$$

for any  $\eta \in S^1$ .

Given a Roe matrix  $\mathcal{Q}_\Phi(W_l, W_r, \eta)$ , let us consider:

$$\widehat{\mathcal{Q}}_\Phi^\pm(W_l, W_r, \eta) = \frac{1}{2} (\mathcal{Q}_\Phi(W_l, W_r, \eta) \pm \mathcal{Q}_\Phi(W_l, W_r, \eta)),$$

where  $\mathcal{Q}_\Phi(W_l, W_r, \eta)$  is a semi-definite positive matrix that can be seen as the viscosity matrix associated to the method.

Now, it is straightforward to define a path-conservative scheme in the sense defined in [Parés \[2006\]](#) based on the previous definition:

$$(23) \quad \mathcal{D}_\phi^\pm(W_l, W_r, \eta) = \widehat{\mathcal{Q}}_\Phi^\pm(W_l, W_r, \eta)(W_r - W_l).$$

Finally, we could also define a path-conservative scheme for the system (1) as follows:

$$(24) \quad \begin{aligned} D_{\Phi}^{\pm}(U_l, U_r, H_l, H_r, \eta) &= \frac{1}{2}(F_{\eta}(W_r) - F_{\eta}(W_l) - B_{\Phi} \cdot (U_r - U_l) \\ &\quad - S_{\Phi}(H_r - H_l) \\ &\quad \pm Q_{\Phi} \cdot (U_r - U_l - A_{\Phi}^{-1} \cdot S_{\Phi}(H_r - H_l))) \end{aligned}$$

where the path is supposed to be given by  $\Phi = (\Phi_U \ \Phi_H)^T$  and

$$\begin{aligned} B_{\Phi} \cdot (U_r - U_l) &= B_{\Phi}(U_l, U_r, \eta) \cdot (U_r - U_l) \\ &= \int_0^1 B_{\eta}(\Phi_U(s; W_l, W_r, \eta)) \frac{\partial \Phi_U}{\partial s}(s; W_l, W_r, \eta) ds \end{aligned}$$

with

$$\begin{aligned} B_{\eta}(U) &= \eta_x B_1(U) + \eta_y B_2(U); \\ S_{\Phi}(H_r - H_l) &= S_{\Phi}(U_l, U_r, \eta)(H_r - H_l) \\ &= \int_0^1 S_{\eta}(\Phi_U(s; W_l, W_r, \eta)) \frac{\partial \Phi_H}{\partial s}(s; W_l, W_r, \eta) ds \end{aligned}$$

with

$$S_{\eta}(U) = \eta_x S_1(U) + \eta_y S_2(U).$$

The matrix  $A_{\Phi}$  is defined as follows

$$A_{\Phi} = A_{\Phi}(U_l, U_r, \eta) = J(U_l, U_r, \eta) + B_{\Phi}(U_l, U_r, \eta)$$

where  $J(U_l, U_r, \eta)$  is a Roe matrix for the flux  $F_{\eta}(U)$ , that is

$$J(U_l, U_r, \eta) \cdot (U_r - U_l) = F_{\eta}(U_r) - F_{\eta}(U_l).$$

**Remark 3.** The previous scheme can be derived from the standard Roe method for system (2), taking into account the structure of the matrices  $\mathcal{R}_{\Phi}^{\pm}(W_l, W_r, \eta)$ . In fact (24) exactly coincides with Roe method for system (1) setting  $Q_{\Phi} = |A_{\Phi}|$ .

**Remark 4.** Notice that the term  $Q_{\Phi} \cdot (U_r - U_l - A_{\Phi}^{-1} \cdot S_{\Phi}(H_r - H_l))$  is not well defined and makes no sense if one of the eigenvalues of  $A_{\Phi}$  vanishes. In this case, two eigenvalues of  $\mathcal{R}_{\Phi}(W_l, W_r, \eta)$  vanish and the problem is said to be resonant. Resonant problems exhibit an additional difficulty, as weak solutions may not be uniquely determined by their initial data. The analysis of this difficulty depends on the considered problem and it is beyond of this review. A general procedure, that formally avoids this difficulty is described in [Castro, Pardo, Parés, and E. F. Toro \[2010\]](#).

Finally, in order to fully define the scheme (24), the matrix  $Q_\Phi(U_l, U_r, \eta)$ , that plays the role of the viscosity matrix, has to be defined. For instance, the standard Roe method is obtained if  $Q_\Phi = |A_\Phi|$ . Note that with this choice one needs to perform the complete spectral decomposition of  $A_\Phi$ . In many situations, as in the case of the multilayer shallow-water system, it is not possible to obtain an easy analytical expression of the eigenvalues and eigenvectors, and a numerical algorithm has to be used to perform the spectral decomposition of matrix  $A_\Phi$ , increasing the computational cost of the Roe method.

A rough approximation is given by the local Lax-Friedrichs (or Rusanov) method, in which:

$$(25) \quad Q_\Phi(U_l, U_r, \eta) = \max(|\lambda_i(U_l, U_r, \eta)|, i = 1, \dots, N)I,$$

$I$  being the identity matrix. Note that this definition of  $Q_\Phi(U_l, U_r, \eta)$  only requires an estimation of the largest wave speed in absolute value. However, this approach gives excessive numerical diffusion for the waves corresponding to the lower eigenvalues.

The strategy that we follow is to consider viscosity matrices of the form

$$(26) \quad Q_\Phi(U_l, U_r) = f(A_\Phi(U_l, U_r, \eta)),$$

where,  $f : \mathbb{R} \mapsto \mathbb{R}$  satisfies the following properties:

- $f(x) \geq 0, \forall x \in \mathbb{R}$ ,
- $f(x)$  is *easy* to evaluate,
- the graph of  $f(x)$  is *close* to the graph of  $|x|$ .

Moreover, if

$$f(0) > 0$$

no entropy-fix techniques are required to avoid the appearance of non-entropy discontinuities at the numerical solutions.

The stability of the scheme is strongly related to the definition of the function  $f(x)$ . In particular, if  $\lambda_1(U_l, U_r, \eta) < \dots < \lambda_N(U_l, U_r, \eta)$  denote the eigenvalues of  $A_\Phi(U_l, U_r, \eta)$  and the usual CFL condition is assumed

$$(27) \quad \Delta t \cdot \max \left\{ \frac{|\lambda_{ij,k}|}{d_{ij}}; i = 1, \dots, N_{\mathcal{T}}, j \in \mathfrak{n}_i, k = 1, \dots, N \right\} = \delta,$$

with  $0 < \delta \leq 1$ , where  $d_{ij}$  is the distance from the center of cell  $V_i$  to the edge  $E_{ij}$ , then the resulting scheme is  $L^\infty$ -stable if  $f(x)$  satisfies the following condition [Castro Díaz and Fernández-Nieto \[2012\]](#):

$$(28) \quad f(x) \geq |x|, \quad \forall x \in [\lambda_1(U_l, U_r, \eta), \lambda_N(U_l, U_r, \eta)],$$

i.e., the graph of the function  $f(x)$  must be above the graph of the absolute value function in the interval containing the eigenvalues.

At the beginning of the eighties, Harten, Lax, and van Leer [1983] proposed to choose  $f$  as the linear polynomial  $p(x)$  that interpolates  $|x|$  at the smallest and largest eigenvalue, which results in a considerable improvement of the local Lax-Friedrichs method. This idea, which is on the basis of the HLL method, has been improved later by several authors (see E. F. Toro [2009] for a review).

To our knowledge, the paper by Degond et al. Degond, Peyrard, Russo, and Villedieu [1999] is the first attempt to construct a simple approximation of  $|A|$  by means of a polynomial that approximates  $|x|$  without interpolation of the absolute value function at the exact eigenvalues. This approach has been extended to a general framework in Castro Díaz and Fernández-Nieto [2012], where the so-called PVM (Polynomial Viscosity Matrix) methods are defined in terms of viscosity matrices based on general polynomial evaluations of a Roe matrix. The idea is to consider viscosity matrices of the form:

$$Q_{\Phi}(U_l, U_r, \eta) = P_r(A_{\Phi}(U_l, U_r, \eta)),$$

where  $P_r(x)$  is a polynomial of degree  $r$ .

A number of well-known schemes can be interpreted as PVM methods: this is the case for Roe, Lax-Friedrichs, Rusanov, HLL Harten, Lax, and van Leer [1983], FORCE E. F. Toro and Billett [2000], MUSTA E. F. Toro [2006] and E. F. Toro and Titarev [2006], etc. (see Castro Díaz and Fernández-Nieto [2012] for details). The numerical scheme introduced in Degond, Peyrard, Russo, and Villedieu [1999] and the Krylov-Riemann solver recently introduced in Torrilhon [2012] can be viewed as particular cases of PVM schemes as well.

In Morales de Luna, Castro Díaz, and Parés [2014], the relation between Simple Riemann Solvers (SRS) and PVM methods is analyzed. It has been shown that every PVM method can be interpreted as a SRS provided that it is based on a polynomial that interpolates the absolute value function at some points. Furthermore, the converse is true under some technical assumptions. Besides its theoretical interest, this relation provides a useful tool to investigate the properties of some well-known numerical methods that are particular cases of PVM methods, as the analysis of certain properties (like positivity preserving) is easier for SRS methods.

Besides the interpretation of well-known numerical schemes as PVM, this framework allows for the development of new ones. For instance, in Fernández-Nieto, Castro Díaz, and Parés [2011] a numerical method based on a polynomial that interpolates three values (the largest and lowest eigenvalues and the maximum of the intermediate ones) has been derived. This numerical method gives excellent results for the two-layer shallow water model. Another interesting family of PVM schemes based on *Chebyshev polynomials*,



which provide optimal uniform approximations to the absolute value function has been proposed in [Castro, Gallardo, and Marquina \[2014\]](#).

A natural extension of the PVM methods consists in the use of rational functions to approach the absolute value, since these functions provide more accurate approximations of this function in the uniform norm. Specifically, in [Castro, Gallardo, and Marquina \[ibid.\]](#) two families of Rational Viscosity Matrix (RVM) methods have been considered, based on the so-called Newman and Halley rational approximations of the absolute value function.

Another interesting application of functional viscosity methods is the derivation of new flux-limiter type schemes similar to the WAF (Weighted Average Flux) method introduced by Toro in [E. Toro \[1989\]](#). In [Castro Díaz, Fernández-Nieto, Narbona-Reina, and de la Asunción \[2014\]](#) a natural extension of the original WAF method has been proposed based on a non-linear combination of two PVM methods.

**2.2 Well-balancing.** In this section the well-balanced property of the scheme (20) is studied. Let us consider the following definitions:

**Definition 1.** Consider a semi-discrete method to approximate (1)

$$(29) \quad \begin{cases} U'_i(t) = \frac{1}{|V_i|} \mathcal{H}(U_j(t), j \in \mathcal{B}_i), \\ U(0) = U_0, \end{cases}$$

where  $U(t) = \{U_i(t)\}_{i=1}^{N_{\mathcal{T}}}$  represents the vector of the approximations to the averaged values of the exact solution;  $U_0 = \{U_i^0\}$  is the vector of the all averages of the initial conditions; and  $\mathcal{B}_i$  are the stencils. Given a smooth stationary solution  $U$  of the system, the numerical scheme is said to be exactly well-balanced for  $U$  if the vector of its cell averages is a critical point of (29), i.e.

$$(30) \quad \mathcal{H}(U_j, j \in \mathcal{B}_i) = 0.$$

Let us also introduce the concept of well-balanced reconstruction operator:

**Definition 2.** Given a smooth stationary solution of (1), a reconstruction operator is said to be well-balanced for  $U(\mathbf{x})$  if the approximation functions  $P_i(\mathbf{x})$  associated to the averaged values of  $U$  are also stationary solutions of the system (1).

**Remark 5.** Here, as  $H(\mathbf{x})$  is a given function, we set that its reconstruction is  $P_i^H(\mathbf{x}) = H(\mathbf{x}) \mathbf{x} \in V_i$ .

The following results can be proved:

**Theorem 2.** *Let  $U$  be a stationary solution of (1) and let us assume that the family of paths  $\Phi(s, W_l, W_r, \eta) = (\Phi_U(s, W_l, W_r, \eta), \Phi_H(s, W_l, W_r, \eta))^T$  connecting two states  $W_l = (U(x_l), H(x_l))^T$  and  $W_r = (U(x_r), H(x_r))^T$  with  $x_l < x_r$  is a reparametrization of  $x \in [x_l, x_r] \mapsto U(x)$ , then the first order FVM scheme is exactly well-balanced for  $U$ .*

**Theorem 3.** *Let  $U$  be a stationary solution of (1). Let us suppose that the first order FVM path-conservative scheme and the reconstruction operator chosen are exactly well-balanced for  $U$ . Then the numerical scheme (20) is also exactly well-balanced for  $U$ .*

**Remark 6.** *Note that if the stationary solution is smooth, then  $U_{ij}^- = U_{ij}^+$  and  $D_\Phi^\pm = 0$ , therefore, the well-balanced property of the high order method only depends on the well-balanced property of the reconstruction operator.*

Notice that standard reconstruction operators are not expected in general to be well-balanced. In [Castro, Gallardo, López-García, and Parés \[2008b\]](#) we propose a general procedure to modify any standard reconstruction operator  $P_i^{U,t}$  in order to be well-balanced for every stationary solution of (1). This procedure summarizes as follows:

Given a family of cell values  $\{U_i(t)\}$ , at every cell  $V_i$ :

1. Look for the stationary solution  $U_i^*(x)$  such that

$$(31) \quad \frac{1}{|V_i|} \int_{V_i} U_i^*(x) dx = U_i(t).$$

2. Apply the reconstruction operator to the cell values  $\{DU_j\}_{j \in \mathbb{B}_i}$  given by

$$DU_j(t) = U_j(t) - \frac{1}{|V_j|} \int_{V_j} U_i^*(x) dx, \quad j \in \mathbb{B}_i,$$

to obtain

$$\widetilde{P}_i^t(x) = P_i(x; \{DU_j(t)\}_{j \in \mathbb{B}_i}).$$

being  $P_i(x)$  a standard reconstruction operator that is exact for the null function.

Note that  $DU_j(t)$ ,  $j \in \mathbb{B}_i$  are the cell averages of the *fluctuations* with respect to the stationary solution and should be zero if  $U_j(t)$ ,  $j \in \mathbb{B}_i$  correspond to the cell average of the stationary solution  $U_i^*(x)$ .

3. Define

$$P_i^t(x) = U_i^*(x) + \widetilde{P}_i^t(x).$$

It can be easily checked that the reconstruction operator  $P_i^t$  is well-balanced for every stationary solution provided that the reconstruction operator  $P_i$  is exact for the null function and it is high-order accurate, and that the stationary solutions  $U_i^*$  are smooth. When

it is not possible to solve the equation (31) the reconstruction reduces to the standard one  $P_i(x)$ .

Finally, let us remark that quadrature formulae also play an important role to preserve the well-balanced properties of the scheme. In fact, the previous results have been established assuming that the integrals are exactly computed. Thus, in order to preserve the well-balanced properties, quadrature formulae should be exact for the stationary solutions. For that purpose, the strategy developed in [Castro, Ortega, and Parés \[2017\]](#) can be used.

### 3 Numerical tests

In this section two numerical tests related to tsunami modeling are presented. In the first one, the well-known Lituya Bay mega-tsunami is modeled and, in the second one, a hypothetical tsunami in the Mediterranean Sea is simulated. Both simulations have been run with the software package HySEA developed by Edanya group (University of Málaga). Tsunami-HySEA and Landslide-HySEA are the numerical models of the HySEA software specifically designed for tsunami simulations. Tsunami-HySEA model simulates with the same code the three parts of an earthquake generated tsunami (generation, propagation, and coastal inundation) using the non-linear shallow-water system on the sphere. In the generation stage, Okada's fault deformation model (see [Okada \[1985\]](#)) is used to predict the initial bottom deformation that is transmitted instantaneously to the sea surface generating the tsunami wave. This method assumes that an earthquake can be regarded as the rupture of a single fault plane. This fault is described by a set of parameters, including dip angle, strike angle, rake angle, fault width, fault length, and fault depth. Landslide-HySEA model implements the natural 2D extension of the 1D two-layer Savage-Hutter model presented in [Fernández-Nieto, Bouchut, Bresch, Castro Díaz, and Mangeney \[2008\]](#), where Cartesian coordinates are used instead of local coordinates at each point of the 2D domain and where no anisotropy effects are taken into account in the normal stress tensor of the solid phase. The mathematical model consists of two systems of equations that are coupled: the model for the slide material is represented by a Savage-Hutter type of model and the water dynamics model is represented by the shallow-water equations. Both models have been implemented on multiGPU architectures to speedup the computations. Modern *Graphics Processing Units (GPUs)* are highly programmable and massively parallel devices which can be used to accelerate considerably numerical computations in a cost-effective way [Brodtkorb, Hagen, and Sætra \[2013\]](#), [Owens, Houston, Luebke, Green, Stone, and Phillips \[2008\]](#), and [M. Ujaldon \[2012\]](#). They offer hundreds or thousands of processing units optimized for massively performing floating-point operations in parallel and have proven to be effective in the acceleration of numerical schemes which exhibit a lot of exploitable fine-grain parallelism. GPU computing consists of using GPUs together

with CPUs to accelerate the solution of compute-intensive science, engineering and enterprise problems. Since the numerical simulations based on PDEs present a lot of exploitable parallelism, there has been an increasing interest in the acceleration of these simulations by using GPU-based computer systems. There is a widespread use of CUDA-based platforms to accelerate numerical solvers for PDEs. See [de la Asunción, Castro, Fernández-Nieto, Mantas, Ortega Acosta, and González-Vida \[2013\]](#), [de la Asunción, Castro, Mantas, and Ortega \[2016\]](#), [de la Asunción and Castro \[2017\]](#), and [Mantas, de la Asunción, and Castro \[2016\]](#) and the references therein for some examples of the use of CUDA-based codes in geophysical flows.

**3.1 Lituya Bay mega-tsunami.** On July 9, 1958, an 8.3 magnitude (rated on the Richter scale) earthquake, along the Fairweather fault, triggered a major subaerial landslide into the Gilbert Inlet at the head of Lituya Bay on the southern coast of Alaska (USA). The landslide impacted the water at a very high speed generating a giant tsunami with the highest recorded wave run-up. The mega-tsunami run-up was up to an elevation of 524 m and caused total destruction of the forest as well as erosion down to the bedrock on a spur ridge, along the slide axis. Many attempts have been made to understand and simulate this mega tsunami. Here, we consider the 2D extension of the two-layer Savage-Hutter system described in [Fernández-Nieto, Bouchut, Bresch, Castro Díaz, and Mangeney \[2008\]](#). This system has been discretized by the second-order IFCP FVM path-conservative scheme described in [Fernández-Nieto, Castro Díaz, and Parés \[2011\]](#). Friction terms are discretized semi-implicitly. For fast computations, this scheme has been implemented on GPUs using CUDA. This two-dimensional scheme and its GPU adaptation and implementation using single numerical precision are described in [de la Asunción, Castro, Mantas, and Ortega \[2016\]](#).

A rectangular grid of  $3,648 \times 1,264 = 4,611,072$  cells with a resolution of  $4 \text{ m} \times 7.5 \text{ m}$  has been designed in order to perform this simulation. We use public domain topo-bathymetric data as well as the review paper [Miller \[1960\]](#) to approximate the Gilbert inlet topo-bathymetry. The parameters describing the properties of the sediment layer and those present in the friction laws have been calibrated with some laboratory experiments. The CFL number is set to 0.9. Figures [1\(a\)-1\(b\)](#) show the simulated free-surface elevation at 39 s. and 120 s. The maximum run-up is reached at 39 s.

While the initial wave moves through the main axis of Lituya Bay, a larger second wave appears as reflection of the first one from the south shoreline (see [Figure 1\(b\)](#)). These waves sweep both sides of the shoreline in their way. In the north shoreline, the wave reaches between 15-20 m height while in the south shoreline the wave reaches values between 20-30 m.

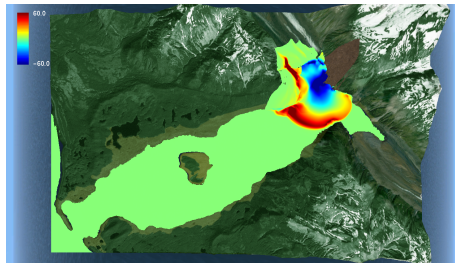
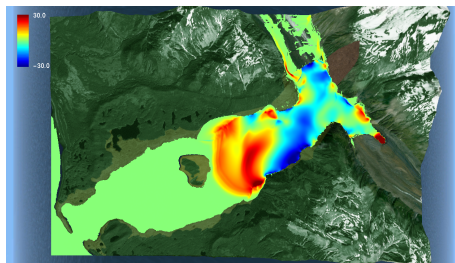
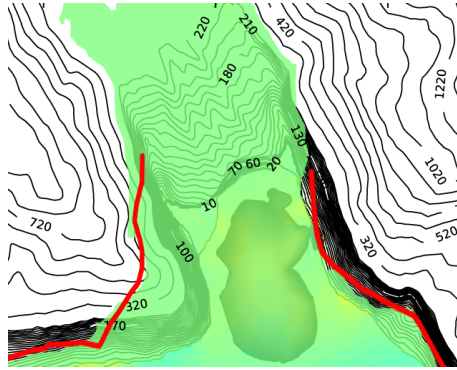
(a) Free surface at  $t = 39$  seconds.(b) Free surface at  $t = 120$  seconds.

Figure 1: Evolution of the Lituya Bay mega-tsunami

Figures 2(a) and 2(b) show the comparison between the computed maximum run-up (inundation) and the measured, provided by NOAA, in red, at the near field and the Cenotaph island, located in the middle of the narrow bay. The impact time and the maximum run-up provided by the simulation are in good agreement with the majority of observations and conclusions described by Miller [1960].

**3.2 Hypothetical tsunami in the eastern Mediterranean basin.** In this test we consider the one-layer shallow-water equations (SWE). The application of SWE to large scale phenomena (of the order of 1000's of km) makes necessary to take into account the curvature of the Earth. Usually, the Earth is approached by a sphere and the equations are written in spherical coordinates. Although the PDE system is similar to the SWEs in the plane using Cartesian coordinates, new source terms appear due to the change of variables. Therefore, the discretization of the system in spherical coordinates goes far beyond a simple adaptation of the numerical methods for the equations written in Cartesian coordinates. In Castro, Ortega, and Parés [2017] a third order path-conservative scheme has been presented. The numerical scheme is exactly well-balanced for the water at rest, that in this particular case, must take into account the curvature of the Earth.



(a) Maximum runup at the Gilbert inlet.



(b) Maximum runup at Cenotaph island.

Figure 2: Comparison between the computed and the observed (in red) maximum runup in Gilbert intlet and Cenotaph island (Lituya Bay event).

In this test we simulate the evolution of a hypothetical tsunami in the eastern Mediterranean basin. A uniform cartesian grid of the rectangular domain in the  $\tilde{\theta}$ - $\tilde{\varphi}$  plane (that is, the longitude and latitude in degrees), given by  $[6.25, 36.25] \times [30.25, 45.65]$  with  $\Delta_{\tilde{\theta}} = \Delta_{\tilde{\varphi}} = 30''$ . The mean radius of the Earth is set to  $R = 6371009.4$  m and the CFL parameter is set to 0.5. Open boundary conditions are prescribed at the four boundaries. The topo-bathymetry of the area has been interpolated from the ETOPO1 Global Relief Model (see [Amante and Eakins \[n.d.\]](#)). Next, a hypothetical seafloor deformation generated by an earthquake of magnitude  $M_s = 8$  has been computed using the Okada model (see [Okada \[1985\]](#)). This seafloor deformation is instantaneously transmitted to the water column to generate the initial tsunami profile (see [Figure 3\(a\)](#)). The initial velocity is set to zero. Concerning the numerical treatment of wet/dry fronts, here we follow the ideas described in [Gallardo, Parés, and Castro \[2007\]](#), that have been adapted to the

reconstruction operator defined in [Gallardo, Ortega, de la Asunción, and Mantas \[2011\]](#). Figures 3 and 4 show the evolution of the tsunami wave propagating along the eastern Mediterranean Sea. Note that after approximately one hour, the waves generated near to the Greek coasts, arrive to the north of Africa and south of Italy (see Figures 3 and 4).

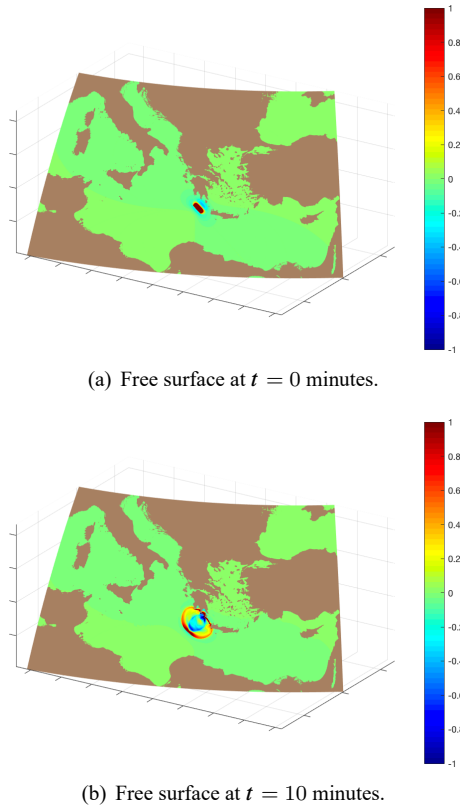


Figure 3: Evolution of a hypothetical tsunami at the eastern Mediterranean basin.

## References

Christopher Amante and Barry W Eakins (n.d.). “ETOPO1 1 arc-minute global relief model: procedures, data sources and analysis” (cit. on p. [3551](#)).

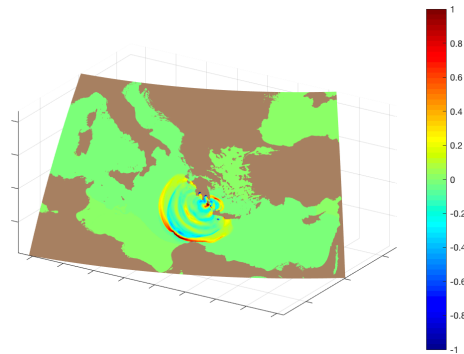
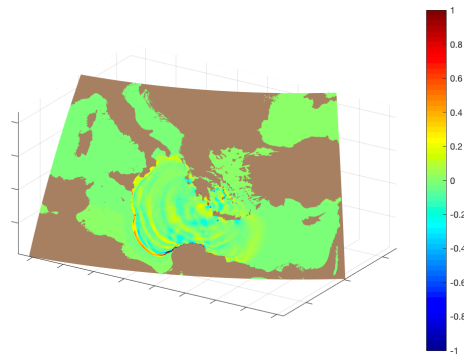
(a) Free surface at  $t = 30$  minutes(b) Free surface at  $t = 60$  minutes

Figure 4: Evolution of a hypothetical tsunami at the eastern Mediterranean basin.

- M. de la Asunción and Manuel J. Castro (2017). “[Simulation of tsunamis generated by landslides using adaptive mesh refinement on GPU](#)”. *J. Comput. Phys.* 345, pp. 91–110. MR: [3667605](#) (cit. on p. [3549](#)).
- Marc de la Asunción, Manuel J. Castro, E. D. Fernández-Nieto, José M. Mantas, Sergio Ortega Acosta, and José Manuel González-Vida (2013). “[Efficient GPU implementation of a two waves TVD-WAF method for the two-dimensional one layer shallow water system on structured meshes](#)”. *Comput. & Fluids* 80, pp. 441–452. MR: [3068086](#) (cit. on p. [3549](#)).
- Marc de la Asunción, Manuel J. Castro, José Miguel Mantas, and Sergio Ortega (2016). “Numerical simulation of tsunamis generated by landslides on multiple GPUs”. *Advances in Engineering Software* 99, pp. 59–72 (cit. on p. [3549](#)).



- Emmanuel Audusse, Marie-Odile Bristeau, Benoît Perthame, and Jacques Sainte-Marie (2011). “[A multilayer Saint-Venant system with mass exchanges for shallow water flows. Derivation and numerical validation](#)”. *ESAIM Math. Model. Numer. Anal.* 45.1, pp. 169–200. MR: [2781135](#) (cit. on p. [3534](#)).
- Scott F Bradford and Nikolaos D Katopodes (1999). “Hydrodynamics of turbid underflows. I: Formulation and numerical analysis”. *Journal of hydraulic engineering* 125.10, pp. 1006–1015 (cit. on p. [3534](#)).
- André R Brodtkorb, Trond R Hagen, and Martin L Sætra (2013). “Graphics processing unit (GPU) programming strategies and trends in GPU computing”. *Journal of Parallel and Distributed Computing* 73.1, pp. 4–13 (cit. on p. [3548](#)).
- Manuel J. Castro Díaz, Yuanzhen Cheng, Alina Chertock, and Alexander Kurganov (2014). “[Solving two-mode shallow water equations using finite volume methods](#)”. *Commun. Comput. Phys.* 16.5, pp. 1323–1354. MR: [3256969](#) (cit. on p. [3534](#)).
- Manuel J. Castro Díaz and E. D. Fernández-Nieto (2012). “[A class of computationally fast first order finite volume solvers: PVM methods](#)”. *SIAM J. Sci. Comput.* 34.4, A2173–A2196. MR: [2970401](#) (cit. on pp. [3542](#), [3544](#), [3545](#)).
- Manuel J. Castro Díaz, E. D. Fernández-Nieto, and A. M. Ferreiro (2008). “[Sediment transport models in shallow water equations and numerical approach by high order finite volume methods](#)”. *Comput. & Fluids* 37.3, pp. 299–316. MR: [2645529](#) (cit. on p. [3534](#)).
- Manuel J. Castro Díaz, E. D. Fernández-Nieto, G. Narbona-Reina, and M. de la Asunción (2014). “[A second order PVM flux limiter method. Application to magnetohydrodynamics and shallow stratified flows](#)”. *J. Comput. Phys.* 262, pp. 172–193. MR: [3163113](#) (cit. on p. [3546](#)).
- Manuel J. Castro Díaz, Enrique Domingo Fernández-Nieto, Tomás Morales de Luna, Gladys Narbona-Reina, and Carlos Parés (2013). “[A HLLC scheme for nonconservative hyperbolic problems. Application to turbidity currents with sediment transport](#)”. *ESAIM Math. Model. Numer. Anal.* 47.1, pp. 1–32. MR: [2968693](#) (cit. on p. [3535](#)).
- Manuel J. Castro, E. D. Fernández-Nieto, A. M. Ferreiro, J. A. García-Rodríguez, and C. Parés (2009). “[High order extensions of Roe schemes for two-dimensional nonconservative hyperbolic systems](#)”. *J. Sci. Comput.* 39.1, pp. 67–114. MR: [2495820](#) (cit. on pp. [3534](#), [3537](#), [3540](#), [3541](#)).
- Manuel J. Castro, José M. Gallardo, Juan A. López-García, and Carlos Parés (2008a). “[Well-balanced high order extensions of Godunov’s method for semilinear balance laws](#)”. *SIAM J. Numer. Anal.* 46.2, pp. 1012–1039. MR: [2383221](#) (cit. on p. [3534](#)).
- (2008b). “[Well-balanced high order extensions of Godunov’s method for semilinear balance laws](#)”. *SIAM J. Numer. Anal.* 46.2, pp. 1012–1039. MR: [2383221](#) (cit. on p. [3547](#)).

- Manuel J. Castro, José M. Gallardo, and Antonio Marquina (2014). “A class of incomplete Riemann solvers based on uniform rational approximations to the absolute value function”. *J. Sci. Comput.* 60.2, pp. 363–389. MR: [3225787](#) (cit. on pp. [3542](#), [3546](#)).
- Manuel J. Castro, José M. Gallardo, and Carlos Parés (2006). “High order finite volume schemes based on reconstruction of states for solving hyperbolic systems with non-conservative products. Applications to shallow-water systems”. *Math. Comp.* 75.255, pp. 1103–1134. MR: [2219021](#) (cit. on p. [3534](#)).
- Manuel J. Castro, T. Morales de Luna, and C. Parés (2017). “Well-balanced schemes and path-conservative numerical methods”. In: *Handbook of numerical methods for hyperbolic problems*. Vol. 18. Handb. Numer. Anal. Elsevier/North-Holland, Amsterdam, pp. 131–175. MR: [3645391](#) (cit. on pp. [3534](#), [3535](#), [3541](#), [3542](#)).
- Manuel J. Castro, Sergio Ortega, and Carlos Parés (2017). “Well-balanced methods for the shallow water equations in spherical coordinates”. *Comput. & Fluids* 157, pp. 196–207. MR: [3706680](#) (cit. on pp. [3548](#), [3550](#)).
- Manuel J. Castro, Alberto Pardo, Carlos Parés, and E. F. Toro (2010). “On some fast well-balanced first order solvers for nonconservative systems”. *Math. Comp.* 79.271, pp. 1427–1472. MR: [2629999](#) (cit. on p. [3543](#)).
- I. Cravero and M. Semplice (2016). “On the accuracy of WENO and CWENO reconstructions of third order on nonuniform meshes”. *J. Sci. Comput.* 67.3, pp. 1219–1246. MR: [3493501](#) (cit. on pp. [3537](#), [3539](#)).
- Gianni Dal Maso, Philippe G. Lefloch, and François Murat (1995). “Definition and weak stability of nonconservative products”. *J. Math. Pures Appl. (9)* 74.6, pp. 483–548. MR: [1365258](#) (cit. on p. [3535](#)).
- Pierre Degond, Pierre-François Peyrard, Giovanni Russo, and Philippe Villedieu (1999). “Polynomial upwind schemes for hyperbolic systems”. *C. R. Acad. Sci. Paris Sér. I Math.* 328.6, pp. 479–483. MR: [1680004](#) (cit. on p. [3545](#)).
- Michael Dumbser, Dinshaw S. Balsara, Eleuterio F. Toro, and Claus-Dieter Munz (2008). “A unified framework for the construction of one-step finite volume and discontinuous Galerkin schemes on unstructured meshes”. *J. Comput. Phys.* 227.18, pp. 8209–8253. MR: [2446488](#) (cit. on p. [3537](#)).
- Michael Dumbser, Manuel J. Castro, Carlos Parés, and Eleuterio F. Toro (2009). “ADER schemes on unstructured meshes for nonconservative hyperbolic systems: applications to geophysical flows”. *Comput. & Fluids* 38.9, pp. 1731–1748. MR: [2645784](#) (cit. on p. [3538](#)).
- Michael Dumbser, Cedric Enaux, and Eleuterio F. Toro (2008). “Finite volume schemes of very high order of accuracy for stiff hyperbolic balance laws”. *J. Comput. Phys.* 227.8, pp. 3971–4001. MR: [2403874](#) (cit. on p. [3534](#)).

- Michael Dumbser and Martin Käser (2007). “Arbitrary high order non-oscillatory finite volume schemes on unstructured meshes for linear hyperbolic systems”. *J. Comput. Phys.* 221.2, pp. 693–723. MR: [2293146](#) (cit. on p. [3537](#)).
- Michael Dumbser, Martin Käser, Vladimir A. Titarev, and Eleuterio F. Toro (2007). “Quadrature-free non-oscillatory finite volume schemes on unstructured meshes for nonlinear hyperbolic systems”. *J. Comput. Phys.* 226.1, pp. 204–243. MR: [2356357](#) (cit. on p. [3537](#)).
- Michael Dumbser and Claus-Dieter Munz (2006). “Building blocks for arbitrary high order discontinuous Galerkin schemes”. *J. Sci. Comput.* 27.1-3, pp. 215–230. MR: [2285777](#).
- Felix M Exner (1925). “Über die wechselwirkung zwischen wasser und geschiebe in flüssen”. *Sitzungsber., Akad. Wissenschaften* (cit. on p. [3534](#)).
- E. D. Fernández-Nieto, F. Bouchut, D. Bresch, Manuel J. Castro Díaz, and A. Mangeney (2008). “A new Savage-Hutter type model for submarine avalanches and generated tsunami”. *J. Comput. Phys.* 227.16, pp. 7720–7754. MR: [2437587](#) (cit. on pp. [3548](#), [3549](#)).
- E. D. Fernández-Nieto, Manuel J. Castro Díaz, and C. Parés (2011). “On an intermediate field capturing Riemann solver based on a parabolic viscosity matrix for the two-layer shallow water system”. *J. Sci. Comput.* 48.1-3, pp. 117–140. MR: [2811694](#) (cit. on pp. [3545](#), [3549](#)).
- E. D. Fernández-Nieto, E. H. Koné, and T. Chacón Rebollo (2014). “A multilayer method for the hydrostatic Navier-Stokes equations: a particular weak solution”. *J. Sci. Comput.* 60.2, pp. 408–437. MR: [3225789](#) (cit. on p. [3534](#)).
- E. D. Fernández-Nieto, C. Lucas, T. Morales de Luna, and S. Cordier (2014). “On the influence of the thickness of the sediment moving layer in the definition of the bedload transport formula in Exner systems”. *Comput. & Fluids* 91, pp. 87–106. MR: [3151293](#) (cit. on p. [3534](#)).
- José M. Gallardo, Sergio Ortega, Marc de la Asunción, and José Miguel Mantas (2011). “Two-dimensional compact third-order polynomial reconstructions. Solving nonconservative hyperbolic systems using GPUs”. *J. Sci. Comput.* 48.1-3, pp. 141–163. MR: [2811695](#) (cit. on pp. [3534](#), [3537](#), [3552](#)).
- José M. Gallardo, Carlos Parés, and Manuel J. Castro (2007). “On a well-balanced high-order finite volume scheme for shallow water equations with topography and dry areas”. *J. Comput. Phys.* 227.1, pp. 574–601. MR: [2361537](#) (cit. on pp. [3534](#), [3551](#)).
- Sigal Gottlieb and Chi-Wang Shu (1998). “Total variation diminishing Runge-Kutta schemes”. *Math. Comp.* 67.221, pp. 73–85. MR: [1443118](#) (cit. on p. [3538](#)).
- AJ Grass (1981). *Sediment transport by waves and currents*. SERC London Cent. Mar. Technol. Report, No. FL29 (cit. on p. [3534](#)).

- Ami Harten, Björn Engquist, Stanley Osher, and Sukumar R. Chakravarthy (1987). “Uniformly high-order accurate essentially nonoscillatory schemes. III”. *J. Comput. Phys.* 71.2, pp. 231–303. MR: [897244](#) (cit. on p. [3537](#)).
- Amiram Harten, Peter D. Lax, and Bram van Leer (1983). “On upstream differencing and Godunov-type schemes for hyperbolic conservation laws”. *SIAM Rev.* 25.1, pp. 35–61. MR: [693713](#) (cit. on p. [3545](#)).
- Boualem Khouider, Andrew J. Majda, and Samuel N. Stechmann (2008). “Nonlinear dynamics of hydrostatic internal gravity waves”. *Theor. Comput. Fluid Dyn* 22, pp. 407–432 (cit. on p. [3534](#)).
- M. Lukáčová-Medvid'ová, S. Noelle, and M. Kraft (2007). “Well-balanced finite volume evolution Galerkin methods for the shallow water equations”. *J. Comput. Phys.* 221.1, pp. 122–147. MR: [2290566](#) (cit. on p. [3534](#)).
- M. Ujaldon (2012). “High performance computing and simulations on the GPU using CUDA”. In: *2012 International Conference on High Performance Computing and Simulation HPCS 2012, Madrid, Spain, July 2-6*, pp. 1–7 (cit. on p. [3548](#)).
- José Miguel Mantas, Marc de la Asunción, and Manuel J. Castro (2016). “An introduction to GPU computing for numerical simulation”. In: *Numerical simulation in physics and engineering*. Vol. 9. SEMA SIMAI Springer Ser. Springer, [Cham], pp. 219–251. MR: [3561485](#) (cit. on p. [3549](#)).
- Antonio Marquina (1994). “Local piecewise hyperbolic reconstruction of numerical fluxes for nonlinear scalar conservation laws”. *SIAM J. Sci. Comput.* 15.4, pp. 892–915. MR: [1278006](#) (cit. on p. [3537](#)).
- Don John Miller (1960). *Giant Waves in Lituya Bay, Alaska: A Timely Account of the Nature and Possible Causes of Certain Giant Waves, with Eyewitness Reports of Their Destructive Capacity*. Professional paper (cit. on pp. [3549](#), [3550](#)).
- T. Morales de Luna, Manuel J. Castro Díaz, C. Parés Madroñal, and E. D. Fernández Nieto (2009). “On a shallow water model for the simulation of turbidity currents”. *Commun. Comput. Phys.* 6.4, pp. 848–882. MR: [2672326](#) (cit. on p. [3534](#)).
- T. Morales de Luna, E. D. Fernández Nieto, and Manuel J. Castro Díaz (2017). “Derivation of a multilayer approach to model suspended sediment transport: application to hyperpycnal and hypopycnal plumes”. *Commun. Comput. Phys.* 22.5, pp. 1439–1485. MR: [3718006](#) (cit. on p. [3534](#)).
- Tomás Morales de Luna, Manuel J. Castro Díaz, and Carlos Parés (2014). “Relation between PVM schemes and simple Riemann solvers”. *Numer. Methods Partial Differential Equations* 30.4, pp. 1315–1341. MR: [3200278](#) (cit. on p. [3545](#)).
- Sebastian Noelle, Normann Pankratz, Gabriella Puppo, and Jostein R. Natvig (2006). “Well-balanced finite volume schemes of arbitrary order of accuracy for shallow water flows”. *J. Comput. Phys.* 213.2, pp. 474–499. MR: [2207248](#) (cit. on p. [3534](#)).

- Sebastian Noelle, Yulong Xing, and Chi-Wang Shu (2007). “High-order well-balanced finite volume WENO schemes for shallow water equation with moving water”. *J. Comput. Phys.* 226.1, pp. 29–58. MR: [2356351](#) (cit. on p. [3534](#)).
- Yoshimitsu Okada (1985). “Surface deformation due to shear and tensile faults in a half-space”. *Bulletin of the seismological society of America* 75.4, pp. 1135–1154 (cit. on pp. [3548](#), [3551](#)).
- John D Owens, Mike Houston, David Luebke, Simon Green, John E Stone, and James C Phillips (2008). “GPU computing”. *Proceedings of the IEEE* 96.5, pp. 879–899 (cit. on p. [3548](#)).
- Carlos Parés (2006). “Numerical methods for nonconservative hyperbolic systems: a theoretical framework”. *SIAM J. Numer. Anal.* 44.1, pp. 300–321. MR: [2217384](#) (cit. on pp. [3539](#), [3542](#)).
- P. Ripa (1993). “Conservation laws for primitive equations models with inhomogeneous layers”. *Geophys. Astrophys. Fluid Dynam.* 70.1–4, pp. 85–111. MR: [1278169](#) (cit. on p. [3534](#)).
- Giovanni Russo and Alexander Khe (2009). “High order well balanced schemes for systems of balance laws”. In: *Hyperbolic problems: theory, numerics and applications*. Vol. 67. Proc. Sympos. Appl. Math. Amer. Math. Soc., Providence, RI, pp. 919–928. MR: [2605287](#) (cit. on p. [3534](#)).
- Chi-Wang Shu (1998). “Essentially non-oscillatory and weighted essentially non-oscillatory schemes for hyperbolic conservation laws”. In: *Advanced numerical approximation of nonlinear hyperbolic equations (Cetraro, 1997)*. Vol. 1697. Lecture Notes in Math. Springer, Berlin, pp. 325–432. MR: [1728856](#) (cit. on p. [3537](#)).
- Chi-Wang Shu and Stanley Osher (1988). “Efficient implementation of essentially nonoscillatory shock-capturing schemes”. *J. Comput. Phys.* 77.2, pp. 439–471. MR: [954915](#) (cit. on pp. [3538](#), [3539](#)).
- (1989). “Efficient implementation of essentially nonoscillatory shock-capturing schemes. II”. *J. Comput. Phys.* 83.1, pp. 32–78. MR: [1010162](#) (cit. on p. [3537](#)).
- V. A. Titarev and E. F. Toro (2005). “ADER schemes for three-dimensional non-linear hyperbolic systems”. *J. Comput. Phys.* 204.2, pp. 715–736. MR: [2131859](#) (cit. on p. [3538](#)).
- E. F. Toro (2006). “MUSTA: a multi-stage numerical flux”. *Appl. Numer. Math.* 56.10–11, pp. 1464–1479. MR: [2245468](#) (cit. on p. [3545](#)).
- E. F. Toro and S. J. Billett (2000). “Centred TVD schemes for hyperbolic conservation laws”. *IMA J. Numer. Anal.* 20.1, pp. 47–79. MR: [1736950](#) (cit. on p. [3545](#)).
- E. F. Toro and V. A. Titarev (2006). “MUSTA fluxes for systems of conservation laws”. *J. Comput. Phys.* 216.2, pp. 403–429. MR: [2235378](#) (cit. on p. [3545](#)).

- EF Toro (1989). “A weighted average flux method for hyperbolic conservation laws”. In: *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*. Vol. 423. 1865, pp. 401–418 (cit. on p. [3546](#)).
- Eleuterio F Toro (2001). *Shock-capturing methods for free-surface shallow flows*. Wiley and Sons Ltd., Chichester (cit. on p. [3534](#)).
- Eleuterio F. Toro (2009). *Riemann solvers and numerical methods for fluid dynamics*. Third. A practical introduction. Springer-Verlag, Berlin, pp. xxiv+724. MR: [2731357](#) (cit. on p. [3545](#)).
- Manuel Torrilhon (2012). “Krylov-Riemann solver for large hyperbolic systems of conservation laws”. *SIAM J. Sci. Comput.* 34.4, A2072–A2091. MR: [2970397](#) (cit. on p. [3545](#)).
- A. I. Volpert (1967). “Spaces BV and quasilinear equations”. *Mat. Sb. (N.S.)* 73 (115), pp. 255–302. MR: [0216338](#) (cit. on p. [3535](#)).
- Y. Xing (2017). “Numerical methods for the nonlinear shallow water equations”. In: *Handbook of numerical methods for hyperbolic problems*. Vol. 18. Handb. Numer. Anal. Elsevier/North-Holland, Amsterdam, pp. 361–384. MR: [3645398](#) (cit. on p. [3534](#)).
- Yulong Xing and Chi-Wang Shu (2006). “High order well-balanced finite volume WENO schemes and discontinuous Galerkin methods for a class of hyperbolic systems with source terms”. *J. Comput. Phys.* 214.2, pp. 567–598. MR: [2216604](#) (cit. on p. [3534](#)).

Received 2017-11-29.

MANUEL J. CASTRO

DPTO. ANÁLISIS MATEMÁTICO, ESTADÍSTICA E INVESTIGACIÓN OPERATIVA Y MATEMÁTICA APLICADA. UNIVERSIDAD DE MÁLAGA. CAMPUS DE TEATINOS, S/N, 29071 MÁLAGA, SPAIN  
[mjcastro@uma.es](mailto:mjcastro@uma.es)

MARC DE LA ASUNCIÓN

DPTO. ANÁLISIS MATEMÁTICO, ESTADÍSTICA E INVESTIGACIÓN OPERATIVA Y MATEMÁTICA APLICADA. UNIVERSIDAD DE MÁLAGA. CAMPUS DE TEATINOS, S/N, 29071 MÁLAGA, SPAIN  
[marcah@uma.es](mailto:marcah@uma.es)

ENRIQUE D. FERNÁNDEZ NIETO

DEPARTAMENTO DE MATEMÁTICA APLICADA I, UNIVERSIDAD DE SEVILLA. E. T. S. ARQUITECTURA. AVDA, REINA MERCEDES, S/N. 41012 SEVILLA, SPAIN  
[edofer@us.es](mailto:edofer@us.es)

JOSÉ M. GALLARDO

DPTO. ANÁLISIS MATEMÁTICO, ESTADÍSTICA E INVESTIGACIÓN OPERATIVA Y MATEMÁTICA APLICADA. UNIVERSIDAD DE MÁLAGA. CAMPUS DE TEATINOS, S/N, 29071 MÁLAGA, SPAIN  
[jmgallardo@uma.es](mailto:jmgallardo@uma.es)

JOSÉ M. GONZÁLEZ VIDA

DPTO. MATEMÁTICA APLICADA. UNIVERSIDAD DE MÁLAGA. CAMPUS DE TEATINOS, S/N, 29071 MÁLAGA, SPAIN  
[jgv@uma.es](mailto:jgv@uma.es)

JORGE MACÍAS  
DPTO. ANÁLISIS MATEMÁTICO, ESTADÍSTICA E INVESTIGACIÓN OPERATIVA Y MATEMÁTICA APLICADA. UNIVERSIDAD  
DE MÁLAGA. CAMPUS DE TEATINOS, S/N, 29071 MÁLAGA, SPAIN  
[jmacias@uma.es](mailto:jmacias@uma.es)

TOMÁS MORALES  
DPTO. DE MATEMÁTICAS. UNIVERSIDAD DE CÓRDOBA. CAMPUS DE RABANALES. 14071 CÓRDOBA, SPAIN  
[tomas.morales@uco.es](mailto:tomas.morales@uco.es)

SERGIO ORTEGA  
LABORATORIO DE MÉTODOS NUMÉRICOS. SCAI. UNIVERSIDAD DE MÁLAGA. CAMPUS DE TEATINOS, S/N, 29071  
MÁLAGA, SPAIN  
[sergio.ortega@uma.es](mailto:sergio.ortega@uma.es)

CARLOS PARÉS  
DPTO. ANÁLISIS MATEMÁTICO, ESTADÍSTICA E INVESTIGACIÓN OPERATIVA Y MATEMÁTICA APLICADA. UNIVERSIDAD  
DE MÁLAGA. CAMPUS DE TEATINOS, S/N, 29071 MÁLAGA, SPAIN  
[pres@uma.es](mailto:pres@uma.es)

# AN INVITATION TO NONLOCAL MODELING, ANALYSIS AND COMPUTATION

QIANG DU (杜强)

## Abstract

This lecture serves as an invitation to further studies on nonlocal models, their mathematics, computation, and applications. We sample our recent attempts in the development of a systematic mathematical framework for nonlocal models, including basic elements of nonlocal vector calculus, well-posedness of nonlocal variational problems, coupling to local models, convergence and compatibility of numerical approximations, and applications to nonlocal mechanics and diffusion. We also draw connections with traditional models and other relevant mathematical subjects.

## 1 Introduction

Nonlocal phenomena are ubiquitous in nature but their effective modeling and simulations can be difficult. In early mathematical and scientific inquiries, making local approximations has been a dominant strategy. Over centuries, popular continuum models are presented as partial differential equations (PDEs) that are expressed by local information in an infinitesimal neighborhood and are derived, in their origins, for smooth quantities. Entering into the digital age, there have been growing interests and capabilities in the modeling of complex processes that exhibit singularities/anomalies and involve nonlocal interactions. Nonlocal continuum models, fueled by the advent in computing technology, have the potential to be alternatives to local PDE models in many applications, although there are many new challenges for mathematicians and computational scientists to tackle.

While mathematical analysis and numerical solution of local PDEs are well established branches of mathematics, the development of rigorous theoretical and computational framework for nonlocal models, relatively speaking, is still a nascent field. This

---

Supported in part by the U.S. NSF grants DMS–1719699, AFOSR FA9550-14-1-0073 MURI Center for Material Failure Prediction Through Peridynamics, OSD/ARO/MURI W911NF-15-1-0562 on Fractional PDEs for Conservation Laws and Beyond: Theory, Numerics and Applications, and the Columbia University.

*MSC2010:* primary 45P05; secondary 65R20, 35R09, 74G15, 76M25, 47G10.

*Keywords:* Nonlocal operators, nonlocal function spaces, nonlocal vector calculus, numerical methods, asymptotic compatibility, heterogeneous localization, nonlocal in time dynamics.



lecture serves as an invitation to further studies on this emerging subject. We refer to [Du \[n.d.\]](#), an NSF-CBMS monograph, for a review on the historical development, recent progress and connections with other mathematical subjects such as multiscale analysis, calculus of variations, functional analysis, fractional PDEs, differential geometry, graph, data and image analysis, deep learning, as well as various applications. Instead of a brief survey, we present here samples of our recent attempts to develop a systematic mathematical framework for nonlocal models, including some basic building blocks, algorithms and applications. In particular, our discussions are centered around nonlocal models with a finite range of interactions typically characterized by a horizon parameter  $\delta$ . Their local ( $\delta \rightarrow 0$ ) and global ( $\delta \rightarrow \infty$ ) limits offer natural links to local and fractional PDEs and their discretization are also tied with graph operators, point clouds and discrete networks. A few questions on nonlocal modeling, analysis and computation are addressed here: how do nonlocal models compare with local and discrete models and how are they connected with each other? what are the ingredients of nonlocal vector calculus? how to develop robust discretization of nonlocal models that are asymptotically compatible with their local limit? how to get well defined trace maps in larger nonlocal function spaces to couple nonlocal and local models? and, how to explain the crossover of diffusion regimes using nonlocal in time dynamics? It is our intention to demonstrate that studies on nonlocal modeling not only provoke the discovery of new mathematics to guide practical modeling efforts, but also provide new perspectives to understand traditional models and new insight into their connections.

## 2 Modeling choices and emergence of nonlocal modeling

Mathematical models have various types, e.g., discrete or continuum, and deterministic or stochastic. Historically, influenced by the great minds like Newton, Leibniz, Maxwell and others, most popular continuum models are those given by PDEs whose simple close-form or approximate solutions have often been utilized. As more recent human endeavors, nevertheless, computer simulations have made discrete models equally prominent.

We consider some simple continuum and discrete equations as illustrations. Let  $u = u(x)$  be a function to be determined on a domain (an interval)  $\Omega \subset \mathbb{R}$ . The differential equation

$$-\mathcal{L}_0 u(x) = -\frac{d^2 u}{dx^2}(x) = f(x, u(x)), \quad \forall x \in \Omega$$

with a prescribed function  $f = f(x, u)$ , represents a local continuum model: it only involves the value and a few derivatives of the solution at any single point  $x$ . By introducing a set of grid points  $\{x_j\}$  in  $\Omega$ , equally spaced with a grid spacing  $h$  and the standard 2nd order center difference operator  $\mathcal{L}_h = D_h^2$  on the grid. We then have a discrete difference

model

$$-\mathfrak{L}_h u(x_j) = -D_h^2 u(x_j) = -\frac{u(x_j + h) - 2u(x_j) + u(x_j - h)}{h^2} = f(x_j, u(x_j)) \quad \forall x_j$$

that, as  $h \rightarrow 0$ , approximates the local continuum model. In comparison, we consider

$$(1) \quad -\mathfrak{L}_\delta u(x) = f(x, u(x)), \quad x \in \Omega,$$

which is a nonlocal model [Du \[2015, 2017b\]](#) and [Du and X. Tian \[2015\]](#) with a nonlocal operator  $\mathfrak{L}_\delta$  defined, for a prescribed nonlocal interaction kernel  $\underline{\omega}_\delta$  associated with a given horizon parameter  $\delta > 0$ , by

$$(2) \quad \mathfrak{L}_\delta u(x) = \int_0^\delta \frac{u(x+s) - 2u(x) + u(x-s)}{s^2} \underline{\omega}_\delta(s) ds.$$

The model (1) is generically nonlocal, particularly if the support of  $\underline{\omega}_\delta$  extends beyond the origin. It, at any  $x$ , involves function values of  $u$  at not only  $x$  but possibly its  $\delta$ -neighborhood. With  $\underline{\omega}_\delta = \underline{\omega}_\delta(s)$  a probability density function,  $\mathfrak{L}_\delta$  can be interpreted as a continuum average (integral) of the difference operator  $D_s^2$  over a continuum of scales  $s \in [0, \delta]$ . This interpretation has various implications as discussed below.

First, differential and discrete equations are special cases of nonlocal equations: let  $\underline{\omega}_\delta(s)$  be the Dirac-delta measure at either  $s = 0$  or  $h$ , we get  $\mathfrak{L}_0 = \frac{d^2}{dx^2}$  or  $\mathfrak{L}_h = D_h^2$  respectively, showing the generality of nonlocal continuum models. A better illustration is via a limiting process, e.g., for smooth  $u$ , small  $\delta$ , and  $\underline{\omega}_\delta$  going to the Dirac-delta at  $s = 0$ , we have

$$\mathfrak{L}_\delta u(x) = \frac{d^2 u}{dx^2}(x) \int_0^\delta \underline{\omega}_\delta(s) ds + c_2 \delta^2 \frac{d^4 u}{dx^4}(x) + \dots \approx \frac{d^2 u}{dx^2} = \mathfrak{L}_0 u(x)$$

showing that nonlocal models may resemble their local continuum limit for smooth quantities of interests (QoI), while encoding richer information for QoIs with singularity.

In addition, with a special class of fractional power-law kernel  $\underline{\omega}_\delta(s) = c_{\alpha,1} |s|^{1-2\alpha}$  for  $0 < \alpha < 1$  and  $\delta = \infty$ ,  $\mathfrak{L}_\delta$  leads to a fractional differential operator [Bucur and Valdinoci \[2016\]](#), [Caffarelli and Silvestre \[2007\]](#), [Nochetto, Otárola, and Salgado \[2016\]](#), [Vázquez \[2017\]](#), and [West \[2016\]](#):

$$\mathfrak{L}_\infty u(x) = c_{\alpha,1} \int_0^\infty D_s^2 u(x) |s|^{1-2\alpha} ds = \left( -\frac{d^2}{dx^2} \right)^\alpha u(x).$$

One may draw further connections from the Fourier symbols of these operators [Du \[n.d.\]](#) and [Du and K. Zhou \[2011\]](#).

Nonlocal models and operators have many variations and extensions. For example, one may define a nonlocal jump (diffusion) operator for a particle density  $u = u(x)$ ,

$$\mathfrak{L}_\delta u(x) = \int (\beta(x, y)u(y) - \beta'(y, x)u(x)) dy,$$

with  $\beta = \beta(x, y)$  and  $\beta' = \beta'(y, x)$  the jumping rates. We can recover (2) if  $\beta(x, y) = \beta'(y, x) = |x - y|^{-2} \omega_\delta(|x - y|)$ , and make connections with stochastic processes [Du, Huang, and Lehoucq \[2014\]](#).

Other extensions include systems for vector and tensor fields such as nonlocal models of mechanics. A representative example is the peridynamic theory [Silling \[2000\]](#) which attempts to offer a unified treatment of balance laws on and off materials discontinuities, see [Bobaru, Foster, Geubelle, and Silling \[2017\]](#) for reviews on various aspects of peridynamics. We briefly describe a simple linear small strain state-based peridynamic model here. Let  $\Omega$  be either  $\mathbb{R}^d$  or a bounded domain in  $\mathbb{R}^d$  with Lipschitz boundary, and  $\Omega^* = \Omega \cup \Omega_I$  where  $\Omega_I$  is an interaction domain. Let  $\mathbf{u} = \mathbf{u}(\mathbf{x}, t) = \mathbf{y}(\mathbf{x}, t) - \mathbf{x}$  denote the displacement field at the point  $\mathbf{x} \in \Omega^*$  and time  $t$  so that  $\mathbf{y} = \mathbf{x} + \mathbf{u}$  gives the deformed position, the peridynamic equation of motion can be expressed by

$$\rho \mathbf{u}_{tt}(\mathbf{x}, t) = \mathfrak{L}_\delta \mathbf{u}(\mathbf{x}, t) + \mathbf{b}(\mathbf{x}, t), \quad \forall \mathbf{x} \in \Omega, \quad t > 0,$$

where  $\rho$  is the constant density,  $\mathbf{b} = \mathbf{b}(\mathbf{x}, t)$  the body force, and  $\mathfrak{L}_\delta \mathbf{u}$  the interaction force derived from the variation of the nonlocal strain energy. Under a small strain assumption, for any  $\mathbf{x}, \mathbf{x}' = \mathbf{x} + \boldsymbol{\xi}$ , the linearized total strain and dilatational strain are given by

$$(3) \quad s(\mathbf{u})(\mathbf{x}', \mathbf{x}) := \mathbf{e}(\boldsymbol{\xi}) \cdot \frac{\boldsymbol{\eta}}{|\boldsymbol{\xi}|} \quad \text{and} \quad \mathfrak{d}_\delta(\mathbf{u})(\mathbf{x}) := \int \omega_\delta(\mathbf{x}', \mathbf{x}) s(\mathbf{u})(\mathbf{x}', \mathbf{x}) d\mathbf{x}'.$$

where  $\mathbf{e}(\boldsymbol{\xi}) = \boldsymbol{\xi}/|\boldsymbol{\xi}|$ ,  $\boldsymbol{\eta} = \mathbf{u}(\mathbf{x} + \boldsymbol{\xi}) - \mathbf{u}(\mathbf{x})$  and the kernel  $\omega_\delta$  has its support over a spherical neighborhood  $|\mathbf{x}' - \mathbf{x}| < \delta$  (with  $\delta$  being the horizon parameter) and is normalized by

$$\int \omega_\delta(\mathbf{x}', \mathbf{x}) d\mathbf{x}' = 1.$$

The linearized deviatoric strain is denoted by  $\mathfrak{S}_\delta(\mathbf{u})(\mathbf{x}', \mathbf{x}) := s(\mathbf{u})(\mathbf{x}', \mathbf{x}) - \mathfrak{d}_\delta(\mathbf{u})(\mathbf{x})$ . Then, the small strain quadratic last energy density functional is given by

$$(4) \quad \mathcal{W}_\delta(\mathbf{x}, \{\boldsymbol{\xi}, \boldsymbol{\eta}\}) = \kappa |\mathfrak{d}_\delta(\mathbf{u})(\mathbf{x})|^2 + \mu \int_{\Omega^*} \omega_\delta(\mathbf{x} + \boldsymbol{\xi}, \mathbf{x}) |s(\mathbf{u})(\mathbf{x} + \boldsymbol{\xi}, \mathbf{x}) - \mathfrak{d}_\delta(\mathbf{u})(\mathbf{x})|^2 d\boldsymbol{\xi}$$

where  $\kappa$  represents the peridynamic *bulk modulus* and  $\mu$  the peridynamic *shear modulus*.

For  $\kappa = \mu$ , we get a nondimensionalized bond-based peridynamic energy density [Mengesha and Du \[2014\]](#) and [Silling \[2000\]](#)

$$(5) \quad \mathcal{W}_\delta(\mathbf{x}, \{\boldsymbol{\xi}, \boldsymbol{\eta}\}) = \int \omega_\delta(\mathbf{x} + \boldsymbol{\xi}, \mathbf{x}) \left| \frac{\boldsymbol{\xi}}{|\boldsymbol{\xi}|} \cdot \frac{\mathbf{u}(\mathbf{x} + \boldsymbol{\xi}) - \mathbf{u}(\mathbf{x})}{|\boldsymbol{\xi}|} \right|^2 d\boldsymbol{\xi},$$

For a scalar function  $u = u(x)$ , we get a simple one dimensional energy density

$$\mathcal{W}_\delta(x, \{u\}) = \int \underline{\omega}_\delta(|y - x|) \frac{|u(y) - u(x)|^2}{|y - x|^2} dy,$$

associated with the nonlocal operator in (1), if a translation invariant and even kernel  $\underline{\omega}_\delta$  is adopted. This special case has often served as a benchmark problem for peridynamics, even though in most practical applications, peridynamic models do take on nonlinear vector forms to account for complex interactions [Du, Tao, and X. Tian \[2017\]](#).

### 3 Nonlocal vector calculus and nonlocal variational problems

We introduce the theory through an example, accompanied by some general discussions.

**A model equation.** The systematic development of the nonlocal vector calculus was originated from the study of peridynamics [Du, Gunzburger, Lehoucq, and K. Zhou \[2013\]](#). Let us consider a time-independent linear bond-based peridynamic model associated with the strain energy (5) given by

$$(6) \quad -\mathcal{L}_\delta \mathbf{u}(\mathbf{x}) = -2 \int_{\Omega^*} \omega_\delta(\mathbf{x} + \boldsymbol{\xi}, \mathbf{x}) \left[ \frac{\mathbf{u}(\mathbf{x} + \boldsymbol{\xi}) - \mathbf{u}(\mathbf{x})}{|\boldsymbol{\xi}|^2} \cdot \mathbf{e}(\boldsymbol{\xi}) \right] \mathbf{e}(\boldsymbol{\xi}) d\boldsymbol{\xi} = \mathbf{b}(\mathbf{x}), \quad \forall \mathbf{x} \in \Omega,$$

where  $\mathbf{u}$  is a displacement field and  $\mathbf{b}$  is a body force. Intuitively, (6) describes the force balance in a continuum body of linear springs, with the spring force aligned with the undeformed bond direction between any pair of points  $\mathbf{x}$  and  $\mathbf{x}' = \mathbf{x} + \boldsymbol{\xi}$ . This gives a nonlocal analog of classical linear elasticity model with a particular Poisson ratio, yet it does not, at the first sight, share the same elegant form of linear elasticity. Nonlocal vector calculus can make the connections between local and nonlocal models more transparent.

**Examples of nonlocal operators.** Let us introduce some nonlocal operators as illustrative examples. First, we define a nonlocal two-point gradient operator  $\mathcal{G}$  for any  $\mathbf{v}: \mathbb{R}^d \rightarrow \mathbb{R}^m$  such that  $\mathcal{G}\mathbf{v}: \Omega^* \times \Omega^* \rightarrow \mathbb{R}^{d \times n}$  is a two-point second-order tensor field given by,

$$(7) \quad (\mathcal{G}\mathbf{v})(\mathbf{x}', \mathbf{x}) = \mathbf{e}(\mathbf{x}' - \mathbf{x}) \otimes \frac{\mathbf{v}(\mathbf{x}') - \mathbf{v}(\mathbf{x})}{|\mathbf{x}' - \mathbf{x}|} \quad \text{where} \quad \mathbf{e}(\mathbf{x}' - \mathbf{x}) = \frac{\mathbf{x}' - \mathbf{x}}{|\mathbf{x}' - \mathbf{x}|}, \quad \forall \mathbf{x}', \mathbf{x} \in \Omega^*.$$

There are two cases of particular interests, namely,  $n = 1$  and  $n = d$ . In the latter case, we also define a nonlocal two-point divergence operator  $\mathcal{D}$  by

$$(8) \quad (\mathcal{D}\mathbf{v})(\mathbf{x}', \mathbf{x}) = \mathbf{e}(\mathbf{x}' - \mathbf{x}) \cdot \frac{\mathbf{v}(\mathbf{x}') - \mathbf{v}(\mathbf{x})}{|\mathbf{x}' - \mathbf{x}|} = \text{Tr}(\mathcal{G}\mathbf{v})(\mathbf{x}, \mathbf{x}'), \quad \text{for } n = d.$$

For peridynamics,  $(\mathcal{D}\mathbf{v})(\mathbf{x}', \mathbf{x})$  corresponds to the linearized strain described in (3).

Next, we define a nonlocal two-point dual divergence operator  $\mathcal{D}^*$  acting on any two-point scalar field  $\Psi: \Omega^* \times \Omega^* \rightarrow \mathbb{R}$  such that  $\mathcal{D}^*\Psi$  becomes a vector field given by,

$$(9) \quad (\mathcal{D}^*\Psi)(\mathbf{x}) = \int_{\Omega^*} (\Psi(\mathbf{x}, \mathbf{x}') + \Psi(\mathbf{x}', \mathbf{x})) \frac{\mathbf{e}(\mathbf{x}' - \mathbf{x})}{|\mathbf{x}' - \mathbf{x}|} d\mathbf{x}', \quad \forall \mathbf{x} \in \Omega^*.$$

We may interpret  $\mathcal{D}^*$  and  $\mathcal{D}$  as adjoint operators to each other in the sense that

$$(10) \quad \int_{\Omega^*} \mathbf{v}(\mathbf{x}) \cdot (\mathcal{D}^*\Psi)(\mathbf{x}) d\mathbf{x} = - \int_{\Omega^*} \int_{\Omega^*} (\mathcal{D}\mathbf{v})(\mathbf{x}', \mathbf{x}) \Psi(\mathbf{x}', \mathbf{x}) d\mathbf{x}' d\mathbf{x}$$

for all  $\mathbf{v}$  and  $\Psi$  that make integrals in (10) well defined. The duality may also be written more canonically as  $(\mathbf{v}, \mathcal{D}^*\Psi)_{\Omega^*} = -(\mathcal{D}\mathbf{v}, \Psi)_{\Omega^* \times \Omega^*}$  where  $(\cdot, \cdot)_{\Omega^*}$  and  $(\cdot, \cdot)_{\Omega^* \times \Omega^*}$  denote  $L^2$  inner products for vector and scalar fields in their respective domains of definition. Similarly, we can define a nonlocal two-point dual gradient operator  $\mathcal{G}^*$  acting on any two-point vector field  $\Psi: \Omega^* \times \Omega^* \rightarrow \mathbb{R}^d$  by the duality that  $(v, \mathcal{G}^*\Psi)_{\Omega^*} = -(\mathcal{G}v, \Psi)_{\Omega^* \times \Omega^*}$  for  $\mathcal{G}$  given by (7).

Some basic elements of nonlocal vector calculus are listed in Table 2 in comparison with the local counterpart. Discussions on concepts like the nonlocal flux and further justifications on labeling  $\mathcal{G}$  and  $\mathcal{D}$  as two-point gradient and divergence can be found in Du [n.d.] and Du, Gunzburger, Lehoucq, and K. Zhou [2013].

Newton's vector calculus	$\Leftrightarrow$	Nonlocal vector calculus
Differential operators, local flux	$\Leftrightarrow$	Nonlocal operators, nonlocal flux
Green's identity, integration by parts		Nonlocal Green's identity (duality)
$\int_{\Omega} \mathbf{u} \cdot \Delta \mathbf{v} - \mathbf{v} \cdot \Delta \mathbf{u} = \int_{\partial\Omega} \mathbf{u} \cdot \partial_n \mathbf{v} - \mathbf{v} \cdot \partial_n \mathbf{u}$	$\Leftrightarrow$	$\iint_{\Omega^* \times \Omega^*} \mathbf{u} \cdot \mathcal{D}^*(\mathcal{D}\mathbf{v}) - \mathbf{v} \cdot \mathcal{D}^*(\mathcal{D}\mathbf{u}) = 0$

Table 1: Elements of vector calculus: local versus nonlocal.

**Reformulation of nonlocal models.** Let the kernel in (6)  $\omega_\delta = \omega_\delta(\mathbf{x}', \mathbf{x}) = \omega_\delta(\mathbf{x}, \mathbf{x}')$  be symmetric. We consider  $\mathcal{D}^*\Psi$  with  $\Psi(\mathbf{x}', \mathbf{x}) = \omega_\delta(\mathbf{x}', \mathbf{x})(\mathcal{D}\mathbf{u})(\mathbf{x}', \mathbf{x})$ . This leads to

$$-\mathcal{L}_\delta \mathbf{u} = -\mathcal{D}^*(\omega_\delta \mathcal{D}\mathbf{u}) = \mathbf{b}$$

a concise reformation of (6) that starts to resemble, in appearance, the PDE form of classical elasticity, with local differential (gradient and divergent) operators replaced by their nonlocal counterparts. Analogously, a scalar nonlocal diffusion equation for a translation

invariant  $\omega_\delta$ , i.e.,  $\omega_\delta(\mathbf{x}', \mathbf{x}) = \underline{\omega}_\delta(\mathbf{x}' - \mathbf{x}) = \underline{\omega}_\delta(\mathbf{x} - \mathbf{x}')$ , and its reformulation can be given by

(11)

$$-\mathcal{L}_\delta v(\mathbf{x}) = -\int_{\Omega^*} \underline{\omega}_\delta(\xi) \frac{v(\mathbf{x} + \xi) - 2v(\mathbf{x}) + v(\mathbf{x} - \xi)}{|\xi|^2} d\xi = f(\mathbf{x}) \Leftrightarrow -\mathcal{G}^*(\underline{\omega}_\delta \mathcal{G}v) = f.$$

similar to the one-dimensional version (2). Moreover, not only nonlocal models can be nicely reformulated like classical PDEs, their mathematical theory may also be developed in a similar fashion along with interesting new twists [Du, Gunzburger, Lehoucq, and K. Zhou \[2013, 2012\]](#) and [Mengesha and Du \[2014, 2015\]](#).

**Variational problems.** Let  $\Omega \subset \mathbb{R}^d$  be a bounded domain with Lipschitz boundary and  $\Omega_I$  be where constraints on the solution are imposed. We consider an energy functional

$$(12) \quad E_\delta(\mathbf{u}) := \frac{1}{2} |\mathbf{u}|_{\mathbf{g}_\delta^2}^2 - (\mathbf{b}, \mathbf{u})_{\Omega^*} \quad \text{with} \quad |\mathbf{u}|_{\mathbf{g}_\delta^2}^2 := \iint_{\Omega^* \times \Omega^*} \underline{\omega}_\delta(\mathbf{x}' - \mathbf{x}) (\mathcal{D}\mathbf{u}(\mathbf{x}', \mathbf{x}))^2 d\mathbf{x}' d\mathbf{x}$$

for a prescribed body force  $\mathbf{b} = \mathbf{b}(\mathbf{x}) \in L^2(\Omega_\delta^*)^d$  and a kernel  $\underline{\omega}_\delta = \underline{\omega}_\delta(\xi)$  satisfying

$$(13) \quad \begin{cases} \underline{\omega}_\delta(\xi) \geq 0 \text{ is radially symmetric, } B_{\sigma\delta}(\mathbf{0}) \subset \text{supp}(\underline{\omega}_\delta) \subset B_\delta(\mathbf{0}) \subset \mathbb{R}^d \\ \text{for } 0 < \sigma < 1, \text{ and } \int_{B_\delta(\mathbf{0})} \underline{\omega}_\delta(\xi) d\xi = 1. \end{cases}$$

Let  $\mathcal{S}_\delta^2$  be the set of  $\mathbf{u} \in L^2(\Omega^*)^d$  with  $\|\mathbf{u}\|_{\mathbf{g}_\delta^2}^2 = \|\mathbf{u}\|_{L^2(\Omega^*)^d}^2 + |\mathbf{u}|_{\mathbf{g}_\delta^2}^2$  finite, which is a separable Hilbert space with an inner product induced by the norm  $\|\cdot\|_{\mathbf{g}_\delta^2}$  [Mengesha and Du \[2014\]](#). For a weakly closed subspace  $V \subset L^2(\Omega^*)^d$  that has no nontrivial affine maps with skew-symmetric gradients, we let  $V_{c,\delta} = \mathcal{S}_\delta^2 \cap V$ . One can establish a compactness result on  $V_{c,\delta}$  [Mengesha and Du \[2014\]](#) and [Mengesha \[2012\]](#):

**Lemma 3.1.** *For a bounded sequence  $\{\mathbf{u}_n\} \in V_{c,\delta}$ ,  $\lim_{n \rightarrow \infty} |\mathbf{u}_n|_{\mathbf{g}_\delta^2} = 0$  gives  $\|\mathbf{u}_n\|_{L^2(\Omega^*)} \rightarrow 0$ .*

This leads to a nonlocal Poincaré inequality and the coercivity of the energy functional.

**Proposition 3.2** (Nonlocal Poincaré). *There exists a positive constant  $C$  such that*

$$\|\mathbf{u}\|_{L^2(\Omega^*)^d} \leq C |\mathbf{u}|_{\mathbf{g}_\delta^2}, \quad \forall \mathbf{u} \in V_{c,\delta}.$$

The well-posedness of the variational problem then follows [Mengesha and Du \[2014\]](#). Moreover, one can get a uniform Poincaré constant, independent of  $\delta$  as  $\delta \rightarrow 0$ , if the nonlocal interaction kernels behave like a Dirac-delta sequence. More specifically, they satisfy that

$$(14) \quad \lim_{\delta \rightarrow 0} \int_{|\xi| > \epsilon} \underline{\omega}_\delta(\xi) d\xi = 0, \quad \forall \epsilon > 0.$$

The assumption is particularly true for a rescaled kernel  $\underline{\omega}_\delta(\xi) = \delta^{-d} \underline{\omega}(\xi/\delta)$  Du [n.d.] and Mengesha and Du [2014].

Note that the above line of analysis can be carried out by extending similar results for the scalar function spaces originated from the celebrated work Bourgain, Brezis, and Mironescu [2001] and further studied in Ponce [2004]. One complication for vector fields is that the energy seminorm only uses a projected difference  $\mathcal{D}\mathbf{u}$  instead of the total difference, see Mengesha [2012] and Mengesha and Du [2014, 2015, 2016] for detailed discussions.

The nonlocal Poincaré inequality and energy coercivity then imply a well-posed variational formulation of the nonlocal model through minimizing  $E_\delta(\mathbf{u})$  over  $\mathbf{u} \in V_{c,\delta}$ . The weak form of the Euler–Lagrange equation is given by

$$B_\delta(\mathbf{u}, \mathbf{v}) := (\underline{\omega}_\delta \mathcal{D}\mathbf{u}, \mathcal{D}\mathbf{v})_{\Omega^* \times \Omega^*} = (\mathbf{u}, \mathbf{b})_{\Omega^*}, \quad \forall \mathbf{v} \in V_{c,\delta}.$$

We note a special case with  $V_{c,\delta} = \mathcal{S}_{0,\delta}$ , the closure of  $C_0^\infty(\Omega)^d$  in  $\mathcal{S}_2^\delta$  with all of its elements satisfying  $\mathbf{u}(\mathbf{x}) = \mathbf{0}$  on  $\Omega_I = \Omega_\delta = \{\mathbf{x} \in \mathbb{R}^d \setminus \Omega \mid \text{dist}(\mathbf{x}, \Omega) < \delta\}$ , corresponding to a problem with a homogeneous nonlocal Dirichlet constraint on a  $\delta$ -layer around  $\Omega$ , see Figure 1.

$$(15) \quad \begin{aligned} -\mathcal{L}_\delta \mathbf{u} &= -\mathcal{D}^*(\underline{\omega}_\delta \mathcal{D})\mathbf{u} = \mathbf{b}, \quad \text{in } \Omega, \\ \mathbf{u} &= \mathbf{0}, \quad \text{in } \Omega_I = \Omega_\delta. \end{aligned} \quad \xrightarrow{\delta \rightarrow 0} \quad \begin{aligned} \mathbf{u} &= \mathbf{0} \text{ on } \partial\Omega \\ -\mathcal{L}_0 \mathbf{u} &= \mathbf{b} \quad \text{in } \Omega \end{aligned}$$

Figure 1: A nonlocal constrained value problem and its local PDE limit

Furthermore, under the assumption (14), we can show that as  $\delta \rightarrow 0$ , the solution of (15), denoted by  $\mathbf{u}_\delta$ , converges in  $L^2(\Omega)$  to the solution  $\mathbf{u}_0 \in H_0^1(\Omega)$  of the equation  $-\mathcal{L}_0 \mathbf{u} = -(d+2)^{-1}(\Delta \mathbf{u} + 2\nabla(\nabla \cdot \mathbf{u})) = \mathbf{b}$  in  $\Omega$  Mengesha and Du [2014], thus compatible with linear elasticity.

**Elements of mathematical foundation of nonlocal models.** Without going into details, we summarize some basic elements in Table 2. For brevity, the illustration is devoted to the case of nonlocal diffusion model and its local limit, with  $\underline{K}$  denoting a generic 2nd order positive definite coefficient tensor, and the same notation  $(\cdot, \cdot)$  for  $L^2$  inner products of scalar, vector and tensor fields over their respective domains.

Local variational problems	$\Leftrightarrow$	Nonlocal variational problems
Local energy: $(\nabla u, \nabla u)$	$\Leftrightarrow$	Nonlocal energy: $(\omega_\delta \mathcal{G}u, \mathcal{G}u)$
Sobolev space $H^1(\Omega)$	$\Leftrightarrow$	Nonlocal function space $\mathcal{S}_2^\delta$
Local balance (PDE): $\left\{ \begin{array}{l} -\nabla \cdot (\underline{K} \nabla u) = f \end{array} \right\}$	$\Leftrightarrow$	Nonlocal balance: $\left\{ \begin{array}{l} -\mathcal{G}^*(\omega_\delta \mathcal{G}u) = f \end{array} \right\}$
Boundary conditions on $\partial\Omega$	$\Leftrightarrow$	Volumetric constraints on $\Omega_I$
Local weak forms: $\left\{ \begin{array}{l} (\underline{K} \nabla u, \nabla u) = (f, v), \forall v \end{array} \right\}$	$\Leftrightarrow$	Nonlocal weak forms: $\left\{ \begin{array}{l} (\omega_\delta \mathcal{G}u, \mathcal{G}v) = (f, v), \forall v \end{array} \right\}$
Classical Poincaré: $\ u\ _{L^2} \leq c \ \nabla u\ _{L^2}$	$\Leftrightarrow$	Nonlocal Poincaré: $\ u\ _{L^2} \leq c \ u\ _{\mathcal{S}_2^\delta}$

Table 2: Elements of variational problems: local versus nonlocal.

**Other variants of nonlocal operators and nonlocal calculus.** As part of the nonlocal vector calculus, there are other on possible variants to the nonlocal operators introduced here, e.g., the one-point nonlocal divergence and nonlocal dual gradient given by

$$\mathcal{D}_\rho \mathbf{v}(\mathbf{x}) = \int_\Omega \rho_\delta(\mathbf{x}' - \mathbf{x}) (\mathcal{D}\mathbf{v})(\mathbf{x}', \mathbf{x}) d\mathbf{x}', \quad \mathcal{G}_\rho^* \mathbf{v}(\mathbf{x}) = \int_\Omega \rho_\delta(\mathbf{x}' - \mathbf{x}) \mathbf{e}(\mathbf{x}' - \mathbf{x}) \otimes \frac{\mathbf{v}(\mathbf{x}') + \mathbf{v}(\mathbf{x})}{|\mathbf{x}' - \mathbf{x}|} d\mathbf{x}'$$

for an averaging kernel  $\rho_\delta$ . With  $\rho_\delta$  approaching a Dirac-delta measure at the origin as  $\delta \rightarrow 0$ ,  $\mathcal{D}_\rho$  and  $\mathcal{G}_\rho^*$  recover the conventional local divergence and gradient operators Du, Gunzburger, Lehoucq, and K. Zhou [2013] and Mengesha and Du [2016]. They also form a duality pair and have been used for robust nonlocal gradient recovery Du, Tao, X. Tian, and J. Yang [2016]. Their use in the so-called correspondence peridynamic materials models could be problematic but more clarifications have been given recently Du and X. Tian [n.d.]. Moreover, these one-point operators are needed to reformulate more general state-based peridynamics Du, Gunzburger, Lehoucq, and K. Zhou [2013] and Mengesha and Du [2015, 2016] where the equation of motion is often expressed by Silling [2010], Silling, Epton, Weckner, Xu, and Askari [2007], and Silling and Lehoucq [2010]

$$\rho \mathbf{u}_{tt} = \int \{ \underline{\mathbb{T}}[\mathbf{x}, \mathbf{u}] \langle \mathbf{x}' - \mathbf{x} \rangle - \underline{\mathbb{T}}[\mathbf{x}', \mathbf{u}] \langle \mathbf{x} - \mathbf{x}' \rangle \} d\mathbf{x}'$$

with  $\underline{\mathbb{T}}[\mathbf{x}, \mathbf{u}] \langle \mathbf{x}' - \mathbf{x} \rangle$  and  $\underline{\mathbb{T}}[\mathbf{x}', \mathbf{u}] \langle \mathbf{x} - \mathbf{x}' \rangle$  denoting the peridynamic force states. In fact,  $\mathcal{D}_\rho \mathbf{u}$  can be used to represent the linear dilational strain in (3). Thus, we once again see that the study of nonlocal models of mechanics further enriches the mathematical theory of nonlocal operators and makes nonlocal vector calculus highly relevant to applications.



## 4 Numerical discretization of nonlocal models

There are many ways to discretize nonlocal models [Du \[2017a\]](#), such as mesh-free [Bessa, Foster, Belytschko, and Liu \[2014\]](#), [Parks, Seleson, Plimpton, Silling, and Lehoucq \[2011\]](#), and [Parks, Littlewood, Mitchell, and Silling \[2012\]](#), quadrature based difference or collocation [Seleson, Du, and Parks \[2016\]](#), [H. Tian, H. Wang, and W. Wang \[2013\]](#), and [X. Zhang, Gunzburger, and Ju \[2016a,b\]](#), finite element [Tao, X. Tian, and Du \[2017\]](#), [H. Tian, Ju, and Du \[2017\]](#), and [K. Zhou and Du \[2010\]](#) and spectral methods [Du and J. Yang \[2017\]](#). In particular, finite difference, finite element and collocation schemes in one dimension were considered in [X. Tian and Du \[2013\]](#), including comparisons and analysis of the differences and similarities. Discontinuous Galerkin approximations have also been discussed, including conforming DG [Chen and Gunzburger \[2011\]](#) and [Ren, C. Wu, and Askari \[2017\]](#) nonconforming DG [X. Tian and Du \[2015\]](#) and local DG [Du, Ju, and Lu \[2018\]](#).

Since nonlocal models are developed as alternatives when conventional continuum PDEs can neither capture the underlying physics nor have meaningful mathematical solutions, we need to place greater emphasis on verification and validation of results from the more tortuous simulations. A common practice for code verification is to consider the case where the nonlocal models can lead to a physically valid and mathematically well-defined local limit on the continuum level and to check if one can numerically reproduce solutions of the local limit by solving nonlocal models with the same given data. Such popular benchmark tests may produce surprising results as discussed here.

**Asymptotical compatibility.** Addressing the consistency on both continuum and discrete levels and ensuring algorithmic robustness have been crucial issues for modeling and code development efforts, especially for a theory like peridynamics that is developed to capture highly complex physical phenomena. In the context of nonlocal models and their local limits, the issues on various convergent paths are illustrated in the diagram shown in [Figure 2 X. Tian and Du \[2014\]](#) (with smaller discretization parameter  $h$  representing finer resolution).

The paths along the diagram edges are for taking limit in one of the parameters while keeping the other fixed: ♠ shows the convergence of solutions of nonlocal continuum models to their local limit as  $\delta \rightarrow 0$ , which has been established for various linear and nonlinear problems; ♣ is a subject of numerical PDE; ◇ assures a convergent discretization to nonlocal problem by design; ♥ is more intriguing, as it is not clear whether the local limit of numerical schemes for nonlocal problems would remain an effective scheme for the local limit of the continuum model. An affirmative answer would lead to a nice *commutative diagram*, or *asymptotic compatibility* (AC) [X. Tian and Du \[ibid.\]](#), one can follow

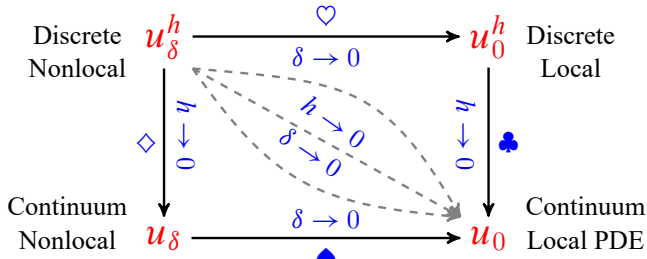


Figure 2: A diagram of possible paths between  $u_\delta$ ,  $u_\delta^h$ ,  $u_0^h$  and  $u_0$  via various limits.

either the paths through those marked with  $\diamond$  and  $\spadesuit$  or ones marked with  $\heartsuit$  and  $\clubsuit$  to get the convergence of  $u_\delta^h$  to  $u_0$ .

AC schemes offer robust and convergent discrete approximations to parameterized problems and preserve the correct limiting behavior. While the variational characterization and framework are distinctive, they are reminiscent in spirit to other studies of convergent approximations in the limiting regimes, see for example [Arnold and Brezzi \[1997\]](#), [Guermond and Kanschat \[2010\]](#), and [Jin \[1999\]](#).

**Getting wrong solution from a convergent numerical scheme.** To motivate the AC schemes, we consider a 1d linear nonlocal problem  $-\mathcal{L}_\delta u_\delta(x) = b(x)$  for  $x \in (0, 1)$ , where  $\mathcal{L}_\delta$  is given by (2) with a special kernel, i.e.,

$$(16) \quad \mathcal{L}_\delta u(x) = \frac{3}{\delta^3} \int_0^\delta (u(x+s) - 2u(x) + u(x-s)) ds = \frac{3}{\delta^3} \int_0^\delta h^2 D_h^2 u(x) dh,$$

We impose the constraint that  $u_\delta(x) = u_0(x)$  for  $x \in (-\delta, 0) \cup (1, 1+\delta)$  where  $u_0$  solves the local limiting problem  $-u_0''(x) = b(x)$  in  $\mathbb{R}$ . On the continuum level, we have  $u_\delta \rightarrow u_0$  as  $\delta \rightarrow 0$  in the appropriate function spaces, as desired. For (16), we may obtain a discrete system if we replace the *continuum difference*  $\mathcal{L}_\delta$  by discrete finite differences through suitable quadrature approximations (leading to the *quadrature based finite difference* discretization as named in [X. Tian and Du \[2013\]](#)). For example, following [Du and X. Tian \[2015\]](#) and [X. Tian and Du \[2013\]](#), we consider a scheme for (16) obtained from a Riemann sum quadrature: for  $1 \leq i \leq N = 1/h$ ,  $\delta = rh$ ,

$$(17) \quad -\mathcal{L}_\delta^h u_i = -\frac{3h}{\delta^3} \sum_{m=1}^r (D_{mh}^2 u)_i = b(x_i),$$

where  $\{u_i\}$  are approximations of  $\{u(x_i)\}$  at nodal points  $\{x_i = ih\}_{i=-r}^{N+r}$ . For any given  $\delta > 0$ , we can show the convergence of the discretization as  $h \rightarrow 0$  for any given  $\delta$  by combining both stability with consistency estimates [X. Tian and Du \[2013\]](#). However, by

considering a special case with  $r = 1$  in (17), we end up with a scheme  $-3(D_h^2 u)_i = b_i$ , which converges to the differential equation  $-3u''(x) = b(x)$  as  $h = \delta \rightarrow 0$ , but not to the correct local limit. In other words, if we set  $h$  and  $\delta$  to zero proportionally, the numerical solution of the discrete scheme for the nonlocal problem yields a convergent approximation to a *wrong* local limit associated with, unfortunately, a consistently over-estimated elastic constant!

The possibility of numerical approximations converging to a wrong solution is alarming; if without prior knowledge, such convergence might be mistakenly used to verify or disapprove numerical simulation, and we see the risks involved due to the wrong local limits produced by discrete solutions to nonlocal models. Although illustrated via a simple example here, it has been shown to be a generic feature of discretizations represented by (17) and other schemes such as the piecewise constant Galerkin finite element approximations, for scalar nonlocal diffusion models and general state-based peridynamic systems [Du and X. Tian \[2015\]](#) and [X. Tian and Du \[2013, 2014\]](#).

**Robust discretization via AC schemes.** On a positive note, the complications due to the use of discrete schemes like (17) can be resolved through other means. For example, it is proposed in [X. Tian and Du \[2013\]](#) that an alternative formulation works much more robustly by suitably adjusting the weights for the second order differences  $\{D_{mh}^2 u\}$  so that the elastic constant always maintain its correct constant value 1, independently of  $r$ ! Hence, as shown in [X. Tian and Du \[ibid.\]](#), we have a scheme that is convergent to the nonlocal model for any fixed  $\delta$  as  $h \rightarrow 0$  and to the correct local limit whenever  $\delta \rightarrow 0$  and  $h \rightarrow 0$  simultaneously, regardless how the two parameters are coupled. Moreover, for a fixed  $h$ , it recovers the standard different scheme for the correct local limit models as  $\delta \rightarrow 0$ . Thus, we have a robust numerical approximation that is free from the risk of going to the wrong continuum solution. Naturally, it is interesting to characterize how such schemes can be constructed in general.

**Quadrature based finite difference AC scheme.** Approximations for multidimensional scalar nonlocal diffusion equations have been developed [Du, Tao, X. Tian, and J. Yang \[n.d.\]](#), which are not only AC but also preserve the discrete maximum principle. We consider a set of nodes (grid points)  $\{\mathbf{x}_j\}$  of a uniform Cartesian mesh with a mesh size  $h$  and a multi-index  $\mathbf{j}$  corresponding to  $\mathbf{x}_j = h\mathbf{j}$ . It is natural to approximate the nonlocal operator in (11) by

$$(18) \quad \mathcal{L}_\delta u(\mathbf{x}_i) \approx \int \mathfrak{L}_h \left( \frac{u(\mathbf{x}_i + \mathbf{z}) - 2u(\mathbf{x}_i) + u(\mathbf{x}_i - \mathbf{z})}{|\mathbf{z}|^2 W(\mathbf{z})} \right) W(\mathbf{z}) \omega_\delta(\mathbf{z}) d\mathbf{z},$$

where  $\mathfrak{L}_h$  represents the piecewise  $d$ -multi-linear interpolation operator in  $\mathbf{z}$  associated with the uniform Cartesian mesh  $\{\mathbf{x}_j = h\mathbf{j}\}$ , but the key that is crucial for the AC property

and the discrete maximum principle is to choose a properly defined nonnegative weight  $W = W(\mathbf{z})$ . The choice adopted in [Du, Tao, X. Tian, and J. Yang \[ibid.\]](#) corresponds to  $W(\mathbf{z}) = 1/|\mathbf{z}|_1$  where  $|\mathbf{z}|_1$  denotes the  $\ell_1$  norm in  $\mathbb{R}^d$ . This particular weight makes the quadrature exact for all quadratic functions. One can then show, through a series of technical calculations, that the resulting numerical solution converges to the solution of the nonlocal model on the order of  $O(h^2)$  for a fixed  $\delta > 0$ , and converges to that of the local limit model on the order of  $O(\delta^2 + h^2)$  as both  $h, \delta \rightarrow 0$  simultaneously, thus demonstrating the AC property.

**AC finite element approximations.** For multidimensional systems, one can extend, as in [X. Tian and Du \[2014\]](#), to more general abstract settings using conforming Galerkin finite element (FE) methods on unstructured meshes. In particular, the concept and theory of asymptotically compatible schemes are introduced for general parametrized variational problems. A special application is to pave a way for identifying robust approximations to linear nonlocal models that are guaranteed to be consistent in the local limit. Specifically, we have the following theorem that agrees with numerical experiments reported in the literature [Bobaru, M. Yang, Alves, Silling, Askari, and Xu \[2009\]](#) and [X. Tian and Du \[2013\]](#).

**Theorem 4.1.** *Let  $\mathbf{u}_\delta$  be the solution of (15) and  $\mathbf{u}_{\delta,h}$  be the conforming Galerkin FE approximation on a regular quasi-uniform mesh with meshing parameter  $h$ . If the FE space  $V_{\delta,h}$  contains all continuous piecewise linear elements, then  $\|\mathbf{u}_{\delta,h} - \mathbf{u}_0\|_{L^2(\Omega)} \rightarrow 0$  as  $\delta \rightarrow 0$  and  $h \rightarrow 0$ . If in addition, the FE subspace is given by a conforming FE space of the local limit PDE model with zero extension outside  $\Omega$  with  $u_{0,h}$  being the FE solution, then on each fixed mesh,  $\|u_{\delta,h} - u_{0,h}\|_{L^2} \rightarrow 0$  as  $\delta \rightarrow 0$ . On the other hand, if  $V_{\delta,h}$  is the piecewise constant space and conforming for (15), then  $\|u_{\delta,h} - u_0\|_{L^2} \rightarrow 0$  if  $h = o(\delta)$  as  $\delta \rightarrow 0$ .*

The above theorem, proved under minimal solution regularity, remains valid for nonlocal diffusion and state-based peridynamic models. The same framework of AC schemes can establish the convergence of numerical approximation to linear fractional diffusion equations (that correspond to  $\delta = \infty$ ) via the approximation of a nonlocal diffusion model with a finite horizon [X. Tian, Du, and Gunzburger \[2016\]](#). For example, consider a scalar fractional diffusion model, for  $\alpha \in (0, 1)$ ,

$$(-\Delta)^\alpha u = f, \text{ on } \Omega, \quad u = 0, \text{ on } \mathbb{R}^d \setminus \Omega, \quad (-\Delta)^\alpha u(\mathbf{x}) = C_{d,\alpha} \int_{\mathbb{R}^d} \frac{u(\mathbf{x}) - u(\mathbf{x}')}{|\mathbf{x} - \mathbf{x}'|^{d+2\alpha}} d\mathbf{x}',$$

and  $C_{d,\alpha}$  is a positive constant dependent on  $d$  and  $\alpha$ . We have that [X. Tian, Du, and Gunzburger \[ibid.\]](#),

**Theorem 4.2.** *Let  $u_\delta$  be the solution of the above fractional diffusion model with the integral truncated to a spherical neighborhood of radius  $\delta > 0$ . Let  $u_\delta^h$  be a conforming Galerkin FE approximation with the discretization parameter  $h$ , then  $\|u_\delta^h - u_\delta\|_{H^\alpha} \rightarrow 0$  as  $h \rightarrow 0$  for any given  $\delta$  and  $\|u_\delta^h - u_\infty\|_{H^\alpha} \rightarrow 0$  as  $\delta \rightarrow \infty$  and  $h \rightarrow 0$ .*

We note that studies of AC schemes have been extended to nonconforming DG FE [X. Tian and Du \[2015\]](#), local DG FE [Du, Ju, and Lu \[2018\]](#), spectral approximation [Du and J. Yang \[2016\]](#) and nonlocal gradient recoveries [Du, Tao, X. Tian, and J. Yang \[2016\]](#). There were also extensions to nonlinear nonlocal models [Du and Huang \[2017\]](#) and [Du and J. Yang \[2016\]](#).

## 5 Nonlocal and local coupling

Nonlocal models can be effective alternatives to local models by accommodating singular solutions, which makes nonlocal models particularly useful to subjects like fracture mechanics. Yet treating nonlocality in simulations may incur more computation. Thus, exploring localization and effective coupling of nonlocal and local models can be helpful in practice. Nevertheless, nonlocal models, unlike local PDEs, generically do not employ local boundary or interface conditions imposed on a co-dimension-1 surface, hence motivating the development of different approaches for local-nonlocal coupling [Li and Lu \[2017\]](#) and [Du, Tao, and X. Tian \[2018\]](#).

**Heterogeneous localization.** A particular mathematical quest for a coupled local and nonlocal model is through heterogeneous localization, as initiated in [X. Tian and Du \[2017\]](#).

The aim is to characterize subspaces of  $L^2(\Omega)$ , denoted by  $\mathcal{S}(\Omega)$ , that are significantly larger than  $H^1(\Omega)$  and have a continuous trace map into  $H^{1/2}(\partial\Omega)$ . One such example is defined as the completion of  $C^1(\Omega)$  with respect to the nonlocal norm for a kernel  $\gamma_\delta$ ,

$$\|u\|_{\mathcal{S}(\Omega)} = \left( \|u\|_{L^2(\Omega)}^2 + |u|_{\mathcal{S}(\Omega)}^2 \right)^{\frac{1}{2}}, \quad \text{with}$$

$$|u|_{\mathcal{S}(\Omega)}^2 = \int_{\Omega} \int_{\Omega \cap B_\delta(\mathbf{x})} \gamma_\delta(\mathbf{x}, \mathbf{y}) \frac{(u(\mathbf{y}) - u(\mathbf{x}))^2}{|\mathbf{y} - \mathbf{x}|^2} d\mathbf{y} d\mathbf{x}.$$

The main findings of [X. Tian and Du \[ibid.\]](#) are that the trace map exists and is continuous on a nonlocal function space  $\mathcal{S}(\Omega)$  if the radius of the support of  $\gamma_\delta$ , i.e., the horizon, is heterogeneously localized as  $\mathbf{x} \rightarrow \partial\Omega$ . By considering such a class of kernels, the study departs from many existing works, such as [Bourgain, Brezis, and Mironescu \[2001\]](#), corresponding to typical translate-invariant kernels. In [X. Tian and Du \[2017\]](#), the class of

kernels under consideration is given by

$$(19) \quad \gamma(\mathbf{x}, \mathbf{y}) = \frac{1}{|\delta(\mathbf{x})|^d} \hat{\gamma} \left( \frac{|\mathbf{y} - \mathbf{x}|}{\delta(\mathbf{x})} \right)$$

where  $\hat{\gamma} = \hat{\gamma}(s)$  is a non-increasing nonnegative function defined for  $s \in (0, 1)$  with a finite  $d - 1$  moment. The heterogeneously defined horizon  $\delta = \delta(\mathbf{x})$  approaches zero when  $\mathbf{x} \rightarrow \Gamma \subset \partial\Omega$ . A simple choice taken in [X. Tian and Du \[ibid.\]](#) is  $\delta(\mathbf{x}) = \sigma \operatorname{dist}(\mathbf{x}, \Gamma)$  for  $\sigma \in (0, 1]$ .

The following proposition has been established in [X. Tian and Du \[ibid.\]](#), which is of independent interests by showing the continuous imbedding of classical Sobolev space  $H^1(\Omega)$  in the new heterogeneously localized nonlocal space  $\mathcal{S}(\Omega)$ . The result generalizes a well-known result of [Bourgain, Brezis, and Mironescu \[2001\]](#) for the case with a constant horizon and translation invariant kernel.

**Proposition 5.1.** *For the kernel in (19) and the horizon  $\delta(\mathbf{x}) = \sigma \operatorname{dist}(\mathbf{x}, \Gamma)$  with  $\sigma \in (0, 1)$ ,  $H^1(\Omega)$  is continuously imbedded in  $\mathcal{S}(\Omega)$  and for any  $u \in H^1(\Omega)$ ,  $\|u\|_{\mathcal{S}(\Omega)} \leq C \|u\|_{H^1(\Omega)}$  where the constant  $C = C(\Omega)$  is independent of  $\sigma$  for  $\sigma$  small.*

**New trace theorems.** A key observation proved in [X. Tian and Du \[2017\]](#) is that, with heterogeneously vanishing interaction neighborhood when  $\mathbf{x} \rightarrow \partial\Omega$ , we expect a well defined continuous trace map from the nonlocal space  $\mathcal{S}(\Omega)$ , which is larger than  $H^1(\Omega)$ , to  $H^{1/2}(\partial\Omega)$ .

**Theorem 5.2** (General trace theorem). *Assume that  $\Omega$  is a bounded simply connected Lipschitz domain in  $\mathbb{R}^d$  ( $d \geq 2$ ) and  $\Gamma = \partial\Omega$ , for a kernel in (19) and the heterogeneously defined horizon given by  $\delta(\mathbf{x}) = \sigma \operatorname{dist}(\mathbf{x}, \Gamma)$  for  $\sigma \in (0, 1]$ . there exists a constant  $C$  depending only on  $\Omega$  such that the trace map  $T$  for  $\Gamma$  satisfies  $\|Tu\|_{H^{\frac{1}{2}}(\Gamma)} \leq C \|u\|_{\mathcal{S}(\Omega)}$ , for any  $u \in \mathcal{S}(\Omega)$ .*

By Proposition 5.1, we see that the above trace theorem is indeed a refinement of the classical trace theorem in the space  $H^1(\Omega)$ , with the latter being a simple consequence.

**An illustrative example with a simple kernel on a stripe domain.** A complete proof of the trace Theorem 5.2 is presented in [X. Tian and Du \[ibid.\]](#). To help understanding what the result conveys and how it compares with other relevant works, it is suggestive to consider a special case.

For  $\Omega$  and  $\Gamma$ , we take a special stripe domain  $\Omega = (0, r) \times \mathbb{R}^{d-1}$  and a portion of its boundary  $\Gamma = \{0\} \times \mathbb{R}^{d-1}$  for a constant  $r > 0$ , see equation (20) and Figure 3.

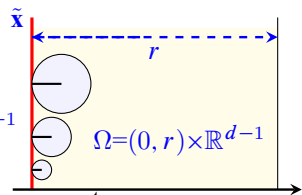
$$(20) \quad \begin{cases} \gamma(\mathbf{x}, \mathbf{y}) = \frac{\chi_{(0,1)}(|\mathbf{y} - \mathbf{x}|)|\mathbf{y} - \mathbf{x}|^2}{|\delta(\mathbf{x})|^{d+2}}, \\ \text{where } \delta(\mathbf{x}) = \text{dist}(\mathbf{x}, \Gamma) = x_1, \\ \forall \mathbf{x} = (x_1, \tilde{\mathbf{x}}), \tilde{\mathbf{x}} \in \mathbb{R}^{d-1}. \end{cases}$$


Figure 3: Nonlocal kernel and depiction of the stripe geometry.

This case serves as not only a helpful step towards proving the more general trace [Theorem 5.2](#) but also an illustrative example on its own. Indeed, this special nonlocal (semi)-norm is

$$(21) \quad |u|_{\mathcal{S}(\Omega)}^2 = \int_{\Omega} \int_{\Omega \cap \{|\mathbf{y} - \mathbf{x}| < |x_1|\}} \frac{(u(\mathbf{y}) - u(\mathbf{x}))^2}{|x_1|^{2+d}} d\mathbf{y} d\mathbf{x}.$$

Clearly, the denominator  $x_1$  penalizes the spatial variation only at  $x_1 = 0$ , thus  $\mathcal{S}(\Omega)$  contains all functions in  $L^2(\bar{\Omega})$  (and possibly discontinuous) for any domain  $\bar{\Omega}$  with its closure being a compact subset of  $\Omega$ . Hence, functions in  $\mathcal{S}(\Omega)$  are generally not expected to have regularity better than  $L^2(\Omega')$  over any strict subdomain  $\Omega'$ . Yet, as elucidated in [X. Tian and Du \[2017\]](#), due to the horizon localization at the boundary, the penalization of spatial variations provides enough regularity for the functions in  $\mathcal{S}(\Omega)$  to have well-defined traces just on the boundary itself. Intuitively, this is a natural consequence of the localization of nonlocal interactions on the boundary. In contrast, a standard norm associated with fractional Sobolev space is defined by

$$|u|_{H^\alpha(\Omega)}^2 = \int_{\Omega} \int_{\Omega} \frac{(u(\mathbf{y}) - u(\mathbf{x}))^2}{|\mathbf{y} - \mathbf{x}|^{2\alpha+d}} d\mathbf{y} d\mathbf{x}.$$

The regularity of the functions is effected by the denominator which vanishes at  $\mathbf{x} = \mathbf{y}$ .

We now state the special trace theorem, see [X. Tian and Du \[ibid.\]](#) for a complete proof.

**Theorem 5.3** (Special trace theorem). *For  $\Omega = (0, r) \times \mathbb{R}^{d-1}$  and  $\Gamma = \{0\} \times \mathbb{R}^{d-1}$ , there exists a constant  $C$  depends only on  $d$  such that for any  $u \in C^1(\bar{\Omega}) \cap \mathcal{S}(\Omega)$ ,*

$$\|u\|_{L^2(\Gamma)} \leq C \left( r^{-1/2} \|u\|_{L^2(\Omega)} + r^{1/2} |u|_{\mathcal{S}(\Omega)} \right), \text{ for } d \geq 1,$$

$$|u|_{H^{1/2}(\Gamma)} \leq C \left( r^{-1} \|u\|_{L^2(\Omega)} + |u|_{\mathcal{S}(\Omega)} \right), \text{ for } d \geq 2.$$

where the nonlocal semi-norm of  $\mathcal{S}(\Omega)$  is as given in (20).

**Coupled local and nonlocal models.** We use  $\Omega_-$  and  $\Omega_+$  to denote two open domains in  $\mathbb{R}^d$  that satisfy  $\overline{\Omega_-} \cap \overline{\Omega_+} = \Gamma$ , a co-dimension-1 interface, and  $\Omega$  to denote their union. We

consider the coupling of a local model on  $\Omega_-$  with a nonlocal model on  $\Omega_+$ , see [Figure 4](#). Let  $\mathcal{S}(\Omega_+)$  be the nonlocal space with heterogeneous localization on the boundary. By the trace theorem, we define the energy (solution) space and the test function space as

$$\mathcal{W}(\Omega) = \{u \in H^1(\Omega_-) \cap \mathcal{S}(\Omega_+) \mid u_- = u_+ \text{ on } \Gamma\}$$

$$\mathcal{W}_0(\Omega) = \{u \in \mathcal{W}(\Omega) \mid u = 0 \text{ on } \partial\Omega\},$$

where  $\{u_{\pm}(x)\}$  denotes the traces of  $u$  defined from  $\Omega_{\pm}$  respectively. From [Proposition 5.1](#), we have the space  $H^1(\Omega)$  continuously imbedded in  $\mathcal{W}(\Omega)$  and  $H_0^1(\Omega)$  also continuously imbedded in  $\mathcal{W}_0(\Omega)$ . For  $u \in \mathcal{W}(\Omega)$ , its norm is defined as  $\|u\|_{\mathcal{W}(\Omega)} = \|u\|_{H^1(\Omega_-)} + \|u\|_{\mathcal{S}(\Omega_+)}$ . For  $g \in H^{1/2}(\partial\Omega)$  and  $f \in L^2(\Omega)$ , we have a coupled nonlocal-to-local model ([22](#)).

$$(22) \quad \begin{array}{ll} \min \left\{ \frac{1}{2} |u|_{H^1(\Omega_-)}^2 + \frac{1}{2} |u|_{\mathcal{S}(\Omega_+)}^2 - (f, u)_{\Omega} \right\}, & -\Delta u = f \\ \text{subject to } u \in \mathcal{W}(\Omega) \text{ and } u|_{\partial\Omega} = g. & u \in H^1(\Omega_-) \end{array} \quad \begin{array}{c} \text{Diagram: A vertical line } \Gamma \text{ separates } \Omega_- \text{ (left) and } \Omega_+ \text{ (right). In } \Omega_+, \text{ there are two overlapping dashed red circles. One circle contains a red dot, and the other contains a blue dot. To the right of the diagram, the text } -\mathcal{L}u = f \text{ and } u \in \mathcal{S}(\Omega_+) \text{ is shown.} \end{array}$$

Figure 4: Variational formulation of a coupled local-nonlocal model.

**Well-posedness of the coupled model.** For ([22](#)) to be well-posed, the coercivity of the energy functional is the key, which is consequence of a Poincaré inequality on  $\mathcal{W}_0(\Omega)$ . The latter can be established in a similar fashion as that on the nonlocal space with the constant horizon (and the local Sobolev space  $H_0^1(\Omega)$  as well). We thus have

**Proposition 5.4.** *The coupled variational problem ([22](#)) has a unique minimizer  $u \in \mathcal{W}_0(\Omega)$ .*

The seamless coupling of the nonlocal and local model means that one could use the same numerical discretization to solve the coupled problems if the heterogeneous localization of horizon can be handled effectively. Indeed, this is where we can circle back to utilize the concept of robust asymptotically compatible schemes [X. Tian \[2017\]](#), [Du, Tao, and X. Tian \[2018\]](#), and [X. Tian and Du \[2014\]](#).

## 6 Nonlocal in time dynamics

Spatial nonlocality is often accompanied by temporal correlations and memory effects. The latter involves nonlocality in time. Let us note first that a major difference in time and



space nonlocality is perhaps the generic time irreversibility. While a local time derivative may be defined by an infinitesimal change either backward to the history or forward to the future, it is more natural to view nonlocal time derivative as only dependent on past history. Thus, it is of much interests to reconsider the basic operators of the nonlocal vector calculus to accommodate the nonlocal interactions that are not symmetric. Of course, the issue of symmetry does not only pertain to changes in time. In earlier works, nonlocal gradients of the *upwind type*, variants of the operators given in [Section 3](#), have been utilized in the modeling of convective effects [H. Tian, Ju, and Du \[2017\]](#) and in the nonlocal formulation of conservation laws [Du and Huang \[2017\]](#) and [Du, Huang, and LeFloch \[2017\]](#). They have also been used to perform nonlocal gradient recovery [Du, Tao, X. Tian, and J. Yang \[2016\]](#). The first rigorous treatment of a nonlocal in time dynamics with a finite memory span, in the spirit of nonlocal vector calculus, was given in [Du, J. Yang, and Z. Zhou \[2017\]](#), which we follow here.

**Nonlocal time derivative and nonlocal-in-time dynamics.** We take the operator

$$(\mathcal{G}_\delta u)(t) = \lim_{\epsilon \rightarrow 0} \int_\epsilon^\delta \frac{u(t) - u(t-s)}{s} \rho_\delta(s) ds, \quad \text{for } t > 0,$$

as the nonlocal time derivative for a nonnegative density kernel  $\rho_\delta$  that is supported in the interval  $[0, \delta)$ . This leads to the study of an abstract nonlocal-in-time dynamics:

$$(23) \quad \mathcal{G}_\delta u + \mathcal{Q}u = f, \quad \forall t \in \Omega_T = (0, T) \subset \mathbb{R}_+, \quad u(t) = g(t), \quad \forall t \in (-\delta, 0) \subset \mathbb{R}_-.$$

for a linear operator  $\mathcal{Q}$  in an abstract space, together with some nonlocal initial (historical) data  $g = g(t)$ . We recall a well-posedness result for (23) corresponding to  $\mathcal{Q} = -\Delta$  on a bounded spatial domain  $\Omega$  with a homogeneous Dirichlet boundary condition [Du, J. Yang, and Z. Zhou \[ibid.\]](#).

**Theorem 6.1.** *For  $f \in L^2(0, T; H^{-1}(\Omega))$ , the problem (23) for  $\mathcal{Q} = -\Delta$  on  $\Omega$  with the homogeneous Dirichlet boundary condition and  $g(x, t) \equiv 0$  has a unique weak solution  $u \in L^2(0, T; H_0^1(\Omega))$ . Moreover, there is a constant  $c$ , independent of  $\delta$ ,  $f$  and  $u$ , such that.  $\|u\|_{L^2(0, T; H_0^1(\Omega))} + \|\mathcal{G}_\delta u\|_{L^2(0, T; H^{-1}(\Omega))} \leq c \|f\|_{L^2(0, T; H^{-1}(\Omega))}$ .*

The nonlocal-in-time diffusion equation may be related to fractional in time sub-diffusion equations like  $\partial_t^\alpha u - \Delta u = 0$  for  $\alpha \in (0, 1)$  [Du, J. Yang, and Z. Zhou \[2017\]](#), [Metzler and Klafter \[2004\]](#), and [Sokolov \[2012\]](#) by taking some special memory kernels [Allen, Caffarelli, and Vasseur \[2016\]](#). Such equations have often been used to describe the continuous time random walk (CTRW) of particles in heterogeneous media, where trapping events occur. In particular, particles get repeatedly immobilized in the environment for a trapping time drawn from the waiting time PDF that has a heavy tail [Metzler and Klafter](#)

[2004]. In general though, (23) provides a new class of models, due to the finite memory span, that serves to bridge anomalous and normal diffusion, with the latter being the limit as  $\delta \rightarrow 0$ . Indeed, the model (23) can also be related to a trapping model, see Du [n.d.], Du, J. Yang, and Z. Zhou [2017], and Du and Z. Zhou [2018] for more detailed studies.

**Crossover of diffusion regimes.** Diffusions in heterogeneous media have important implications in many applications. Using single particle tracking, recent studies have revealed many examples of anomalous diffusion, such as sub-diffusion with a slower spreading process in more constricted environment Berkowitz, Klafter, Metzler, and Scher [2002], He, Song, Su, Geng, Ackerson, Peng, and Tong [2016], and Jeon, Monne, Javanainen, and Metzler [2012]. Meanwhile, the origins and models of anomalous diffusion might differ significantly Korabel and Barkai [2010], McKinley, Yao, and Forest [2009], and Sokolov [2012]. On one hand, new experimental standards have been called for Saxton [2012]. On the other hand, there are needs for in-depth studies of mathematical models, many of which are non-conventional and non-local Du, Huang, and Lehoucq [2014], Du, Gunzburger, Lehoucq, and K. Zhou [2012], and Sokolov [2012].

Motivated by recent experimental reports on the crossover between initial transient sub-diffusion and long time normal diffusion in various settings He, Song, Su, Geng, Ackerson, Peng, and Tong [2016], the simple dynamic equation (23) with  $\mathcal{A} = -\Delta$  provides an effective description of the diffusion process encompassing these regimes Du and Z. Zhou [2018]. For model (23), the memory effect dominates initially, but as time goes on, the fixed memory span becomes less significant over the long life history. As a result, the transition from sub-diffusion to normal diffusion occurs naturally. This phenomenon can be illustrated by considering the mean square displacement (MSD)  $m(t)$  which can be explicitly computed Du and Z. Zhou [ibid.]. In Figure 5, we plot a solution of  $\mathcal{G}_\delta m(t) = 2$ , i.e., the mean square displacement of the nonlocal solution for  $f \equiv 0$  and  $\rho_\delta(s) = (1 - \alpha)\delta^{\alpha-1}s^{-\alpha}$  with  $\alpha = 0.2$  and  $\delta = 0.5$ . The result again illustrates the analytically suggested transition from the early fractional anomalous diffusion regime to the later standard diffusion regime. This "transition" or "crossover" behavior have been seen in many applications, e.g. diffusions in lipid bilayer systems of varying chemical compositions Jeon, Monne, Javanainen, and Metzler [2012, Fig.2], and lateral motion of the acetylcholine receptors on live muscle cell membranes He, Song, Su, Geng, Ackerson, Peng, and Tong [2016, Figs.3, 4].

## 7 Discussion and conclusion

Nonlocal models, arguably more general than their local or discrete analogs, are designed to account for nonlocal interactions explicitly and to remain valid for complex

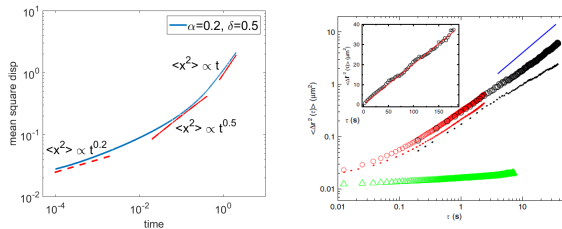


Figure 5: The MSD plot for (23) and the MSD curve He, Song, Su, Geng, Ackerson, Peng, and Tong [2016, Fig. 3] showing the crossover from sub-diffusion to normal diffusion for immobile AChRs.

systems involving possibly singular solutions. They have the potential to be alternatives and bridges to existing local continuum and discrete models. Their increasing popularity in applications makes the development of a systematic/axiomatic mathematical framework for nonlocal models necessary and timely. This work attempts to answer a few questions on nonlocal modeling, analysis and computation, particularly for models involving a finite-range nonlocal interactions and vector fields. To invite further studies on the subject, it might be more enticing to identify some issues worthy further investigation and to explore connections with other relevant topics. This is the purpose here, but before we proceed, we note that there are already many texts and online resources devoted to nonlocal models (scalar fractional equations in particular, see for example more recent books Bucur and Valdinoci [2016], Vázquez [2017], and West [2016] and [http://www.ma.utexas.edu/mediawiki/index.php/Starting\\_page](http://www.ma.utexas.edu/mediawiki/index.php/Starting_page)). We also refer to Du [n.d.] for more details and references on topics discussed below.

**Nonlocal exterior calculus and geometry.** While an analogy has been drawn between traditional local calculus and the nonlocal vector calculus involving nonlocal operators and fluxes, nonlocal integration by parts and nonlocal conservation laws, the nonlocal framework still needs to be updated or revamped. For example, a geometrically intrinsic framework for nonlocal exterior calculus and nonlocal forms on manifolds is not yet available. It would be of interests to develop nonlocal geometric structures that are more general than both discrete complexes and smooth Riemannian manifolds. In connection with such investigations, there are relevant studies on metric spaces Burago, Ivanov, and Kurylev [2014] and Fefferman, Ivanov, Kurylev, Lassas, and Narayanan [2015], Laplace-Beltrami Belkin and Niyogi [2008] and Lévy [2006], and combinatorial Hodge theory with scalar nonlocal forms Bartholdi, Schick, N. Smale, and S. Smale [2012]. We also made attempts like Le [2012] to introduce nonlocal vector forms, though more coherent constructions are desired. Given the close relations between local continuum models of

mechanics and differential geometry, one expects to find deep and intrinsic connections between nonlocal mechanics and geometry.

**Nonlocal models, kernel methods, graph and data.** Discrete, graph, network models and various kernel based methods in statistics often exhibit nonlocality. Exploring their continuum limits and localization can offer fundamental insights. In this direction, we mention some works related to graph Laplacians, diffusion maps, spectral clustering and so on [Coifman and Lafon \[2006\]](#), [Singer and H.-T. Wu \[2017\]](#), [Spielman \[2010\]](#), [Trillos and Slepčev \[2016\]](#), and [van Gennip and A. L. Bertozzi \[2012\]](#). These subjects are also connected with the geometric analysis already mentioned and applications such as image and data analysis and learning [Buades, Coll, and Morel \[2010\]](#), [Gilboa and Osher \[2008\]](#), and [Lou, X. Zhang, Osher, and A. Bertozzi \[2010\]](#). For instance, one can find, for applications to image analysis, the notion of nonlocal means [Buades, Coll, and Morel \[2010\]](#) and nonlocal (NL) gradient operator [Gilboa and Osher \[2008\]](#) together with a graph divergence all defined for scalar fields. Indeed, there have been much works on nonlocal calculus for scalar quantities, see [Du \[n.d.\]](#) for more detailed comparisons.

**Nonlocal function spaces, variational problems and dynamic systems.** While there have been a vast amount of studies on nonlocal functional spaces, related variational problems and dynamic systems, such as [Ambrosio, De Philippis, and Martinazzi \[2011\]](#), [Bourgain, Brezis, and Mironescu \[2001\]](#), [Bucur and Valdinoci \[2016\]](#), [Caffarelli and Silvestre \[2007\]](#), [Silvestre \[2014\]](#), and [West \[2016\]](#), the majority of them have focused on scalar quantities of interests and are often associated with fractional differential operators, fractional calculus, fractional Sobolev spaces and fractional PDEs having global interactions. On the other hand, motivated by applications in mechanics, our recent works can serve as a starting point of further investigations on nonlocal functional analysis of vector and tensor fields and systems of nonlocal models. For example, one may consider nonlocal extensions to the variational theory of nonlinear elasticity [Ball \[2010\]](#) and use them to develop better connections with atomistic modeling. One may further consider nonlocal spaces that can account for anisotropies and heterogeneities in both state and configuration variables. Extensions of the new trace theorems on heterogeneously localized nonlocal spaces to various vector field forms are also topics of more subsequent research. For instance, one may investigate possible nonlocal generalization of the trace theorems on the normal component of vector fields in the  $H(\text{div})$  space [Buffa and Ciarlet \[2001\]](#). Moreover, there are also interesting questions related to nonlocal models of fluid mechanics, including the nonlocal Navier-Stokes equations involving fractional order derivatives [Constantin and Vicol \[2012\]](#) and more recently analyzed nonlocal analogs of the linear incompressible Stokes equation as presented in the following forms, together with a comparison with

their classical form in the local limit:

$$(24) \quad \begin{cases} -\mathcal{L}_\delta \mathbf{u} + \mathcal{G}_\delta p = \mathbf{b}, \\ -\mathfrak{D}_\delta \mathbf{u} = 0, \end{cases} \quad \text{or} \quad \begin{cases} -\mathcal{L}_\delta \mathbf{u} + \mathcal{G}_\delta p = \mathbf{b}, \\ -\mathfrak{D}_\delta \mathbf{u} - \delta^2 \hat{\mathcal{L}}_\delta p = 0, \end{cases} \quad \text{and} \quad \begin{cases} -\Delta \mathbf{u} + \nabla p = \mathbf{b}, \\ \nabla \cdot \mathbf{u} = 0, \end{cases}$$

where  $\mathcal{L}_\delta$  and  $\hat{\mathcal{L}}_\delta$  are nonlocal diffusion operators,  $\mathcal{G}_\delta$  and  $\mathfrak{D}_\delta$  are one-point nonlocal gradient and divergence operators, similar to ones described in [Section 3](#). There are surely more questions about the extensions to time-dependent and nonlinear systems.

**Nonlocal, multiscale and stochastic modeling.** Nonlocality arises naturally from model reductions and has appeared (either knowingly or implicitly) in many early works (such as the Mori-Zwanzig formalism [Chorin, Hald, and Kupferman \[2002\]](#)). Nonlocal modeling could play more prominent roles in multiscale and stochastic modeling, ranging from bridging atomistic and continuum models, to data-driver model reductions of dynamic systems. There are also strong connections of nonlocal models with hydrodynamic descriptions of collective behavior and flocking hydrodynamics [Motsch and Tadmor \[2014\]](#) and [Shvydkoy and Tadmor \[2017\]](#). Exploring nonlocal models in diffusion and dispersal processes has also received much attention [Fuentes, Kuperman, and Kenkre \[2003\]](#), [Kao, Lou, and Shen \[2010\]](#), and [Massaccesi and Valdinoci \[2017\]](#), with the resulting nonlocal models having strong ties with stochastic processes, particularly, non-Gaussian and non-Markovian behaviors [Kumagai \[2014\]](#) and [Zaburdaev, Denisov, and Klafter \[2015\]](#). Stochastic nonlocal modeling is certainly an interesting subject on its own. In addition, inverse problems related to nonlocal models are also essential research subjects of both theoretical and practical interests and they can also be connected with various design and control problems.

**Nonlocal modeling, numerical analysis and simulation.** Numerical simulations of nonlocal models bring new computational challenges, from discretization to efficient solvers. To elevate the added cost associated with nonlocal interactions, it is of interests to explore a whole host of strategies, including local and nonlocal coupling [Li and Lu \[2017\]](#) and [Du, Tao, and X. Tian \[2018\]](#), adaptive grids [Du, L. Tian, and Zhao \[2013\]](#), multigrid and fast solvers [Du and Z. Zhou \[2017\]](#) and [H. Wang and H. Tian \[2012\]](#), some of them are less examined than others and most of topics remain to be further studied. The subject is naturally linked to sparse and low rank approximations that would allow one to explore the nonlocal structure to achieve efficient evaluation of nonlocal interactions as well as the solution of associated algebraic systems. Scalable algorithms via domain decomposition or other strategies that can particularly handle the information exchange (communications between processors) involving nonlocal interactions are interesting and important research questions. Let us also mention that nonlocal models can also become effective

tools to analyze numerical schemes that were initially developed to solve local PDEs. For example, to understand the interplay between the smoothing length and the particle spacing in the context of smoothed particle hydrodynamics [Gingold and Monaghan \[1977\]](#) and [Monaghan \[2005\]](#), nonlocal continuum systems (24) can help providing a rigorous mathematical foundation for improving the stability and robustness of the discretization [Du and X. Tian \[2017\]](#). Another example is concerned with discretization schemes for multidimensional local diffusion equations through the nonlocal integral formulation [Du, Tao, X. Tian, and J. Yang \[n.d.\]](#) and [Nochetto and W. Zhang \[2017\]](#), a topic linked with approximations of fully nonlinear elliptic equations such as the Monge-Ampère. An open question there is whether or not there are discretization schemes on unstructured meshes which can preserve the discrete maximum principles and are asymptotically compatible for general anisotropic and heterogeneous diffusion equations.

**Thinking nonlocally, acting locally.** The pushes for nonlocal modeling come from several fronts. Foremost, the development of nonlocal models is driven by the interests in studying singular/anomalous/stochastic/multiscale behavior of complex systems where nonlocal models can potentially unify and bridge different models. Nowadays, the imminent growth of nonlocal modeling may also be attributed to the inescapable presence of nonlocality in the daily human experience. The emergence of augmented reality, information technology and data science as well as intelligent computing has been fueling the popularity of nonlocal modeling as the world is getting more than ever remotely and non-locally networked together. With extreme computing capabilities beyond doing simple analytical approximations, we could be ready to tackle nonlocal interactions directly. Yet, despite the huge lift in computing power, exploring simple representations and closure relations via local, sparse, low rank or low dimensional approximations is still of great theoretical interest and practical significance. We thus conclude by saying that promoting the role of nonlocal modeling is to not only argue for the need to think nonlocally and to retain nonlocal features wherever necessary, but also point out the importance in utilize local models wherever feasible, hence to act locally, as our goal is to have the efficiency and robustness of mathematical modeling and numerical simulations while maintaining their generality and predicability.

**Acknowledgments.** The author would like to thank the organizing committee of ICM 2018 for the invitation. Some of the more detailed discussions presented here was taken from the joint publications with Xiaochuan Tian, Tadele Mengesha, Max Gunzburger, Richard Lehoucq, Kun Zhou, Zhi Zhou and Jiang Yang. The author is grateful to them and many other students and collaborators on the subject (an extended list of them is given in [Du \[n.d.\]](#)).

## References

- Mark Allen, Luis Caffarelli, and Alexis Vasseur (2016). “[A parabolic problem with a fractional time derivative](#)”. *Arch. Ration. Mech. Anal.* 221.2, pp. 603–630. MR: [3488533](#) (cit. on p. [3576](#)).
- Luigi Ambrosio, Guido De Philippis, and Luca Martinazzi (2011). “[Gamma-convergence of nonlocal perimeter functionals](#)”. *Manuscripta Math.* 134.3–4, pp. 377–403. MR: [2765717](#) (cit. on p. [3579](#)).
- Douglas N. Arnold and Franco Brezzi (1997). “[Locking-free finite element methods for shells](#)”. *Math. Comp.* 66.217, pp. 1–14. MR: [1370847](#) (cit. on p. [3569](#)).
- John M Ball (2010). “Progress and puzzles in nonlinear elasticity”. In: *Poly-, quasi- and rank-one convexity in applied mechanics*. Springer, pp. 1–15 (cit. on p. [3579](#)).
- Laurent Bartholdi, Thomas Schick, Nat Smale, and Steve Smale (2012). “[Hodge theory on metric spaces](#)”. *Found. Comput. Math.* 12.1. Appendix B by Anthony W. Baker, pp. 1–48. MR: [2886155](#) (cit. on p. [3578](#)).
- Mikhail Belkin and Partha Niyogi (2008). “[Towards a theoretical foundation for Laplacian-based manifold methods](#)”. *J. Comput. System Sci.* 74.8, pp. 1289–1308. MR: [2460286](#) (cit. on p. [3578](#)).
- Brian Berkowitz, Joseph Klafter, Ralf Metzler, and Harvey Scher (2002). “Physical pictures of transport in heterogeneous media: Advection-dispersion, random-walk, and fractional derivative formulations”. *Water Resources Research* 38.10, pp. 9–1 (cit. on p. [3577](#)).
- M. A. Bessa, J. T. Foster, T. Belytschko, and Wing Kam Liu (2014). “[A meshfree unification: reproducing kernel peridynamics](#)”. *Comput. Mech.* 53.6, pp. 1251–1264. MR: [3201938](#) (cit. on p. [3568](#)).
- Florin Bobaru, John T. Foster, Philippe H. Geubelle, and Stewart A. Silling, eds. (2017). *Handbook of peridynamic modeling*. Advances in Applied Mathematics. CRC Press, Boca Raton, FL, pp. xxxviii+548. MR: [3560288](#) (cit. on p. [3562](#)).
- Florin Bobaru, Mijia Yang, Leonardo Frota Alves, Stewart A. Silling, Ebrahim Askari, and Jifeng Xu (2009). “Convergence, adaptive refinement, and scaling in 1D peridynamics”. *International Journal for Numerical Methods in Engineering* 77.6, pp. 852–877 (cit. on p. [3571](#)).
- Jean Bourgain, Haim Brezis, and Petru Mironescu (2001). “Another look at Sobolev spaces”. In: *Optimal control and partial differential equations*. IOS, Amsterdam, pp. 439–455. MR: [3586796](#) (cit. on p. [3566](#), [3572](#), [3573](#), [3579](#)).
- A. Buades, B. Coll, and J. M. Morel (2010). “[Image denoising methods. A new nonlocal principle](#)”. *SIAM Rev.* 52.1. Reprint of “A review of image denoising algorithms, with a new one” [MR2162865], pp. 113–147. MR: [2608636](#) (cit. on p. [3579](#)).

- Claudia Bucur and Enrico Valdinoci (2016). *Nonlocal diffusion and applications*. Vol. 20. Lecture Notes of the Unione Matematica Italiana. Springer, [Cham]; Unione Matematica Italiana, Bologna, pp. xii+155. MR: [3469920](#) (cit. on pp. [3561](#), [3578](#), [3579](#)).
- A. Buffa and P. Ciarlet Jr. (2001). “On traces for functional spaces related to Maxwell’s equations. I. An integration by parts formula in Lipschitz polyhedra”. *Math. Methods Appl. Sci.* 24.1, pp. 9–30. MR: [1809491](#) (cit. on p. [3579](#)).
- Dmitri Burago, Sergei Ivanov, and Yaroslav Kurylev (2014). “A graph discretization of the Laplace-Beltrami operator”. *J. Spectr. Theory* 4.4, pp. 675–714. MR: [3299811](#) (cit. on p. [3578](#)).
- Luis Caffarelli and Luis Silvestre (2007). “An extension problem related to the fractional Laplacian”. *Comm. Partial Differential Equations* 32.7-9, pp. 1245–1260. MR: [2354493](#) (cit. on pp. [3561](#), [3579](#)).
- X. Chen and Max Gunzburger (2011). “Continuous and discontinuous finite element methods for a peridynamics model of mechanics”. *Comput. Methods Appl. Mech. Engrg.* 200.9-12, pp. 1237–1250. MR: [2796157](#) (cit. on p. [3568](#)).
- Alexandre J. Chorin, Ole H. Hald, and Raz Kupferman (2002). “Optimal prediction with memory”. *Phys. D* 166.3-4, pp. 239–257. MR: [1915310](#) (cit. on p. [3580](#)).
- Ronald R. Coifman and Stéphane Lafon (2006). “Diffusion maps”. *Appl. Comput. Harmon. Anal.* 21.1, pp. 5–30. MR: [2238665](#) (cit. on p. [3579](#)).
- Peter Constantin and Vlad Vicol (2012). “Nonlinear maximum principles for dissipative linear nonlocal operators and applications”. *Geom. Funct. Anal.* 22.5, pp. 1289–1321. MR: [2989434](#) (cit. on p. [3579](#)).
- Qiang Du (n.d.). *Nonlocal modeling, analysis and computation, NSF-CBMS Monograph*. To appear in *SIAM* (cit. on pp. [3560](#), [3561](#), [3564](#), [3566](#), [3577](#)–[3579](#), [3581](#)).
- (2015). “From peridynamics to stochastic jump process: illustrations of nonlocal balance laws and nonlocal calculus framework”. *Scientia Sinica Mathematica* 45.7, pp. 939–952 (cit. on p. [3561](#)).
  - (2017a). “Local limits and asymptotically compatible discretizations”. In: *Handbook of peridynamic modeling*. Adv. Appl. Math. CRC Press, Boca Raton, FL, pp. 87–108. MR: [3618714](#) (cit. on p. [3568](#)).
  - (2017b). “Nonlocal calculus of variations and well-posedness of peridynamics”. In: *Handbook of peridynamic modeling*. Adv. Appl. Math. CRC Press, Boca Raton, FL, pp. 63–85. MR: [3618713](#) (cit. on p. [3561](#)).
- Qiang Du, Max Gunzburger, Richard B. Lehoucq, and Kun Zhou (2012). “Analysis and approximation of nonlocal diffusion problems with volume constraints”. *SIAM Rev.* 54.4, pp. 667–696. MR: [3023366](#) (cit. on pp. [3565](#), [3577](#)).
- (2013). “A nonlocal vector calculus, nonlocal volume-constrained problems, and nonlocal balance laws”. *Math. Models Methods Appl. Sci.* 23.3, pp. 493–540. MR: [3010838](#) (cit. on pp. [3563](#)–[3565](#), [3567](#)).



- Qiang Du and Zhan Huang (2017). “Numerical solution of a scalar one-dimensional monotonicity-preserving nonlocal nonlinear conservation law”. *J. Math. Res. Appl.* 37.1, pp. 1–18. MR: [3642662](#) (cit. on pp. [3572](#), [3576](#)).
- Qiang Du, Zhan Huang, and Philippe G. LeFloch (2017). “Nonlocal conservation laws. A new class of monotonicity-preserving models”. *SIAM J. Numer. Anal.* 55.5, pp. 2465–2489. MR: [3715381](#) (cit. on p. [3576](#)).
- Qiang Du, Zhan Huang, and Richard B. Lehoucq (2014). “Nonlocal convection-diffusion volume-constrained problems and jump processes”. *Discrete Contin. Dyn. Syst. Ser. B* 19.2, pp. 373–389. MR: [3170190](#) (cit. on pp. [3562](#), [3577](#)).
- Qiang Du, L. Ju, and J. Lu (2018). “A discontinuous Galerkin method for one-dimensional time-dependent nonlocal diffusion problems”. *Mathematics of Computation* (cit. on pp. [3568](#), [3572](#)).
- Qiang Du, Yunzhe Tao, and Xiaochuan Tian (2017). “A Peridynamic Model of Fracture Mechanics with Bond-Breaking”. *Journal of Elasticity*, pp. 1–22 (cit. on p. [3563](#)).
- (2018). “Nonlocal models with heterogeneous localization and their application to seamless local-nonlocal coupling”. In preparation (cit. on pp. [3572](#), [3575](#), [3580](#)).
- Qiang Du, Yunzhe Tao, Xiaochuan Tian, and Jiang Yang (n.d.). “Asymptotically compatible numerical approximations of multidimensional nonlocal diffusion models and nonlocal Green’s functions”. To appear in *IMA J. Numerical Analysis* (cit. on pp. [3570](#), [3571](#), [3581](#)).
- (2016). “Robust a posteriori stress analysis for quadrature collocation approximations of nonlocal models via nonlocal gradients”. *Comput. Methods Appl. Mech. Engrg.* 310, pp. 605–627. MR: [3548574](#) (cit. on pp. [3567](#), [3572](#), [3576](#)).
- Qiang Du, Li Tian, and Xuying Zhao (2013). “A convergent adaptive finite element algorithm for nonlocal diffusion and peridynamic models”. *SIAM J. Numer. Anal.* 51.2, pp. 1211–1234. MR: [3045653](#) (cit. on p. [3580](#)).
- Qiang Du and Xiaochuan Tian (n.d.). “Stability of nonlocal Dirichlet integrals and implications for peridynamic correspondence material modeling”. To appear in *SIAM J. Applied Mathematics* (cit. on p. [3567](#)).
- (2015). “Robust discretization of nonlocal models related to peridynamics”. In: *Mesh-free methods for partial differential equations VII*. Vol. 100. Lect. Notes Comput. Sci. Eng. Springer, Cham, pp. 97–113. MR: [3587379](#) (cit. on pp. [3561](#), [3569](#), [3570](#)).
- (2017). “Mathematics of smoothed particle hydrodynamics, part I: nonlocal Stokes equation” (cit. on p. [3581](#)).
- Qiang Du and Jiang Yang (2016). “Asymptotically compatible Fourier spectral approximations of nonlocal Allen-Cahn equations”. *SIAM J. Numer. Anal.* 54.3, pp. 1899–1919. MR: [3514714](#) (cit. on p. [3572](#)).

- (2017). “Fast and accurate implementation of Fourier spectral approximations of non-local diffusion operators and its applications”. *J. Comput. Phys.* 332, pp. 118–134. MR: [3591174](#) (cit. on p. [3568](#)).
- Qiang Du, Jiang Yang, and Zhi Zhou (2017). “Analysis of a nonlocal-in-time parabolic equation”. *Discrete Contin. Dyn. Syst. Ser. B* 22.2, pp. 339–368. MR: [3639119](#) (cit. on pp. [3576](#), [3577](#)).
- Qiang Du and Kun Zhou (2011). “Mathematical analysis for the peridynamic nonlocal continuum theory”. *ESAIM Math. Model. Numer. Anal.* 45.2, pp. 217–234. MR: [2804637](#) (cit. on p. [3561](#)).
- Qiang Du and Z. Zhou (2018). “A nonlocal-in-time dynamic system for anomalous diffusion”. Preprint (cit. on p. [3577](#)).
- Qiang Du and Zhi Zhou (2017). “Multigrid finite element method for nonlocal diffusion equations with a fractional kernel” (cit. on p. [3580](#)).
- Charles Fefferman, Sergei Ivanov, Yaroslav Kurylev, Matti Lassas, and Hariharan Narayanan (Aug. 2015). “Reconstruction and interpolation of manifolds I: The geometric Whitney problem”. arXiv: [1508.00674](#) (cit. on p. [3578](#)).
- MA Fuentes, MN Kuperman, and VM Kenkre (2003). “Nonlocal interaction effects on pattern formation in population dynamics”. *Physical review letters* 91.15, p. 158104 (cit. on p. [3580](#)).
- Yves van Gennip and Andrea L. Bertozzi (2012). “T-convergence of graph Ginzburg-Landau functionals”. *Adv. Differential Equations* 17.11-12, pp. 1115–1180. MR: [3013414](#) (cit. on p. [3579](#)).
- Guy Gilboa and Stanley Osher (2008). “Nonlocal operators with applications to image processing”. *Multiscale Model. Simul.* 7.3, pp. 1005–1028. MR: [2480109](#) (cit. on p. [3579](#)).
- R. Gingold and J. J. Monaghan (1977). “Smoothed particle hydrodynamics: theory and application to non-spherical stars”. *Monthly Notices of the Royal Astronomical Society* 181, pp. 375–389 (cit. on p. [3581](#)).
- Jean-Luc Guermond and Guido Kanschat (2010). “Asymptotic analysis of upwind discontinuous Galerkin approximation of the radiative transport equation in the diffusive limit”. *SIAM J. Numer. Anal.* 48.1, pp. 53–78. MR: [2608358](#) (cit. on p. [3569](#)).
- Wei He, Hao Song, Yun Su, Ling Geng, Bruce J Ackerson, HB Peng, and Penger Tong (2016). “Dynamic heterogeneity and non-Gaussian statistics for acetylcholine receptors on live cell membrane”. *Nature communications* 7, p. 11701 (cit. on pp. [3577](#), [3578](#)).
- Jae-Hyung Jeon, Hector Martinez-Seara Monne, Matti Javanainen, and Ralf Metzler (2012). “Anomalous diffusion of phospholipids and cholesterol in a lipid bilayer and its origins”. *Physical review letters* 109.18, p. 188103 (cit. on p. [3577](#)).
- Shi Jin (1999). “Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations”. *SIAM J. Sci. Comput.* 21.2, pp. 441–454. MR: [1718639](#) (cit. on p. [3569](#)).

- Chiu-Yen Kao, Yuan Lou, and Wenxian Shen (2010). “[Random dispersal vs. non-local dispersal](#)”. *Discrete Contin. Dyn. Syst.* 26.2, pp. 551–596. MR: [2556498](#) (cit. on p. [3580](#)).
- Nickolay Korabel and Eli Barkai (2010). “Paradoxes of subdiffusive infiltration in disordered systems”. *Physical review letters* 104.17, p. 170603 (cit. on p. [3577](#)).
- Takashi Kumagai (2014). “Anomalous random walks and diffusions: From fractals to random media”. In: *Proceedings of ICM*. Vol. IV. Citeseer, pp. 75–94 (cit. on p. [3580](#)).
- Thinh Duc Le (2012). “Nonlocal Exterior Calculus on Riemannian Manifolds”. PhD thesis. Penn State University (cit. on p. [3578](#)).
- Bruno Lévy (2006). “Laplace-beltrami eigenfunctions towards an algorithm that” understands” geometry”. In: *Shape Modeling and Applications, 2006. SMI 2006. IEEE International Conference on*. IEEE, pp. 13–21 (cit. on p. [3578](#)).
- Xingjie Helen Li and Jianfeng Lu (2017). “[Quasi-nonlocal coupling of nonlocal diffusions](#)”. *SIAM J. Numer. Anal.* 55.5, pp. 2394–2415. MR: [3709890](#) (cit. on pp. [3572](#), [3580](#)).
- Yifei Lou, Xiaoqun Zhang, Stanley Osher, and Andrea Bertozzi (2010). “[Image recovery via nonlocal operators](#)”. *J. Sci. Comput.* 42.2, pp. 185–197. MR: [2578033](#) (cit. on p. [3579](#)).
- Annalisa Massaccesi and Enrico Valdinoci (2017). “[Is a nonlocal diffusion strategy convenient for biological populations in competition?](#)” *J. Math. Biol.* 74.1-2, pp. 113–147. MR: [3590678](#) (cit. on p. [3580](#)).
- Scott A McKinley, Lingxing Yao, and M Gregory Forest (2009). “Transient anomalous diffusion of tracer particles in soft matter”. *Journal of Rheology* 53.6, pp. 1487–1506 (cit. on p. [3577](#)).
- Tadele Mengesha (2012). “[Nonlocal Korn-type characterization of Sobolev vector fields](#)”. *Commun. Contemp. Math.* 14.4, pp. 1250028, 28. MR: [2965673](#) (cit. on pp. [3565](#), [3566](#)).
- Tadele Mengesha and Qiang Du (2014). “[The bond-based peridynamic system with Dirichlet-type volume constraint](#)”. *Proc. Roy. Soc. Edinburgh Sect. A* 144.1, pp. 161–186. MR: [3164542](#) (cit. on pp. [3562](#), [3565](#), [3566](#)).
- (2015). “[On the variational limit of a class of nonlocal functionals related to peridynamics](#)”. *Nonlinearity* 28.11, pp. 3999–4035. MR: [3424902](#) (cit. on pp. [3565](#)–[3567](#)).
  - (2016). “[Characterization of function spaces of vector fields and an application in nonlinear peridynamics](#)”. *Nonlinear Anal.* 140, pp. 82–111. MR: [3492730](#) (cit. on pp. [3566](#), [3567](#)).
- Ralf Metzler and Joseph Klafter (2004). “[The restaurant at the end of the random walk: recent developments in the description of anomalous transport by fractional dynamics](#)”. *J. Phys. A* 37.31, R161–R208. MR: [2090004](#) (cit. on p. [3576](#)).
- J. J. Monaghan (2005). “[Smoothed particle hydrodynamics](#)”. *Rep. Progr. Phys.* 68.8, pp. 1703–1759. MR: [2158506](#) (cit. on p. [3581](#)).

- Sebastien Motsch and Eitan Tadmor (2014). “[Heterophilious dynamics enhances consensus](#)”. *SIAM Rev.* 56.4, pp. 577–621. MR: [3274797](#) (cit. on p. [3580](#)).
- Ricardo H. Nochetto, Enrique Otárola, and Abner J. Salgado (2016). “[A PDE approach to space-time fractional parabolic problems](#)”. *SIAM J. Numer. Anal.* 54.2, pp. 848–873. MR: [3478958](#) (cit. on p. [3561](#)).
- Ricardo H Nochetto and Wujun Zhang (2017). “Discrete ABP estimate and convergence rates for linear elliptic equations in non-divergence form”. *Foundations of Computational Mathematics*, pp. 1–57 (cit. on p. [3581](#)).
- Michael L. Parks, D. J. Littlewood, J. A. Mitchell, and Stewart A. Silling (2012). *Peridigm users’ guide*. Tech. rep. (cit. on p. [3568](#)).
- Michael L. Parks, Pablo Seleson, Steven J. Plimpton, Stewart A. Silling, and Richard B. Lehoucq (2011). “Peridynamics with LAMMPS: A User Guide, v0. 3 Beta”. *Sandia Report (2011–8253)* (cit. on p. [3568](#)).
- Augusto C. Ponce (2004). “[An estimate in the spirit of Poincaré’s inequality](#)”. *J. Eur. Math. Soc. (JEMS)* 6.1, pp. 1–15. MR: [2041005](#) (cit. on p. [3566](#)).
- Bo Ren, CT Wu, and E Askari (2017). “A 3D discontinuous Galerkin finite element method with the bond-based peridynamics model for dynamic brittle failure analysis”. *International Journal of Impact Engineering* 99, pp. 14–25 (cit. on p. [3568](#)).
- Michael J Saxton (2012). “Wanted: a positive control for anomalous subdiffusion”. *Bio-physical journal* 103.12, pp. 2411–2422 (cit. on p. [3577](#)).
- Pablo Seleson, Qiang Du, and Michael L. Parks (2016). “[On the consistency between nearest-neighbor peridynamic discretizations and discretized classical elasticity models](#)”. *Comput. Methods Appl. Mech. Engrg.* 311, pp. 698–722. MR: [3564707](#) (cit. on p. [3568](#)).
- Roman Shvydkoy and Eitan Tadmor (2017). “Eulerian dynamics with a commutator forcing”. *Transactions of Mathematics and its Applications* 1.1 (cit. on p. [3580](#)).
- Stewart A. Silling (2000). “[Reformulation of elasticity theory for discontinuities and long-range forces](#)”. *J. Mech. Phys. Solids* 48.1, pp. 175–209. MR: [1727557](#) (cit. on p. [3562](#)).
- (2010). “[Linearized theory of peridynamic states](#)”. *J. Elasticity* 99.1, pp. 85–111. MR: [2592410](#) (cit. on p. [3567](#)).
- Stewart A. Silling, M. Epton, O. Weckner, J. Xu, and E. Askari (2007). “[Peridynamic states and constitutive modeling](#)”. *J. Elasticity* 88.2, pp. 151–184. MR: [2348150](#) (cit. on p. [3567](#)).
- Stewart A. Silling and Richard B. Lehoucq (2010). “Peridynamic theory of solid mechanics”. In: *Advances in applied mechanics*. Vol. 44. Elsevier, pp. 73–168 (cit. on p. [3567](#)).
- Luis Silvestre (2014). “Regularity estimates for parabolic integro-differential equations and applications”. In: *Proceedings of the International Congress of Mathematicians—Seoul 2014. Vol. III*. Kyung Moon Sa, Seoul, pp. 873–894. MR: [3729056](#) (cit. on p. [3579](#)).

- Amit Singer and Hau-Tieng Wu (2017). “Spectral convergence of the connection Laplacian from random samples”. *Inf. Inference* 6.1, pp. 58–123. MR: [3636868](#) (cit. on p. [3579](#)).
- Igor M Sokolov (2012). “Models of anomalous diffusion in crowded environments”. *Soft Matter* 8.35, pp. 9043–9052 (cit. on pp. [3576](#), [3577](#)).
- Daniel A. Spielman (2010). “Algorithms, graph theory, and linear equations in Laplacian matrices”. In: *Proceedings of the International Congress of Mathematicians. Volume IV*. Hindustan Book Agency, New Delhi, pp. 2698–2722. MR: [2827990](#) (cit. on p. [3579](#)).
- Yunzhe Tao, Xiaochuan Tian, and Qiang Du (2017). “Nonlocal diffusion and peridynamic models with Neumann type constraints and their numerical approximations”. *Appl. Math. Comput.* 305, pp. 282–298. MR: [3621707](#) (cit. on p. [3568](#)).
- Hao Tian, Lili Ju, and Qiang Du (2017). “A conservative nonlocal convection-diffusion model and asymptotically compatible finite difference discretization”. *Comput. Methods Appl. Mech. Engrg.* 320, pp. 46–67. MR: [3646345](#) (cit. on pp. [3568](#), [3576](#)).
- Hao Tian, Hong Wang, and Wenqia Wang (2013). “An efficient collocation method for a non-local diffusion model”. *Int. J. Numer. Anal. Model.* 10.4, pp. 815–825. MR: [3125859](#) (cit. on p. [3568](#)).
- Xiaochuan Tian (2017). *Nonlocal models with a finite range of nonlocal interactions*. Thesis (Ph.D.)—Columbia University. ProQuest LLC, Ann Arbor, MI, p. 221. MR: [3641030](#) (cit. on p. [3575](#)).
- Xiaochuan Tian and Qiang Du (2013). “Analysis and comparison of different approximations to nonlocal diffusion and linear peridynamic equations”. *SIAM J. Numer. Anal.* 51.6, pp. 3458–3482. MR: [3143839](#) (cit. on pp. [3568](#)–[3571](#)).
- (2014). “Asymptotically compatible schemes and applications to robust discretization of nonlocal models”. *SIAM J. Numer. Anal.* 52.4, pp. 1641–1665. MR: [3231986](#) (cit. on pp. [3568](#), [3570](#), [3571](#), [3575](#)).
  - (2015). “Nonconforming discontinuous Galerkin methods for nonlocal variational problems”. *SIAM J. Numer. Anal.* 53.2, pp. 762–781. MR: [3323545](#) (cit. on pp. [3568](#), [3572](#)).
  - (2017). “Trace theorems for some nonlocal function spaces with heterogeneous localization”. *SIAM J. Math. Anal.* 49.2, pp. 1621–1644. MR: [3640623](#) (cit. on pp. [3572](#)–[3574](#)).
- Xiaochuan Tian, Qiang Du, and Max Gunzburger (2016). “Asymptotically compatible schemes for the approximation of fractional Laplacian and related nonlocal diffusion problems on bounded domains”. *Adv. Comput. Math.* 42.6, pp. 1363–1380. MR: [3571209](#) (cit. on p. [3571](#)).
- Nicolás García Trillos and Dejan Slepčev (2016). “A variational approach to the consistency of spectral clustering”. *Applied and Computational Harmonic Analysis* (cit. on p. [3579](#)).

- Juan Luis Vázquez (2017). *The mathematical theories of diffusion: Nonlinear and fractional diffusion*. Springer Lecture Notes in Mathematics, CIME Subseries (cit. on pp. [3561](#), [3578](#)).
- Hong Wang and Hao Tian (2012). “A fast Galerkin method with efficient matrix assembly and storage for a peridynamic model”. *J. Comput. Phys.* 231.23, pp. 7730–7738. MR: [2972865](#) (cit. on p. [3580](#)).
- Bruce J. West (2016). *Fractional calculus view of complexity*. Tomorrow’s science, With a foreword by Chris Arney. CRC Press, Boca Raton, FL, pp. xviii+285. MR: [3444156](#) (cit. on pp. [3561](#), [3578](#), [3579](#)).
- V. Zaburdaev, S. Denisov, and J. Klafter (2015). “Lévy walks”. *Rev. Modern Phys.* 87.2, pp. 483–530. MR: [3403266](#) (cit. on p. [3580](#)).
- Xiaoping Zhang, Max Gunzburger, and Lili Ju (2016a). “Nodal-type collocation methods for hypersingular integral equations and nonlocal diffusion problems”. *Comput. Methods Appl. Mech. Engrg.* 299, pp. 401–420. MR: [3434921](#) (cit. on p. [3568](#)).
- (2016b). “Quadrature rules for finite element approximations of 1D nonlocal problems”. *J. Comput. Phys.* 310, pp. 213–236. MR: [3457967](#) (cit. on p. [3568](#)).
- Kun Zhou and Qiang Du (2010). “Mathematical and numerical analysis of linear peridynamic models with nonlocal boundary conditions”. *SIAM J. Numer. Anal.* 48.5, pp. 1759–1780. MR: [2733097](#) (cit. on p. [3568](#)).

Received 2017-12-03.

QIANG DU (杜强)

[qd2125@columbia.edu](mailto:qd2125@columbia.edu)

DEPARTMENT OF APPLIED PHYSICS AND APPLIED MATHEMATICS

COLUMBIA UNIVERSITY

NEW YORK, NY 10027

USA



# AN INTRODUCTION TO MULTILEVEL MONTE CARLO METHODS

MICHAEL B. GILES

## Abstract

In recent years there has been very substantial growth in stochastic modelling in many application areas, and this has led to much greater use of Monte Carlo methods to estimate expected values of output quantities from stochastic simulation. However, such calculations can be expensive when the cost of individual stochastic simulations is very high. Multilevel Monte Carlo greatly reduces the computational cost by performing most simulations with low accuracy at a correspondingly low cost, with relatively few being performed at high accuracy and a high cost.

This article reviews the key ideas behind the multilevel Monte Carlo method. Some applications are discussed to illustrate the flexibility and generality of the approach, and the challenges in its numerical analysis.

## 1 Introduction

Stochastic modelling and simulation is an important and growing area in applied mathematics and scientific computing. One large application area is in computational finance, in quantitative risk management and the pricing of financial derivatives. Another is Uncertainty Quantification in engineering and science, which has led to new journals and annual conferences.

When the dimensionality of the uncertainty (i.e. the number of uncertain input variables) is low, it can be appropriate to model the uncertainty using the Fokker-Planck PDE and use stochastic Galerkin, stochastic collocation or polynomial chaos methods [Xiu and Karniadakis \[2002\]](#), [Babuška, Tempone, and Zouraris \[2004\]](#), [Babuška, Nobile, and Tempone \[2010\]](#), and [Gunzburger, Webster, and Zhang \[2014\]](#). When the level of uncertainty is low, and its effect is largely linear, then moment methods can be an efficient and accurate way in which to quantify the effects on uncertainty [Putko, Taylor, Newman, and](#)

---

The author acknowledges the financial support of the U.K. Engineering and Physical Sciences Research Council.

*MSC2010:* primary 65C05; secondary 65C30, 65C50, 60H35, 60H10, 60H15.



[Green \[2002\]](#). However, when the uncertainty is high-dimensional and strongly nonlinear, Monte Carlo simulation often remains the preferred approach.

At its simplest, Monte Carlo simulation is extremely simple. To estimate  $\mathbb{E}[P]$ , the expected value of a scalar output quantity of interest, a simple Monte Carlo estimate is just an equally-weighted average of the values  $P(\omega)$  for  $N$  independent samples  $\omega$  coming from the given probability space  $(\Omega, \mathcal{F}, \mathbb{P})$ ,

$$N^{-1} \sum_{n=1}^N P(\omega^{(n)}).$$

The variance of this estimate is  $N^{-1} \mathbb{V}[P]$ , so the RMS (root-mean-square) error is  $O(N^{-1/2})$  and an accuracy of  $\varepsilon$  requires  $N = O(\varepsilon^{-2})$  samples. This is the weakness of Monte Carlo simulation; its computational cost can be very high, particularly when each sample  $P(\omega)$  might require the approximate solution of a PDE, or a computation with many timesteps.

One approach to addressing this high cost is the use of Quasi-Monte Carlo (QMC) methods, in which the samples are not chosen randomly and independently, but are instead selected very carefully to reduce the error. In the best cases, the error may be  $O(N^{-1})$ , up to logarithmic terms, giving a very substantial reduction in the number of samples required for a given accuracy [Dick, Kuo, and Sloan \[2013\]](#).

In this article, we cover a different approach to improving the computational efficiency, the multilevel Monte Carlo (MLMC) method. This is important when the cost of computing the individual samples is very high, but it is possible to compute approximate values at a much lower cost. We also briefly discuss the combination of MLMC with QMC.

This article provides only a short introduction to the subject and some of the corresponding literature. For a more comprehensive overview of multilevel Monte Carlo methods, the author has recently written a 70-page review with a much more extensive list of references [Giles \[2015\]](#). There is also a webpage [web-page](#) with a list of active research groups and their publications.

## 2 Multilevel Monte Carlo

**2.1 MLMC with exact simulation.** The key idea in Multilevel Monte Carlo is also very simple. Suppose we are interested in estimating  $\mathbb{E}[P_L(\omega)]$ , and it is possible to exactly simulate  $P_L(\omega)$  but it is very costly. Suppose also that there is a sequence  $P_0(\omega), \dots, P_{L-1}(\omega)$  which approximates  $P_L(\omega)$  with increasing accuracy, but also increasing cost. In this case, instead of directly estimating  $\mathbb{E}[P_L]$  we can use the trivial identity

$$\mathbb{E}[P_L] = \mathbb{E}[P_0] + \sum_{\ell=1}^L \mathbb{E}[P_\ell - P_{\ell-1}],$$

to construct the following unbiased estimator for  $\mathbb{E}[P_L]$ ,

$$N_0^{-1} \sum_{n=1}^{N_0} P_0^{(0,n)} + \sum_{\ell=1}^L \left\{ N_\ell^{-1} \sum_{n=1}^{N_\ell} \left( P_\ell^{(\ell,n)} - P_{\ell-1}^{(\ell,n)} \right) \right\}$$

where  $P_\ell^{(\ell,n)}$  is shorthand for  $P_\ell(\omega^{(\ell,n)})$ , with the inclusion of the level  $\ell$  in the superscript  $(\ell, n)$  indicating that independent samples are used at each level of correction. The important point is that by using the same  $\omega^{(\ell,n)}$  we aim to ensure that  $P_\ell^{(\ell,n)} - P_{\ell-1}^{(\ell,n)}$  is small for larger values of  $\ell$ , so that relatively few samples are needed on finer levels to estimate  $\mathbb{E}[P_\ell - P_{\ell-1}]$ .

If we define  $C_0, V_0$  to be the cost and variance of one sample of  $P_0$ , and  $C_\ell, V_\ell$  to be the cost and variance of one sample of  $P_\ell - P_{\ell-1}$ , then the overall cost and variance of the multilevel estimator is  $\sum_{\ell=0}^L N_\ell C_\ell$  and  $\sum_{\ell=0}^L N_\ell^{-1} V_\ell$ , respectively. Ignoring the fact that the  $N_\ell$  are integers, for a fixed cost the variance is minimised by choosing  $N_\ell$  to minimise

$$\sum_{\ell=0}^L (N_\ell^{-1} V_\ell + \mu^2 N_\ell C_\ell)$$

for some value of the Lagrange multiplier  $\mu^2$ , which gives

$$(1) \quad N_\ell = \mu \sqrt{V_\ell / C_\ell}.$$

To achieve an overall variance of  $\varepsilon^2$  then requires that  $\mu = \varepsilon^{-2} \sum_{\ell=0}^L \sqrt{V_\ell C_\ell}$ .

Rounding up (1) to the nearest integer improves the overall variance and increases the cost by at most  $\sum_{\ell=0}^L C_\ell$ , so that the variance of  $\varepsilon^2$  can be achieved at a total cost which is bounded by

$$(2) \quad C = \varepsilon^{-2} \left( \sum_{\ell=0}^L \sqrt{V_\ell C_\ell} \right)^2 + \sum_{\ell=0}^L C_\ell.$$

It is important to note whether the product  $V_\ell C_\ell$  increases or decreases with  $\ell$ . If it increases, then the dominant contribution to the cost comes from  $V_L C_L$  and we have  $C \approx \varepsilon^{-2} V_L C_L$ , whereas if it decreases then the dominant contribution comes from  $V_0 C_0$  and  $C \approx \varepsilon^{-2} V_0 C_0$ . This contrasts with the standard Monte Carlo cost of approximately  $\varepsilon^{-2} V_0 C_L$ , assuming that the cost of computing  $P_L$  is similar to the cost of computing  $P_L - P_{L-1}$  and  $\mathbb{V}[P_L] \approx \mathbb{V}[P_0]$ .

This shows that in the first case the MLMC cost is reduced by factor  $V_L/V_0$ , corresponding to the ratio of the variances  $\mathbb{V}[P_L - P_{L-1}]$  and  $\mathbb{V}[P_L]$ , whereas in the second case it

is reduced by factor  $C_0/C_L$ , the ratio of the costs of computing  $P_0$  and  $P_L - P_{L-1}$ . If the product  $V_\ell C_\ell$  does not vary with level, then the total cost is approximately  $\varepsilon^{-2} L^2 V_0 C_0 = \varepsilon^{-2} L^2 V_L C_L$ .

**2.2 MLMC with inexact simulation.** In almost all MLMC applications, it is not possible to exactly simulate the quantity of interest  $P(\omega)$ , often because the calculation of  $P(\omega)$  requires the approximate solution of a PDE or SDE. Instead, what we have is an infinite sequence of approximations  $P_\ell, \ell = 0, 1, \dots$  which approximate  $P$  with increasing accuracy and cost. If  $Y$  is an approximation to  $\mathbb{E}[P]$ , then a standard piece of theory gives the MSE (mean square error) as

$$(3) \quad \text{MSE} \equiv \mathbb{E}[(Y - \mathbb{E}[P])^2] = \mathbb{V}[Y] + (\mathbb{E}[Y] - \mathbb{E}[P])^2.$$

If  $Y$  is now the multilevel estimator

$$(4) \quad Y = \sum_{\ell=0}^L Y_\ell, \quad Y_\ell = N_\ell^{-1} \sum_{n=1}^{N_\ell} (P_\ell^{(\ell,n)} - P_{\ell-1}^{(\ell,n)}),$$

with  $P_{-1} \equiv 0$ , then

$$(5) \quad \mathbb{E}[Y] = \mathbb{E}[P_L], \quad \mathbb{V}[Y] = \sum_{\ell=0}^L N_\ell^{-1} V_\ell, \quad V_\ell \equiv \mathbb{V}[P_\ell - P_{\ell-1}].$$

To ensure that the MSE is less than  $\varepsilon^2$ , it is sufficient to ensure that  $\mathbb{V}[Y]$  and  $(\mathbb{E}[P_L - P])^2$  are both less than  $\frac{1}{2}\varepsilon^2$ . Combining this idea with a geometric sequence of levels in which the cost increases exponentially with level, while both the weak error  $\mathbb{E}[P_L - P]$  and the multilevel correction variance  $V_\ell$  decrease exponentially, leads to the following theorem:

**Theorem 1.** *Let  $P$  denote a random variable, and let  $P_\ell$  denote the corresponding level  $\ell$  numerical approximation. If there exist independent estimators  $Y_\ell$  based on  $N_\ell$  Monte Carlo samples, each with expected cost  $C_\ell$  and variance  $V_\ell$ , and positive constants  $\alpha, \beta, \gamma, c_1, c_2, c_3$  such that  $\alpha \geq \frac{1}{2} \min(\beta, \gamma)$  and*

- i)  $|\mathbb{E}[P_\ell - P]| \leq c_1 2^{-\alpha \ell}$
- ii)  $\mathbb{E}[Y_\ell] = \begin{cases} \mathbb{E}[P_0], & \ell = 0 \\ \mathbb{E}[P_\ell - P_{\ell-1}], & \ell > 0 \end{cases}$
- iii)  $V_\ell \leq c_2 2^{-\beta \ell}$
- iv)  $C_\ell \leq c_3 2^{\gamma \ell}$ ,

then there exists a positive constant  $c_4$  such that for any  $\varepsilon < e^{-1}$  there are values  $L$  and  $N_\ell$  for which the multilevel estimator

$$Y = \sum_{\ell=0}^L Y_\ell,$$

has a mean-square-error with bound

$$MSE \equiv \mathbb{E} \left[ (Y - \mathbb{E}[P])^2 \right] < \varepsilon^2$$

with an expected computational complexity  $C$  with bound

$$C \leq \begin{cases} c_4 \varepsilon^{-2}, & \beta > \gamma, \\ c_4 \varepsilon^{-2} |\log \varepsilon|^2, & \beta = \gamma, \\ c_4 \varepsilon^{-2-(\gamma-\beta)/\alpha}, & \beta < \gamma. \end{cases}$$

The statement of the theorem is a slight generalisation of the original theorem in [Giles \[2008b\]](#). It corresponds to the theorem and proof in [Cliffe, Giles, Scheichl, and Teckentrup \[2011\]](#), except for the minor change to expected costs to allow for applications in which the simulation cost of individual samples is itself random. Note that if condition iii) is tightened slightly to be a bound on  $\mathbb{E}[(P_\ell - P_{\ell-1})^2]$ , which is usually the quantity which is bounded in numerical analysis, then it would follow immediately that  $\alpha \geq \frac{1}{2}\beta$ .

The essence of the proof is very straightforward. If we have  $V_\ell = O(2^{-\beta\ell})$  and  $C_\ell = O(2^{\gamma\ell})$ , then the analysis in [Section 2.1](#) shows that the optimal number of samples  $N_\ell$  on level  $\ell$  is proportional to  $2^{-(\beta+\gamma)\ell/2}$ , and therefore the total cost on level  $\ell$  is proportional to  $2^{(\gamma-\beta)\ell/2}$ . The result then follows from the requirement that  $L$  is chosen so that  $(\mathbb{E}[Y] - \mathbb{E}[P])^2 < \frac{1}{2}\varepsilon^2$ , and the constant of proportionality for  $N_\ell$  is chosen so that  $\mathbb{V}[Y] < \frac{1}{2}\varepsilon^2$ .

The result of the theorem merits some discussion. In the case  $\beta > \gamma$ , the dominant computational cost is on the coarsest levels where  $C_\ell = O(1)$  and  $O(\varepsilon^{-2})$  samples are required to achieve the desired accuracy. This is the standard result for a Monte Carlo approach using i.i.d. samples; to do better would require an alternative approach such as the use of Latin hypercube sampling or quasi-Monte Carlo methods.

In the case  $\beta < \gamma$ , the dominant computational cost is on the finest levels. Because of condition i), we have  $2^{-\alpha L} = O(\varepsilon)$ , and hence  $C_L = O(\varepsilon^{-\gamma/\alpha})$ . If  $\beta = 2\alpha$ , which is usually the best that can be achieved since typically  $\mathbb{V}[P_\ell - P_{\ell-1}]$  is similar in magnitude to  $\mathbb{E}[(P_\ell - P_{\ell-1})^2]$  which is greater than  $(\mathbb{E}[P_\ell - P_{\ell-1}])^2$ , then the total cost is  $O(C_L)$ , corresponding to  $O(1)$  samples on the finest level, which is the best that can be achieved.

The dividing case  $\beta = \gamma$  is the one for which both the computational effort, and the contributions to the overall variance, are spread approximately evenly across all of the levels; the  $|\log \varepsilon|^2$  term corresponds to the  $L^2$  factor in the corresponding discussion at the end of [Section 2.1](#).

One comment on the Theorem is that it assumes lots of properties, and then from these determines relatively easily some conclusions for the efficiency of the MLMC approach. In real applications, the tough challenge is in proving that the assumptions are valid, and in particular determining the values of the parameters  $\alpha, \beta, \gamma$ . Furthermore, the Theorem assumes knowledge of the constants  $c_1, c_2, c_3$ . In practice,  $c_1$  and  $c_2$  are almost never known, and instead have to be estimated based on empirical estimates of the weak error and the multilevel correction variance.

[Equation \(4\)](#) gives the natural choice for the multilevel correction estimator  $Y_\ell$ . However, the multilevel theorem allows for the use of other estimators, provided they satisfy the restriction of condition ii) which ensures that  $\mathbb{E}[Y] = \mathbb{E}[P_L]$ . Examples of this will be given later in this article. In each case, the objective in constructing a more complex estimator is to achieve a greatly reduced variance  $\mathbb{V}[Y_\ell]$  so that fewer samples are required.

**2.3 Randomised MLMC for unbiased estimation.** A very interesting extension was introduced by Rhee & Glynn in [Rhee and Glynn \[2015\]](#). Rather than choosing the finest level of simulation  $L$  based on the desired accuracy, and then using the optimal number of samples on each level based on an estimate of the variance, the “single term” estimator in [Rhee and Glynn \[ibid.\]](#) instead uses  $N$  samples in total, and for each sample the level on which the simulation is performed is selected randomly, with level  $\ell$  being chosen with probability  $p_\ell$ .

The estimator is

$$Y = \frac{1}{N} \sum_{n=1}^N \frac{1}{p_{\ell^{(n)}}} (P_{\ell^{(n)}}^{(n)} - P_{\ell^{(n)}-1}^{(n)})$$

with the level  $\ell^{(n)}$  for each sample being selected randomly with the relevant probability. Alternatively, their estimator can be expressed as

$$Y = \sum_{\ell=0}^{\infty} \left( \frac{1}{p_\ell N} \sum_{n=1}^{N_\ell} (P_\ell^{(n)} - P_{\ell-1}^{(n)}) \right).$$

where  $N_\ell$ , the number of samples from level  $\ell$ , is a random variable with

$$\sum_{\ell=0}^{\infty} N_\ell = N, \quad \mathbb{E}[N_\ell] = p_\ell N.$$

Note that in this form it is very similar in appearance to the standard MLMC estimator. The beauty of their estimator is that it is naturally unbiased, since

$$\begin{aligned}\mathbb{E}[Y] &= \mathbb{E}\left[\frac{1}{p_{\ell'}}(P_{\ell'} - P_{\ell'-1})\right] \\ &= \sum_{\ell=0}^{\infty} p_{\ell} \mathbb{E}\left[\frac{1}{p_{\ell'}}(P_{\ell'} - P_{\ell'-1}) \mid \ell' = \ell\right] = \sum_{\ell=0}^{\infty} \mathbb{E}[P_{\ell} - P_{\ell-1}] = \mathbb{E}[P].\end{aligned}$$

The choice of probabilities  $p_{\ell}$  is crucial. For both the variance and the expected cost to be finite, it is necessary that

$$\sum_{\ell=0}^{\infty} \frac{1}{p_{\ell}} V_{\ell} < \infty, \quad \sum_{\ell=0}^{\infty} p_{\ell} C_{\ell} < \infty.$$

Under the conditions of the MLMC Theorem, this is possible when  $\beta > \gamma$  by choosing  $p_{\ell} \propto 2^{-(\gamma+\beta)\ell/2}$ , so that

$$\frac{1}{p_{\ell}} V_{\ell} \propto 2^{-(\beta-\gamma)\ell/2}, \quad p_{\ell} C_{\ell} \propto 2^{-(\beta-\gamma)\ell/2}.$$

It is not possible when  $\beta \leq \gamma$ , and for these cases the estimators constructed in [Rhee and Glynn \[ibid.\]](#) have infinite expected cost.

**2.4 Multilevel Richardson-Romberg extrapolation.** Richardson extrapolation is a very old technique in numerical analysis. Given a numerical approximation  $P_h$  based on a discretisation parameter  $h$  which leads to an error

$$P_h - P = a h^{\alpha} + O(h^{2\alpha}),$$

it follows that  $P_{2h} - P = a (2h)^{\alpha} + O(h^{2\alpha})$ , and hence the extrapolated value

$$\tilde{P} = \frac{2^{\alpha}}{2^{\alpha}-1} P_h - \frac{1}{2^{\alpha}-1} P_{2h}$$

satisfies  $\tilde{P} - P = O(h^{2\alpha})$ . Lemaire & Pagès take this approach much further [Lemaire and Pagès \[2017\]](#). Assuming that the weak error has a regular expansion

$$\mathbb{E}[P_{\ell}] - \mathbb{E}[P] = \sum_{n=1}^L a_n 2^{-n\alpha\ell} + o(2^{-\alpha\ell L}),$$

they first determine the unique set of weights  $w_{\ell}$ ,  $\ell = 0, 1, \dots, L$  such that

$$\sum_{\ell=0}^L w_{\ell} = 1, \quad \sum_{\ell=0}^L w_{\ell} 2^{-n\alpha\ell} = 0, \quad n = 1, \dots, L,$$

so that

$$\left( \sum_{\ell=0}^L w_{\ell} \mathbb{E}[P_{\ell}] \right) - \mathbb{E}[P] \equiv \sum_{\ell=0}^L w_{\ell} (\mathbb{E}[P_{\ell}] - \mathbb{E}[P]) = o(2^{-\alpha L^2}).$$

Next, they re-arrange terms to give

$$\sum_{\ell=0}^L w_{\ell} \mathbb{E}[P_{\ell}] = \sum_{\ell=0}^L v_{\ell} \mathbb{E}[P_{\ell} - P_{\ell-1}]$$

where as usual  $P_{-1} \equiv 0$ , and the coefficients  $v_{\ell}$  are defined by  $w_{\ell} = v_{\ell} - v_{\ell+1}$ , with  $v_{L+1} \equiv 0$ , and hence

$$v_{\ell} = \sum_{\ell'=\ell}^L w_{\ell'}.$$

This leads to their Multilevel Richardson-Romberg extrapolation estimator,

$$Y = \sum_{\ell=0}^L Y_{\ell}, \quad Y_{\ell} = v_{\ell} N_{\ell}^{-1} \sum_n (P_{\ell}^{(\ell,n)} - P_{\ell-1}^{(\ell,n)}).$$

Because the remaining error is  $o(2^{-\alpha L^2})$ , rather than the usual  $O(2^{-\alpha L})$ , it is possible to obtain the usual  $O(\varepsilon)$  weak error with a value of  $L$  which is approximately the square root of the usual value. Hence, in the case  $\beta = \gamma$  they prove that the overall cost is reduced to  $O(\varepsilon^{-2} |\log \varepsilon|)$ , while for  $\beta < \gamma$  the cost is reduced much more to  $O(\varepsilon^{-2} 2^{(\gamma-\beta)\sqrt{|\log_2 \varepsilon|/\alpha}})$ . This analysis is supported by numerical results which demonstrate considerable savings [Lemaire and Pagès \[2017\]](#), and therefore this is a very useful extension to the standard MLMC approach when  $\beta \leq \gamma$ .

**2.5 Multi-Index Monte Carlo.** In standard MLMC, there is a one-dimensional set of levels, with a scalar level index  $\ell$ , although in some applications changing  $\ell$  can change more than one aspect of the computation (such as both timestep and spatial discretisation in a parabolic SPDE application, or timestep and number of sub-samples in a nested simulation). Multi-Index Monte Carlo developed by [Haji-Ali, Nobile, and Tempone \[2016\]](#) generalises this to “levels” being defined in multiple directions, so that the level “index”  $\ell$  is now a vector of integer indices. This is illustrated in [Figure 1](#) for a 2D MIMC application.

In MLMC, if we define the backward difference  $\Delta P_{\ell} \equiv P_{\ell} - P_{\ell-1}$  with  $P_{-1} \equiv 0$ , as usual, then the telescoping sum which lies at the heart of MLMC is

$$\mathbb{E}[P] = \sum_{\ell \geq 0} \mathbb{E}[\Delta P_{\ell}].$$

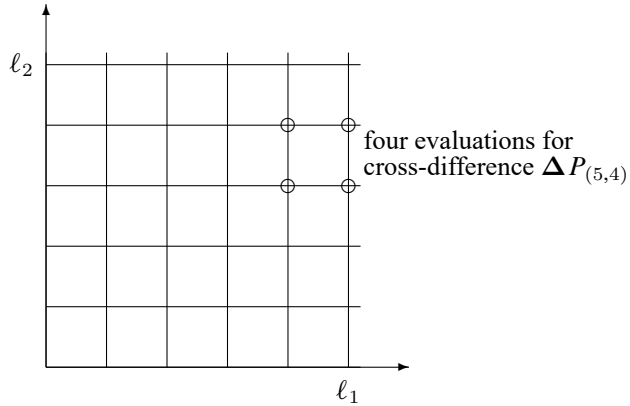


Figure 1: “Levels” in 2D multi-index Monte Carlo application

Generalising this to  $D$  dimensions, we can first define a backward difference operator in one particular dimension,  $\Delta_d P_\ell \equiv P_\ell - P_{\ell - e_d}$  where  $e_d$  is the unit vector in direction  $d$ . Then defining the cross-difference

$$\Delta P_\ell \equiv \left( \prod_{d=1}^D \Delta_d \right) P_\ell$$

the telescoping sum becomes

$$\mathbb{E}[P] = \sum_{\ell \geq 0} \mathbb{E}[\Delta P_\ell].$$

As an example, [Figure 1](#) marks the four locations at which  $P_\ell$  must be computed to determine the value of  $\Delta P_{(5,4)}$  in the 2D application.

Instead of summing  $\mathbb{E}[\Delta P_\ell]$  over the full domain  $\ell \geq 0$ , the sum is instead truncated to a summation region  $\mathcal{L}$ . It might seem natural that this should be rectangular, as illustrated on the left in [Figure 2](#), so that

$$\sum_{\ell \in \mathcal{L}} \Delta P_\ell = P_L$$

where  $L$  is the outermost point on the rectangle. However, [Haji-Ali, Nobile, and Tempone \[ibid.\]](#) proves that it is often better to use a region  $\mathcal{L}$  of the form  $\ell \cdot \mathbf{n} \leq L$  for a particular choice of direction vector  $\mathbf{n}$  with strictly positive components. In 2D, this corresponds to a triangular region, as illustrated on the right in [Figure 2](#). This is very similar to the use of the sparse grid combination technique in high-dimensional PDE approximations [Bungartz](#)



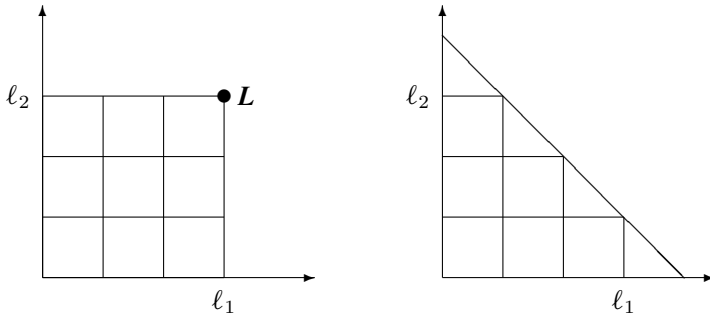


Figure 2: Two choices of 2D MIMC summation region  $\mathcal{L}$ .

and Griebel [2004], and indeed MIMC can be viewed as a combination of this approach with Monte Carlo sampling.

The benefits of MIMC over the standard MLMC can be very substantial. They are perhaps best illustrated by an elliptic PDE or SPDE example, in which  $D$  corresponds to the number of spatial dimensions. Using the standard MLMC approach,  $\beta$ , the rate of convergence of the multilevel variance, will usually be independent of  $D$ , but  $\gamma$ , the rate of increase in the computational cost, will increase at least linearly with  $D$ . Therefore, in a high enough dimension we will have  $\beta \leq \gamma$  and therefore the overall computational complexity will be less (often much less) than the optimal  $O(\varepsilon^{-2})$ . However, using MIMC it is possible to achieve the optimal complexity independent of the value of  $D$ . Hence, in the same way that sparse grids offer the possibility of dimension-independent complexity for deterministic PDE applications, MIMC offers the possibility of dimension-independent complexity for SPDEs and other high-dimensional stochastic applications.

MIMC is the first multi-dimensional generalisation of MLMC, but it is not the only one. Other possibilities include using sparse grid samples within a standard MLMC formulation, and nested MLMC in which there is an outer MLMC telescoping sum in one “direction”, and then each of its expectations is expressed as an MLMC telescoping sum in a second “direction”. These ideas, and the inclusion of quasi-Monte Carlo sampling, are discussed in Giles, Kuo, and Sloan [2018].

**2.6 MLQMC.** The final part of this theory section concerns the use of quasi-Monte Carlo (QMC) sampling in place of standard Monte Carlo. The key change in MLQMC is that  $N_\ell$  is now the size of a set of QMC points used on level  $\ell$ . This set of points is not constructed randomly and independently, but is instead constructed very carefully to provide a relatively uniform coverage of a unit hypercube integration region which is then mapped into the required domain, for example by mapping unit interval uniformly

distributed random variables to standard Normal random variables [Dick, Kuo, and Sloan \[2013\]](#). In the best cases, this results in the numerical integration error being approximately  $O(N_\ell^{-1})$  rather than the usual  $O(N_\ell^{-1/2})$  error which comes from Monte Carlo sampling.

Using just one set of  $N_\ell$  points gives good accuracy, but no confidence interval. To regain a confidence interval one uses randomised QMC in which the set of points is collectively randomised in a way which ensures that the averages obtained from each set of points are independent. Using 32 randomisations, for example, yields 32 set averages for the quantity of interest,  $Y_\ell$ , and from these the variance of their average,  $V_\ell$ , can be estimated in the usual way.

Since  $V_\ell$  is now defined to be the variance of the average of the set averages, the aim now is to choose the  $N_\ell$  in a way which ensures that

$$(6) \quad \sum_{\ell=0}^L V_\ell \leq \frac{1}{2} \varepsilon^2.$$

We can not use the same Lagrange multiplier approach as before to determine the optimal  $N_\ell$ . Instead, we note that many QMC methods work naturally with  $N_\ell$  as a power of 2. Doubling  $N_\ell$  will usually eliminate a large fraction of the variance, so the greatest reduction in total variance relative to the additional computational effort is achieved by doubling  $N_\ell$  on the level  $\ell^*$  given by

$$(7) \quad \ell^* = \arg \max_{\ell} \frac{V_\ell}{N_\ell C_\ell}.$$

This approach was first developed in [Giles and Waterhouse \[2009\]](#), with application to stochastic differential equations (SDEs) using QMC samples based on extensible rank-1 lattices [Dick, Pillichshammer, and Waterhouse \[2008\]](#). QMC is known to be most effective for low-dimensional applications, and the numerical results were very encouraging for SDE applications in which the dominant computational cost was on the coarsest levels of resolution. However, there was no supporting theory for this research. More recently, there has been considerable research on applications and the underlying theoretical foundations for MLQMC methods applied to PDEs with stochastic coefficients [Niu, Hickernell, Müller-Gronbach, and Ritter \[2011\]](#), [Kuo, Schwab, and Sloan \[2015\]](#), and [Dick, Kuo, and Sloan \[2013\]](#). These theoretical developments are very encouraging, showing that under certain conditions they lead to multilevel methods with a complexity which is  $O(\varepsilon^{-p})$  with  $p < 2$ .

### 3 SDEs

The original multilevel path simulation paper [Giles \[2008b\]](#) treated stochastic differential equations

$$dS_t = a(S_t, t) dt + b(S_t, t) dW,$$

using the simple Euler-Maruyama discretisation with a uniform timestep  $h$  and Brownian increments  $\Delta W_n$ ,

$$\widehat{S}_{(n+1)h} = \widehat{S}_{nh} + a(\widehat{S}_{nh}, nh) h + b(\widehat{S}_{nh}, nh) \Delta W_n.$$

The multilevel Monte Carlo implementation is very simple. On level  $\ell$ , the uniform timestep is taken to be  $h_\ell = M^{-\ell} h_0$ , for some integer  $M$ . The timestep  $h_0$  on the coarsest level is often taken to be the interval length  $T$ , so that there is just one timestep for the entire interval, but this is not required, and in some applications using such a large timestep may lead to numerical results which are so inaccurate that they are not helpful in reducing the variance.

The multilevel coupling is achieved by using the same underlying driving Brownian path for the coarse and fine paths; this is accomplished by summing the Brownian increments for the fine path timesteps to obtain the Brownian increments for the coarse timesteps. The multilevel estimator is then the natural one defined in (4), with the specific payoff approximation  $P_\ell$  depending on the particular application.

Provided the SDE satisfies the usual conditions (see Theorem 10.2.2 in [Kloeden and Platen \[1992\]](#)), the strong error for the Euler discretisation with timestep  $h$  is  $O(h^{1/2})$ , so that

$$\mathbb{E} \left[ \sup_{[0, T]} \|S_t - \widehat{S}_t\|^2 \right] = O(h),$$

where  $\widehat{S}_t$  is a piecewise constant interpolation of the discrete values  $\widehat{S}_{nh}$ .

For financial options for which the payoff is a Lipschitz function of  $S_t$ , with constant  $K$ , we have

$$\mathbb{V}[P - P_\ell] \leq \mathbb{E}[(P - P_\ell)^2] \leq K^2 \mathbb{E} \left[ \sup_{[0, T]} \|S_t - \widehat{S}_t\|^2 \right],$$

where  $K$  is the Lipschitz constant, and

$$V_\ell \equiv \mathbb{V}[P_\ell - P_{\ell-1}] \leq 2 \left( \mathbb{V}[P - P_\ell] + \mathbb{V}[P - P_{\ell-1}] \right),$$

and hence  $V_\ell = O(h_\ell)$ .

option	Euler-Maruyama		Milstein	
	numerics	analysis	numerics	analysis
Lipschitz	$O(h)$	$O(h)$	$O(h^2)$	$O(h^2)$
Asian	$O(h)$	$O(h)$	$O(h^2)$	$O(h^2)$
lookback	$O(h)$	$O(h)$	$O(h^2)$	$o(h^{2-\delta})$
barrier	$O(h^{1/2})$	$o(h^{1/2-\delta})$	$O(h^{3/2})$	$o(h^{3/2-\delta})$
digital	$O(h^{1/2})$	$O(h^{1/2} \log h )$	$O(h^{3/2})$	$o(h^{3/2-\delta})$

Table 1: Observed and theoretical convergence rates for the multilevel correction variance for scalar SDEs, using the Euler-Maruyama and Milstein discretisations.  $\delta$  is any strictly positive constant.

If  $h_\ell = 4^{-\ell}h_0$ , as in the numerical examples in [Giles \[2008b\]](#), then this gives  $\alpha = 2$ ,  $\beta = 2$  and  $\gamma = 2$ . This is found to be better than using  $h_\ell = 2^{-\ell}h_0$  with twice as many timesteps on each successive level, which gives  $\alpha = 1$ ,  $\beta = 1$  and  $\gamma = 1$ . In either case, [Theorem 1](#) gives the complexity to achieve a RMS error of  $\varepsilon$  to be  $O(\varepsilon^{-2}|\log \varepsilon|^2)$ , which has been proved to be optimal for a class of Lipschitz path-dependent functions [Creutzig, Dereich, Müller-Gronbach, and Ritter \[2009\]](#).

The more accurate Milstein approximation achieves first order strong convergence, giving  $V_\ell = O(h_\ell^2)$  for certain Lipschitz payoff functions [Giles \[2008a\]](#). Further challenges are encountered with digital and barrier options for which the payoff functions are a discontinuous function of the path  $S_t$ . In such cases, a small difference between the coarse and fine path approximations can nevertheless produce a large value for  $P_\ell - P_{\ell-1}$ . Techniques have been developed to partially address this [Giles \[ibid.\]](#). [Table 1](#) summarises the observed variance convergence rate in numerical experiments for a number of different financial options; the Asian option is based on the average value of the underlying asset, the lookback is based on its maximum or minimum value, the barrier is a discontinuous function of the maximum or minimum, and the digital is a discontinuous function of the final value. The table also displays the theoretical numerical analysis results which have been obtained [Avikainen \[2009\]](#), [Giles, Higham, and Mao \[2009\]](#), and [Giles, Debrabant, and Rößler \[2013\]](#).

There is insufficient space in this article to discuss in detail the many other extensions and generalisations in applying MLMC to SDEs. We will simply mention a few, and further details and references can be obtained from [Giles \[2015\]](#).

There are difficulties in implementing the Milstein approximation for multi-dimensional SDEs when they require the simulation of Lévy areas. In this case, there is a special “antithetic” MLMC estimator which eliminates to leading order the error due to the omission of the Lévy areas.

The classic analysis of SDE approximations assumes that the drift function  $a(S_t, t)$  and volatility  $b(S_t, t)$  are both globally Lipschitz functions of  $S_t$ . Some important applications have drift functions, such as  $-S_t - S_t^3$ , which are only locally Lipschitz. These require adaptive timestepping, or some other technique, to maintain numerical stability, and this causes additional difficulties for MLMC.

There are also extensions to jump-diffusion SDEs, in which there is an additional Poisson jump process, and SDEs driven by increments of a more general Lévy process instead of a Brownian motion.

Finally, in some applications the output quantity of interest is not the expected value of a scalar quantity but a function such as the density of that output, or the cumulative distribution function,  $\text{CDF}(x) = \mathbb{P}[X < x] = \mathbb{E}[H(x - X)]$  where  $H(x)$  is the Heaviside function.

## 4 PDEs and SPDEs

Applying MLMC to stochastic PDEs and PDEs with random data or stochastic coefficients was a natural follow-on to the use for SDEs. Indeed, there was more scope for computational savings because the cost of a single sample increases more rapidly with grid resolution for SPDEs with higher space-time dimension. There has been a variety of papers on elliptic [Barth, Schwab, and Zollinger \[2011\]](#) and [Cliffe, Giles, Scheichl, and Teckentrup \[2011\]](#), parabolic [Barth, Lang, and Schwab \[2013\]](#) and [Giles and Reisinger \[2012\]](#) and hyperbolic [Mishra, Schwab, and Šukys \[2012\]](#) PDEs and SPDEs, as well as for mixed elliptic-hyperbolic systems [Efendiev, Iliev, and Kronsbein \[2013\]](#) and [Müller, Jenny, and Meyer \[2013\]](#).

In almost all of this work, the construction of the multilevel estimator is quite natural, using a geometric sequence of grids and the natural estimators for  $P_\ell - P_{\ell-1}$ . It is the numerical analysis of the variance of the multilevel estimator which is often very challenging, but in the simplest cases it can be straightforward.

Consider, for example, a  $D$ -dimensional elliptic PDE,  $\nabla^2 u = f(x, \omega)$ , where the r.h.s. forcing term is stochastic, depending on a number of random variables. If  $f$  is sufficiently smooth, then a standard piecewise linear finite element method might achieve second order accuracy for a large class of output functionals. Hence, if the computational grid has spacing proportional to  $2^{-\ell}$  in each direction then the error would be  $O(2^{-2\ell})$  and the MLMC variance  $V_\ell$  would be  $O(2^{-4\ell})$ . Using an efficient multigrid solver, the computational cost would be approximately proportional to the total number of grid points, which is  $O(2^{D\ell})$ . Hence this MLMC application has  $\alpha = 2, \beta = 4, \gamma = D$ .

This means that a RMS accuracy of  $\varepsilon$  can be achieved at  $O(\varepsilon^{-2})$  cost for  $D < 4$ , while the cost is  $O(\varepsilon^{-2} |\log \varepsilon|^2)$  for  $D = 4$ , and  $O(\varepsilon^{-D/2})$  for  $D > 4$ . By comparison, to achieve

an accuracy of  $\varepsilon$  for a single deterministic calculation requires  $2^{-2\ell} \sim \varepsilon$ , and hence the cost is  $O(\varepsilon^{-D/2})$ , which for  $D > 4$  is the same order as the cost of estimating the expectation in the random setting.

The largest amount of research on multilevel for SPDEs has been for elliptic PDEs with random coefficients. The PDE typically has the form

$$-\nabla \cdot (\kappa(\mathbf{x}, \omega) \nabla p(\mathbf{x}, \omega)) = 0, \quad \mathbf{x} \in D.$$

with Dirichlet or Neumann boundary conditions on the boundary  $\partial D$ . For sub-surface flow problems, such as the modelling of groundwater flow in nuclear waste repositories, the diffusivity (or permeability)  $\kappa$  is often modelled as a lognormal random field, i.e.  $\log \kappa$  is a Gaussian field with a uniform mean and a covariance function  $R(\mathbf{x}, \mathbf{y})$ . Samples of  $\log \kappa$  can be provided by a Karhunen-Loève expansion:

$$\log \kappa(\mathbf{x}, \omega) = \sum_{n=0}^{\infty} \sqrt{\theta_n} \xi_n(\omega) f_n(\mathbf{x}),$$

where  $\theta_n$  and  $f_n$  are the eigenvalues and eigenfunctions defined by

$$\int R(\mathbf{x}, \mathbf{y}) f_n(\mathbf{y}) \, \mathrm{d}\mathbf{y} = \theta_n f_n(\mathbf{x}),$$

and  $\xi_n$  are independent unit Normal random variables. However, it can be more efficient to generate them using a circulant embedding technique based on the use of FFTs.

There is no space to detail the huge range of other applications. The one other point to note here is that they are not all based on a geometric hierarchy of approximations. A non-geometric example is the use of a reduced basis approximation in which the approximate solution  $u$  at a set of discrete grid points for an arbitrary set of random inputs  $\omega$  is written as

$$u(\omega) = \sum_{k=1}^K c_k(\omega) u_k$$

where the  $c_k$  are a set of scalar coefficients, and the  $u_k$  are a fixed set of vectors, for example corresponding to solutions of the PDE for particular values of the random inputs. In such a reduced basis approximation, the accuracy improves if one increases the number of basis functions,  $K$ , but so too does the cost. Neither behaves in a simple way such that there is an obvious way in which to prescribe  $K_\ell$  as a function of level, and therefore numerical optimisation can be used instead [Vidal-Codina, Nguyen, Giles, and Peraire \[2015\]](#).

## 5 Continuous-time Markov chains

A very interesting and important application of MLMC has been to continuous-time Markov Chain simulation [Anderson and Higham \[2012\]](#). Such models arise in the context of stochastic chemical reactions, when species concentrations are extremely low and so stochastic effects become significant. When there is just one chemical species which is being spontaneously created at a rate which depends on the current number of molecules  $x$ , the “tau-leaping” method (which is essentially the Euler-Maruyama method, approximating the reaction rate as being constant throughout the timestep) gives the discrete equation

$$x_{n+1} = x_n + P(h\lambda(x_n)),$$

where the state  $x_n$  is an integer,  $h$  is the timestep,  $\lambda(x_n)$  is the reaction rate (or propensity function), and  $P(t)$  represents a unit-rate Poisson random variable over time interval  $[0, t]$ . If this equation defines the fine path in the multilevel simulation, then the coarse path, with double the timestep, is given by

$$x_{n+2}^c = x_n^c + P(2h\lambda(x_n^c))$$

for even timesteps  $n$ .

The question then is how to couple the coarse and fine path simulations in a MLMC calculation. The key observation in [Anderson and Higham \[ibid.\]](#), is that for any  $t_1, t_2 > 0$ , the sum of two independent Poisson variates  $P(t_1)$ ,  $P(t_2)$  is equivalent in distribution to  $P(t_1 + t_2)$ . Based on this, the first step is to express the coarse path Poisson variate as the sum of two independent Poisson variates,  $P(h\lambda(\mathbf{x}_n^c))$  corresponding to the first and second fine path timesteps. For the first of the two fine timesteps, the coarse and fine path Poisson variates are coupled by defining two Poisson variates based on the minimum of the two reactions rates, and the absolute difference,

$$P_1 = P\left(h \min(\lambda(\mathbf{x}_n), \lambda(\mathbf{x}_n^c))\right), \quad P_2 = P\left(h |\lambda(\mathbf{x}_n) - \lambda(\mathbf{x}_n^c)|\right),$$

and then using  $P_1$  as the Poisson variate for the path with the smaller rate, and  $P_1 + P_2$  for the path with the larger rate. This elegant approach naturally gives a small difference in the Poisson variates when the difference in rates is small, and leads to a very effective multilevel algorithm with a correction variance which is  $O(h)$ , leading to an  $O(\varepsilon^{-2} |\log \varepsilon|^2)$  complexity.

In their paper [Anderson and Higham \[ibid.\]](#), Anderson & Higham treat more general systems with multiple species and multiple reactions. They also include an additional coupling at the finest level to the exact Stochastic Simulation Algorithm developed by [Gillespie \[1976\]](#) which updates the reaction rates after every single reaction. Hence, their

overall multilevel estimator is unbiased, unlike the estimators discussed earlier for SDEs, and the complexity is reduced to  $O(\varepsilon^{-2})$  because the number of levels remains fixed as  $\varepsilon \rightarrow 0$ . They give a complete numerical analysis of the variance of their multilevel algorithm; this has been further sharpened in more recent work [Anderson, Higham, and Sun \[2014\]](#). Because stochastic chemical simulations typically involve 1000's of reactions, the multilevel method is particularly effective in this context, providing computational savings in excess of a factor of 100 [Anderson and Higham \[2012\]](#).

They also give an interesting numerical example in which an approximate model with fewer reactions/reactants is used as a control variate for the full system. This kind of multilevel modelling is another possibility which could be considered in a wide variety of circumstances.

## 6 Nested simulation

In nested simulations we are interested in estimating quantities of the form

$$\mathbb{E}_Z \left[ f \left( \mathbb{E}_W [g(Z, W)] \right) \right]$$

where  $\mathbb{E}_Z$  represents an expectation with respect to  $Z$ , an outer random variable, and  $\mathbb{E}_W [g(Z, W)]$  is a conditional expectation with respect to an independent inner random variable  $W$ . For example, in some financial applications,  $Z$  represents different risk *scenarios*,  $\mathbb{E}_W [g(Z, W)]$  represents the conditional value of a portfolio, and  $f$  corresponds to the loss in excess of a certain level, so that  $\mathbb{E}_Z \left[ f \left( \mathbb{E}_W [g(Z, W)] \right) \right]$  is the expected shortfall.

This can be simulated using nested Monte Carlo simulation with  $N$  outer samples  $Z^{(n)}$ ,  $M$  inner samples  $W^{(m,n)}$  and a standard Monte Carlo estimator:

$$Y = N^{-1} \sum_{n=1}^N f \left( M^{-1} \sum_{m=1}^M g(Z^{(n)}, W^{(m,n)}) \right)$$

Note that to improve the accuracy of the estimate we need to increase both  $M$  and  $N$ , and this will significantly increase the cost.

An MLMC implementation is straightforward; on level  $\ell$  we can use  $M_\ell = 2^\ell$  inner samples. To construct a low variance estimate for  $\mathbb{E}[P_\ell - P_{\ell-1}]$  where

$$\mathbb{E}[P_\ell] \equiv \mathbb{E}_Z \left[ f \left( M_\ell^{-1} \sum_m g(Z, W^{(m)}) \right) \right],$$



we can use an *antithetic* approach and split the  $M_\ell$  samples of  $W$  for the “fine” value into two subsets of size  $M_{\ell-1}$  for the “coarse” value:

$$Y_{\ell=N_\ell^{-1} \sum_{n=1}^{N_\ell} \left\{ f \left( M_\ell^{-1} \sum_{m=1}^{M_\ell} g(Z^{(n)}, W^{(m,n)}) \right) - \frac{1}{2} f \left( M_{\ell-1}^{-1} \sum_{m=1}^{M_{\ell-1}} g(Z^{(n)}, W^{(m,n)}) \right) - \frac{1}{2} f \left( M_{\ell-1}^{-1} \sum_{m=M_{\ell-1}+1}^{M_\ell} g(Z^{(n)}, W^{(m,n)}) \right) \right\}}$$

Note that this has the correct expectation, i.e.  $\mathbb{E}[Y_\ell] = \mathbb{E}[P_\ell - P_{\ell-1}]$ .

If we now define

$$M_{\ell-1}^{-1} \sum_{m=1}^{M_{\ell-1}} g(Z^{(n)}, W^{(m,n)}) = \mathbb{E}[g(Z^{(n)}, W)] + \Delta g_1^{(n)},$$

$$M_{\ell-1}^{-1} \sum_{m=M_{\ell-1}+1}^{M_\ell} g(Z^{(n)}, W^{(m,n)}) = \mathbb{E}[g(Z^{(n)}, W)] + \Delta g_2^{(n)},$$

then if  $f$  is twice differentiable a Taylor series expansion gives

$$Y_\ell \approx -\frac{1}{4N_\ell} \sum_{n=1}^{N_\ell} f'' \left( \mathbb{E}[g(Z^{(n)}, W)] \right) \left( \Delta g_1^{(n)} - \Delta g_2^{(n)} \right)^2$$

By the Central Limit Theorem,  $\Delta g_1^{(n)}, \Delta g_2^{(n)} = O(M_\ell^{-1/2})$  and therefore

$$f'' \left( \mathbb{E}[g(Z^{(n)}, W)] \right) \left( \Delta g_1^{(n)} - \Delta g_2^{(n)} \right)^2 = O(M_\ell^{-1}).$$

It follows that  $\mathbb{E}[Y_\ell] = O(M_\ell^{-1})$  and  $V_\ell = O(M_\ell^{-2})$ . For the MLMC theorem, this corresponds to  $\alpha = 1$ ,  $\beta = 2$ ,  $\gamma = 1$ , so the complexity is  $O(\varepsilon^{-2})$ .

This approach has been used for a financial credit derivative application [Bujok, Hamblly, and Reisinger \[2015\]](#), but in that case the function  $f$  was piecewise linear, not twice differentiable, and so the rate of variance convergence was slightly lower, with  $\beta = 1.5$ . However, this is still sufficiently large to achieve an overall complexity which is  $O(\varepsilon^{-2})$ .

Current research in this area is addressing the challenges of functions  $f$  which are discontinuous, and Multi-Index Monte Carlo or nested MLMC for applications in which there are additional “dimensions” to the problem, such as the number of timesteps in an SDE simulation in the inner conditional expectation.

## 7 Variable precision arithmetic

This final category of applications is included to illustrate the flexibility and generality of the MLMC approach. In the latest Intel CPUs, each core has a vector unit which can perform 16 single precision or 8 double precision operations with one instruction. Hence, single precision computations can be twice as fast as double precision on CPUs. The latest GPUs (graphics processing units) take this idea even further, including a half-precision capability which is twice as fast as single-precision.

This leads naturally to the idea of a 2-level MLMC calculation on CPUs, or a 3-level calculation on GPUs, with the different levels corresponding to different levels of floating point precision. To ensure that the MLMC telescoping sum is correctly respected, all MLMC summations should be performed in double precision, which we can view as being “exact”. It is also important that the random numbers are generated consistently, so that the distribution of half-precision random numbers used on level 0, is equivalent to the distribution of “coarse sample” half-precision random numbers obtained on level 1 by first generating single-precision random numbers are then truncating them down to half-precision.

This approach has been generalised in research which exploits FPGAs (field-programmable gate arrays) which can perform computations with a user-specified number of bits to represent floating-point or fixed-point numbers. Thus, it is possible to implement a multilevel treatment in which the number of bits used increases with level [Brugger, de Schryver, Wehn, Omland, Hefter, Ritter, Kostiuk, and Korn \[2014\]](#).

## 8 Conclusions

The last ten years has seen considerable progress in the theoretical development, application and analysis of multilevel Monte Carlo methods. On the theoretical side, the key extensions are to unbiased randomised estimators for applications with a rapid rate of variance convergence; Richardson-Romberg extrapolation for improved computational efficiency when the rate of variance convergence is low; and multi-index Monte Carlo (MIMC), generalising multilevel to multiple “directions” in which approximations can be refined. On the practical side, the range of applications is growing steadily, including the examples given in this article and others such as reliability and rare event simulation, and MCMC and Bayesian inverse methods. There has also been excellent progress on the numerical analysis of the MLMC variances for the full range of applications.

This review has attempted to emphasise the conceptual simplicity of the multilevel approach; in essence it is simply a recursive control variate strategy, using cheap inaccurate approximations to some random output quantity as a control variate for more accurate but more costly approximations. In practice, the challenge is first to develop a tight coupling

between successive approximation levels, to minimise the variance of the difference in the output obtained from each level, and then to develop a corresponding numerical analysis.

**Acknowledgments.** The author is very grateful to the many collaborators and students he has worked with on MLMC applications over the past 10 years, and in particular to Dr. Abdul-Lateef Haji-Ali for his comments on the paper.

## References

- David F. Anderson and Desmond J. Higham (2012). “Multilevel Monte Carlo for continuous time Markov chains, with applications in biochemical kinetics”. *Multiscale Model. Simul.* 10.1, pp. 146–179. MR: [2902602](#) (cit. on pp. [3604](#), [3605](#)).
- David F. Anderson, Desmond J. Higham, and Yu Sun (2014). “Complexity of multilevel Monte Carlo tau-leaping”. *SIAM J. Numer. Anal.* 52.6, pp. 3106–3127. MR: [3504598](#) (cit. on p. [3605](#)).
- Rainer Avikainen (2009). “On irregular functionals of SDEs and the Euler scheme”. *Finance Stoch.* 13.3, pp. 381–401. MR: [2519837](#) (cit. on p. [3601](#)).
- Ivo Babuška, Fabio Nobile, and Raúl Tempone (2010). “A stochastic collocation method for elliptic partial differential equations with random input data”. *SIAM Rev.* 52.2, pp. 317–355. MR: [2646806](#) (cit. on p. [3589](#)).
- Ivo Babuška, Raúl Tempone, and Georgios E. Zouraris (2004). “Galerkin finite element approximations of stochastic elliptic partial differential equations”. *SIAM J. Numer. Anal.* 42.2, pp. 800–825. MR: [2084236](#) (cit. on p. [3589](#)).
- Andrea Barth, Annika Lang, and Christoph Schwab (2013). “Multilevel Monte Carlo method for parabolic stochastic partial differential equations”. *BIT* 53.1, pp. 3–27. MR: [3029293](#) (cit. on p. [3602](#)).
- Andrea Barth, Christoph Schwab, and Nathaniel Zollinger (2011). “Multi-level Monte Carlo finite element method for elliptic PDEs with stochastic coefficients”. *Numer. Math.* 119.1, pp. 123–161. MR: [2824857](#) (cit. on p. [3602](#)).
- Christian Brügger, Christian de Schryver, Norbert Wehn, Steffen Omland, Mario Hefter, Klaus Ritter, Anton Kostiuk, and Ralf Korn (2014). “Mixed precision multilevel Monte Carlo on hybrid computing systems”. In: *Computational Intelligence for Financial Engineering & Economics (CIFER), 2104 IEEE Conference on*. IEEE, pp. 215–222 (cit. on p. [3607](#)).
- K. Bujok, B. M. Hambly, and C. Reisinger (2015). “Multilevel simulation of functionals of Bernoulli random variables with application to basket credit derivatives”. *Methodol. Comput. Appl. Probab.* 17.3, pp. 579–604. MR: [3377850](#) (cit. on p. [3606](#)).

- Hans-Joachim Bungartz and Michael Griebel (2004). “Sparse grids”. *Acta Numer.* 13, pp. 147–269. MR: [2249147](#) (cit. on p. [3597](#)).
- K. A. Cliffe, Michael B. Giles, R. Scheichl, and A. L. Teckentrup (2011). “Multilevel Monte Carlo methods and applications to elliptic PDEs with random coefficients”. *Comput. Vis. Sci.* 14.1, pp. 3–15. MR: [2835612](#) (cit. on pp. [3593](#), [3602](#)).
- Jakob Creutzig, Steffen Dereich, Thomas Müller-Gronbach, and Klaus Ritter (2009). “Infinite-dimensional quadrature and approximation of distributions”. *Found. Comput. Math.* 9.4, pp. 391–429. MR: [2519865](#) (cit. on p. [3601](#)).
- Josef Dick, Frances Y. Kuo, and Ian H. Sloan (2013). “High-dimensional integration: the quasi-Monte Carlo way”. *Acta Numer.* 22, pp. 133–288. MR: [3038697](#) (cit. on pp. [3590](#), [3599](#)).
- Josef Dick, Friedrich Pillichshammer, and Benjamin J. Waterhouse (2008). “The construction of good extensible rank-1 lattices”. *Math. Comp.* 77.264, pp. 2345–2373. MR: [2429889](#) (cit. on p. [3599](#)).
- Y. Efendiev, O. Iliev, and C. Kronsbein (2013). “Multilevel Monte Carlo methods using ensemble level mixed MsFEM for two-phase flow and transport simulations”. *Comput. Geosci.* 17.5, pp. 833–850. MR: [3104637](#) (cit. on p. [3602](#)).
- Michael B. Giles (2008a). “Improved multilevel Monte Carlo convergence using the Milstein scheme”. In: *Monte Carlo and quasi-Monte Carlo methods 2006*. Springer, Berlin, pp. 343–358. MR: [2479233](#) (cit. on p. [3601](#)).
- (2008b). “Multilevel Monte Carlo path simulation”. *Oper. Res.* 56.3, pp. 607–617. MR: [2436856](#) (cit. on pp. [3593](#), [3600](#), [3601](#)).
- (2015). “Multilevel Monte Carlo methods”. *Acta Numer.* 24, pp. 259–328. MR: [3349310](#) (cit. on pp. [3590](#), [3601](#)).
- Michael B. Giles, Kristian Debrabant, and Andreas Rößler (Feb. 2013). “Numerical analysis of multilevel Monte Carlo path simulation using the Milstein discretisation”. arXiv: [1302.4676](#) (cit. on p. [3601](#)).
- Michael B. Giles, Desmond J. Higham, and Xuerong Mao (2009). “Analysing multi-level Monte Carlo for options with non-globally Lipschitz payoff”. *Finance Stoch.* 13.3, pp. 403–413. MR: [2519838](#) (cit. on p. [3601](#)).
- Michael B. Giles, Frances Y. Kuo, and I. H. Sloan (2018). “Combining sparse grids, multilevel MC and QMC for elliptic PDEs with random coefficients”. In: *Monte Carlo and quasi-Monte Carlo methods 2016*. Ed. by P.W. Glynn and A. Owen. Springer Proc. Math. Stat. Springer, Heidelberg (cit. on p. [3598](#)).
- Michael B. Giles and Christoph Reisinger (2012). “Stochastic finite differences and multilevel Monte Carlo for a class of SPDEs in finance”. *SIAM J. Financial Math.* 3.1, pp. 572–592. MR: [2968046](#) (cit. on p. [3602](#)).

- Michael B. Giles and Benjamin J. Waterhouse (2009). “Multilevel quasi-Monte Carlo path simulation”. In: *Advanced financial modelling*. Vol. 8. Radon Ser. Comput. Appl. Math. Walter de Gruyter, Berlin, pp. 165–181. MR: [2648461](#) (cit. on p. [3599](#)).
- Daniel T. Gillespie (1976). “A general method for numerically simulating the stochastic time evolution of coupled chemical reactions”. *J. Computational Phys.* 22.4, pp. 403–434. MR: [0503370](#) (cit. on p. [3604](#)).
- Max D. Gunzburger, Clayton G. Webster, and Guannan Zhang (2014). “Stochastic finite element methods for partial differential equations with random input data”. *Acta Numer.* 23, pp. 521–650. MR: [3202242](#) (cit. on p. [3589](#)).
- Abdul-Lateef Haji-Ali, Fabio Nobile, and Raúl Tempone (2016). “Multi-index Monte Carlo: when sparsity meets sampling”. *Numerische Mathematik* 132.4, pp. 767–806. MR: [3474489](#) (cit. on pp. [3596](#), [3597](#)).
- Peter E. Kloeden and Eckhard Platen (1992). *Numerical solution of stochastic differential equations*. Vol. 23. Applications of Mathematics (New York). Springer-Verlag, Berlin, pp. xxxvi+632. MR: [1214374](#) (cit. on p. [3600](#)).
- Frances Y. Kuo, Christoph Schwab, and Ian H. Sloan (2015). “Multi-level quasi-Monte Carlo finite element methods for a class of elliptic PDEs with random coefficients”. *Found. Comput. Math.* 15.2, pp. 411–449. MR: [3320930](#) (cit. on p. [3599](#)).
- Vincent Lemaire and Gilles Pagès (2017). “Multilevel Richardson-Romberg extrapolation”. *Bernoulli* 23.4A, pp. 2643–2692. MR: [3648041](#) (cit. on pp. [3595](#), [3596](#)).
- S. Mishra, Ch. Schwab, and J. Šukys (2012). “Multi-level Monte Carlo finite volume methods for nonlinear systems of conservation laws in multi-dimensions”. *J. Comput. Phys.* 231.8, pp. 3365–3388. MR: [2897628](#) (cit. on p. [3602](#)).
- Florian Müller, Patrick Jenny, and Daniel W. Meyer (2013). “Multilevel Monte Carlo for two phase flow and Buckley-Leverett transport in random heterogeneous porous media”. *J. Comput. Phys.* 250, pp. 685–702. MR: [3079555](#) (cit. on p. [3602](#)).
- Ben Niu, Fred J. Hickernell, Thomas Müller-Gronbach, and Klaus Ritter (2011). “Deterministic multi-level algorithms for infinite-dimensional integration on  $\mathbb{R}^N$ ”. *J. Complexity* 27.3-4, pp. 331–351. MR: [2793867](#) (cit. on p. [3599](#)).
- Michele M Putko, Arthur C Taylor, Perry A Newman, and Lawrence L Green (2002). “Approach for input uncertainty propagation and robust design in CFD using sensitivity derivatives”. *Journal of Fluids Engineering* 124.1, pp. 60–69 (cit. on p. [3589](#)).
- Chang-Han Rhee and Peter W. Glynn (2015). “Unbiased estimation with square root convergence for SDE models”. *Oper. Res.* 63.5, pp. 1026–1043. MR: [3422533](#) (cit. on pp. [3594](#), [3595](#)).
- F. Vidal-Codina, N. C. Nguyen, Michael B. Giles, and J. Peraire (2015). “A model and variance reduction method for computing statistical outputs of stochastic elliptic partial differential equations”. *J. Comput. Phys.* 297, pp. 700–720. MR: [3361685](#) (cit. on p. [3603](#)).

Dongbin Xiu and George Em Karniadakis (2002). “[The Wiener-Askey polynomial chaos for stochastic differential equations](#)”. *SIAM J. Sci. Comput.* 24.2, pp. 619–644. MR: [1951058](#) (cit. on p. [3589](#)).

Received 2017-11-30.

MICHAEL B. GILES: MATHEMATICAL INSTITUTE, UNIVERSITY OF OXFORD

[Mike.Giles@maths.ox.ac.uk](mailto:Mike.Giles@maths.ox.ac.uk)



# NUMERICAL MATHEMATICS OF QUASICRYSTALS

KAI JIANG (蒋凯) AND PINGWEN ZHANG (张平文)

## Abstract

Quasicrystals are one kind of fascinating aperiodic structures, and give a strong impact on material science, solid state chemistry, condensed matter physics and soft matters. The theory of quasicrystals, included in aperiodic order, has grown rapidly in mathematical and physical areas over the past few decades. Many scientific problems have been explored with the efforts of physicists and mathematicians. However, there are still lots of open problems which might to be solved by the close collaboration of physicists, mathematicians and computational mathematicians. In this article, we would like to bridge the physical quasicrystals and mathematical quasicrystals from the perspective of numerical mathematics.

## 1 Introduction

Crystals are one of the most important structures in material science, solid-state physics, condensed matter physics, and soft matters. Before 1980s, by traditional criterion of long-range order, i.e., periodicity, materials can be divided into two categories: crystal and non-crystalline. Crystal materials have periodicity, while the non-crystalline does not. Two ingredients of the description of periodic crystals are lattice and morphology. The periodic lattice is a pure mathematical concept characterized by space group symmetry. In the 19th century, 230 space groups (219 distinct types) in three dimensions, were determined based on periodicity. Then the “classical” crystallography, in which the allowed rotational symmetry is only 1-, 2-, 3-, 4-, or 6-fold symmetry, was perfectly completed. The diversity of morphologies in real world depends on the specific building block of structures in corresponding scales. All of rest materials belong to non-crystalline materials, e.g. glass, coal, coke, and plastic. Non-crystalline materials have short-range order,

---

The work is supported by the Natural Science Foundation of China (Grant No. 21274005, No. 11421101, and No. 11771368).

*MSC2010:* primary 65Z05; secondary 52C23, 11K70, 11K60.

*Keywords:* Quasicrystals, Cut-and-project scheme, Almost periodic functions, Projection method.



lack symmetry, and usually are isotropic. By this classification, all structures in real world seem to be fully known at that time.

In 1982, however, an unexpected pattern with fivefold symmetric diffraction, which is not compatible with three-dimensional periodicity, was found by Shechtman in rapidly quenched aluminum manganese alloy [Shechtman, Blech, Gratias, and Cahn \[1984\]](#). This discovery raised much interest. Until now, many stable quasicrystals have been found in more than a hundred of different metal alloys [N. Wang, Chen, and Kuo \[1987\]](#), [Tsai \[2008\]](#), and [Steurer \[2004\]](#). Moreover, quasicrystals have been also discovered in a host of soft-matter materials [Zeng, Ungar, Liu, Percec, Dulcey, and Hobbs \[2004\]](#), even in nature [Bindi, Steinhardt, Yao, and Lu \[2009\]](#). These discoveries lead people to realize that there are many different kinds of structures between periodic crystals and non-crystalline phases. This is more important than quasicrystals appear at first sight, because non-periodic order shows both new features and new horizon line. As a consequence, the crystal has been redefined as that if it has a sharp diffraction pattern in reciprocal space “[Report of the executive committee for 1991](#)” [\[1992\]](#). In fact, beside quasicrystals, incommensurate modulated phases, incommensurate composites and incommensurate magnetic structures have been confirmed as members in the family of aperiodic structures. It is fair to say that a classification of a hierarchy of aperiodic order, and structures between periodic crystals and amorphous phases, has not been achieved yet.

With the efforts of physicists and mathematicians, many scientific problems of quasicrystals have been solved in the past few decades. However, there are still many open problems in the field of quasicrystals, many of which lie in both fields of physics and mathematics. This inspires us to bridge the connection between physical quasicrystals and mathematical quasicrystals. In this paper, we would like to briefly review the related studies on both fields and understand their connection from the viewpoint of numerical mathematics. The paper is organized as follows. In [Section 2](#), we present a short introduction of physical quasicrystals, and related mathematical knowledge of quasicrystals. In [Section 3](#), we review common computational methods of quasicrystal, especially the projection method. Finally in [Section 4](#), the relationship among mathematical, physical, and numerical quasicrystals is discussed. Also some perspectives and unsolved problems are drawn.

## 2 Quasicrystals: From Physics to Mathematics

**2.1 What is a physical quasicrystal ?** Since the symmetry is one of the most important properties in experimental discovery, a prevailing definition of quasicrystals is that a  $d$ -dimensional quasicrystal is not periodic, and has any finite subgroup of  $O(d)$  as its point

group. However, this definition is too restrictive to excludes any important and interesting collections of quasicrystals. For example, there are no quasicrystals in one dimension according to the above definition. To contain these aperiodic structures who exhibit all the well-known properties of quasicrystals, a quasicrystal is defined as quasiperiodic crystal which was firstly proposed by Levine and Steinhardt [Levine and Steinhardt \[1986\]](#), [Mermin \[1991\]](#), and [Lifshitz \[2003\]](#). A  $d$ -dimensional quasiperiodic crystal means that its density function can be expanded as

$$(1) \quad f(\mathbf{x}) = \sum_{\mathbf{k} \in \Lambda} \hat{f}(\mathbf{k}) e^{i\mathbf{k} \cdot \mathbf{x}}, \quad \mathbf{x} \in \mathbb{R}^d,$$

where the spectrum set  $\Lambda = \{\mathbf{k} = \sum_{j=1}^n h_j \mathbf{b}_j, \mathbf{b}_j \in \mathbb{R}^d, h_j \in \mathbb{Z}\}$  is a finitely generated  $\mathbb{Z}$ -module of rank  $n$ . If  $n = d$ , it is a periodic crystal.

Many mathematical works on quasicrystals have been independently explored before the discovery of quasicrystals. In recent years, it has been found that quasicrystals have a fundamental connection to many areas of mathematics, e.g. algebra [de Bruijn \[1981a,b\]](#), discrete geometry [Senechal \[1995\]](#), number theory and harmonic analysis [Salem \[1963\]](#), [Meyer \[1972\]](#), and [Lev and Olevskii \[2015\]](#), crystallography [Mermin \[1991\]](#) and [Steurer and Deloudi \[2009\]](#), diffraction theory [Baake and Grimm \[2011, 2012, 2013\]](#), dynamical systems [Hou and You \[2012\]](#), sampling theory [Meyer \[2012\]](#). In particular, it might be a way to treat the well-known Riemann Hypothesis [Dyson \[2009\]](#).

To the best of our knowledge, there are two significant pioneer works in mathematics. One is the Penrose's work from discrete geometry [Penrose \[1974\]](#). Another is Meyer's monograph on the connection between algebraic number theory and harmonic analysis [Meyer \[1972\]](#). In fact, it contains the abstract theory of the cut-and-projection method, in terms of the full generality of locally compact Abelian groups. Meyer's work is also an extension of the theory of almost periodic functions to the setting of point sets. In the following context, we briefly review the mentioned mathematical works.

**2.2 Penrose tiling.** The best known work might be the fivefold symmetric planar tiling due to Penrose from Oxford [Penrose \[1974\]](#), now referred to as Penrose tiling. This work stems from the Hilbert's 18<sup>th</sup> problem which raises issues of the filling of space with congruent shapes. Hilbert was probably assuming – incorrectly, as it turned out – that no anisohedral prototile could exist in two dimensions. In other words, there do not exist finite prototiles that cover entire space without void. Based on the previous work of many mathematicians, including H. Wang, R. Berger, and D. Knuth, in 1974, Penrose found an approach to pave the whole plane with only two rhombi in an aperiodic way [Senechal \[1995\]](#).

In particular, the simplest rhombic Penrose tiling includes two prototiles: one thick rhombus (with angles  $2\pi/5$  and  $3\pi/5$ ) and one thin rhombus (with angles  $4\pi/5$  and

$\pi/5$ ). Using the above rhombi, Penrose found a matching rule to pave the plane space. The resulted pattern, also known as Penrose tiling, is a quasicrystal with 5-fold symmetry without any period. It was later shown that this pattern is compatible with a projection from 5-space and has a pure Bragg spectrum [de Bruijn \[1981a,b\]](#). More tilings with other symmetries, as well as three dimensional tilings have also been constructed and studied intensively [Senechal \[1995\]](#), [Baake and Grimm \[2013\]](#), and [Grünbaum and Shephard \[1987\]](#).

**2.3 Almost periodic functions.** Harmonic analysis was originally devoted to periodic functions. The analysis of a periodic function perfectly depends on the knowledge of one period. A natural aperiodic generalization of continuous periodic functions is almost periodic functions, originally from Bohr's work in 1920s [Bohr \[1925\]](#). Its definition is given as follows.

**Definition 1** ( $\varepsilon$ -almost period). *Let  $f$  be a complex-valued function on  $\mathbb{R}$  and let  $\varepsilon > 0$ . An  $\varepsilon$ -almost period of  $f$  is a number  $\tau$  such that*

$$\sup_{x \in \mathbb{R}} |f(x - \tau) - f(x)| < \varepsilon.$$

**Definition 2** (Almost periodic function). *A complex-valued function  $f$  on  $\mathbb{R}$  is almost periodic, if it is continuous and if for every  $\varepsilon > 0$ , there exists  $L = L(\varepsilon, f) > 0$  such that every interval of length  $L$  on  $\mathbb{R}$  contains an  $\varepsilon$ -almost period of  $f$ .*

Obviously, continuous periodic functions are almost periodic. Beside that, another simplest almost periodic function is  $f(x) = \sin(x) + \sin(\sqrt{2}x)$ . However,  $f$  is not periodic since  $f(x) = 0$  only for  $x = 0$ . Denote  $AP(\mathbb{R})$  to be the space of almost periodic functions on  $\mathbb{R}$ . It is known that  $AP(\mathbb{R})$  is a closed subalgebra of  $L^\infty(\mathbb{R})$  and that almost periodic functions are uniformly continuous [Katznelson \[2004\]](#). Given  $f \in AP(\mathbb{R})$ , the mean value  $M(f)$  of  $f$  is defined by

$$(2) \quad M(f) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f(x) dx.$$

The limit of the above definition exists when  $f \in AP(\mathbb{R})$ . Then, we define the inner product  $\langle \cdot, \cdot \rangle_{\mathcal{L}^2}$  on  $AP(\mathbb{R})$  as

$$(3) \quad \langle f, g \rangle_{\mathcal{L}^2} = M(f \bar{g}) = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f(x) \bar{g}(x) dx,$$

for  $f, g \in AP(\mathbb{R})$ . This inner product is well defined since  $f\bar{g}$  is also in  $AP(\mathbb{R})$ . Moreover, it induces the  $\mathfrak{L}^2$ -norm, for  $f \in AP(\mathbb{R})$ ,

$$(4) \quad \|f\|_{\mathfrak{L}^2} = \lim_{T \rightarrow \infty} \left( \frac{1}{2T} \int_{-T}^T |f(x)|^2 dx \right)^{1/2}.$$

We now define the Fourier coefficient of  $f \in AP(\mathbb{R})$  by

$$(5) \quad \hat{f}(\lambda) = \langle f, e^{i\lambda x} \rangle_{\mathfrak{L}^2} = M(fe^{-i\lambda x}).$$

By Bessel's inequality, we have  $\sum_{\lambda \in \mathbb{R}} |\hat{f}(\lambda)|^2 \leq \|f\|_{\mathfrak{L}^2}^2 < \infty$ . This implies that  $\hat{f}(\lambda) = 0$  except for countable many values of  $\lambda$ 's. We define its frequency set  $\sigma(f) = \{\lambda \in \mathbb{R} : \hat{f}(\lambda) \neq 0\}$  and write

$$(6) \quad f(x) \approx \sum_{\lambda \in \sigma(f)} \hat{f}(\lambda) e^{i\lambda x},$$

where the right-hand side is referred to as Fourier series associated to  $f$ . It is known that the complex exponentials  $\{e^{i\lambda x}\}_{\lambda \in \mathbb{R}}$  form an orthonormal basis of  $AP(\mathbb{R})$ . Moreover, the Parseval's identity

$$(7) \quad \|f\|_{\mathfrak{L}^2}^2 = \sum_{\lambda \in \mathbb{R}} |\hat{f}(\lambda)|^2,$$

holds for  $f \in AP(\mathbb{R})$ . The following theorem gives the convergence of the Fourier series to an almost periodic function.

**Theorem 1** (Theorem 1.20 in [Corduneanu \[1968\]](#)). *Let  $f \in AP(\mathbb{R})$ . If the Fourier series in (6) associated to  $f$  converges uniformly, then it converges to  $f$ .*

**2.4 Meyer's work.** Let  $\Lambda$  be a set of real numbers, we say  $\Lambda$  is a coherent set of frequencies if there exist a  $C > 0$  and a compact set  $\mathfrak{K}$  such that

$$(8) \quad \sup_{x \in \mathbb{R}} |P(x)| \leq C \sup_{x \in \mathfrak{K}} |P(x)|,$$

for all trigonometric sums  $P(x)$  whose frequencies belong to  $\Lambda$ . In 1970s, Meyer considered a primal issue: how to construct coherent set  $\Lambda$  and what  $\Lambda$ 's properties are. A trivial example of coherent set is  $\Lambda = \mathbb{Z}$ . As a consequence,  $P(x)$  is the periodic function. More general coherent sets can be characterized by the remarkable Diophantine approximation property. The quasicrystal is an interesting by-product during the period of studying this primal issue.

To characterize the Diophantine approximation property, a new concept of harmonious set is introduced into the subject of harmonic analysis on locally compact Abelian groups. In this paper, we focus on the space of  $\mathbb{R}^n$  which is a specific example of local compact Abelian groups. More general results can be found in Meyer's book [Meyer \[1972\]](#).

**Definition 3** (Harmonious set). *A set  $\Lambda = \{\lambda_j\}_{j \geq 1}$  is harmonious if for each positive  $\varepsilon$ , there exists a function  $T(\varepsilon, \lambda_1, \dots) = T > 0$ , such that each interval of real numbers of length  $T$  at least contains a solution  $t$  satisfying the infinite system of Diophantine inequalities*

$$(9) \quad |t\lambda_j - [t\lambda_j]| \leq \varepsilon, \quad j \geq 1.$$

$[\cdot]$  is the nearest integer of  $\cdot$ .

Another useful concept is the Delaunay set we will use later.

**Definition 4** (Delaunay set). *A subset  $\Lambda \subset \mathbb{R}^n$  is a Delaunay set if there exist two radii  $R_2 > R_1 > 0$  such that each ball with radius  $R_1$ , whatever be its location, shall contain at most one point in  $\Lambda$  while each ball with radius  $R_2$ , whatever be its location, shall contain at least one point in  $\Lambda$ .*

The first requirement can be equivalently given by the *uniformly discrete* formulation: there exists a positive  $r$  such that  $\lambda, \lambda' \in \Lambda$  and  $\lambda \neq \lambda'$  imply  $|\lambda - \lambda'| \geq r$ . A collection of points fulfilling the second condition is referred to as *relatively dense*.

An interesting problem is to characterize a harmonious set by its arithmetical properties. Considering additive properties, if  $\Lambda \subset \mathbb{R}^d$  is a Delaunay set and  $F$  is a finite subset of  $\mathbb{R}^d$ , then  $\Lambda$  is harmonious if and only if  $\Lambda - \Lambda \subset \Lambda + F$ , where  $\Lambda - \Lambda$  denotes the set of Minkowski difference  $\lambda_1 - \lambda_2$ , for arbitrary  $\lambda_1, \lambda_2 \in \Lambda$ . This property naturally gives the first mathematical definition of quasicrystals.

**Definition 5.** *A mathematical quasicrystal  $\Lambda$  is a Delaunay set in  $\mathbb{R}^n$  such that  $\Lambda - \Lambda \subset \Lambda + F$  where  $F$  is a finite set.*

In many literatures, a point set satisfying the [Definition 5](#) is referred to as Meyer set. An equivalent concept is the quasiregular set that is a Delaunay set  $\Lambda$  such that  $\Lambda - \Lambda$  is also a Delaunay set [Lagarias \[1996\]](#). However, [Definition 5](#) does not directly yield the Diophantine approximation characterization of quasicrystals.

In 1970s, to maintain the Diophantine approximation property, Meyer proposed an elegant method to construct relatively dense harmonious sets, or in terms of model sets [Meyer \[1972\]](#). A model set  $\Lambda \subset \mathbb{R}^n$  is defined as follows. Consider a lattice  $\mathfrak{D} \subset \mathbb{R}^{n+m} = \mathbb{R}^n \times \mathbb{R}^m$  where  $m$  is an integer and  $m = 0$  is not excluded. If  $(x, y) = X \in \mathbb{R}^n \times \mathbb{R}^m$ , we write  $x = p_1(X)$  and  $y = p_2(X)$ . Assume that  $p_1 : \mathfrak{D} \mapsto p_1(\mathfrak{D})$  is a 1-1 mapping and that  $p_2(\mathfrak{D})$  is a dense subgroup of  $\mathbb{R}^m$ .

**Definition 6** (Model set). *Keeping the above notations, let  $\mathcal{B}$  be any bounded set in  $\mathbb{R}^m$ . Then the model set  $\Lambda$  defined by  $\mathfrak{D}$  and  $\mathcal{B}$  is the collection of all  $\Lambda = p_1(d)$  such that  $d \in \mathfrak{D}$  and  $p_2(d) \in \mathcal{B}$ .*

This is the well-known cut-and-project scheme in the field of quasicrystals. A famous guiding example is the Fibonacci chain which can be obtained as a projection of square lattice  $\mathbb{Z}^2$  within a strip. The chain space is the line with slope  $(\sqrt{5} - 1)/2$ . Cut-and-project method appears in various different disguises in the literature, each of which has its own merits. In 1981, [de Bruijn \[1981a,b\]](#) devised an algebraic description of the rhombic Penrose tilings, based on the dualisation of a pentagrid. It is straightforward to generalize this method to produce planar rhombic tilings with arbitrary rotational symmetry. In fact, Bruijn's method is equivalent to the cut-and-project approach with a suitable choice of the cut window [Gähler and Rhyner \[1986\]](#).

The relationship between model sets and quasicrystals is given as follows.

**Theorem 2** (Theorem 1 in [Meyer \[1995\]](#)). *Let  $\Lambda \subset \mathbb{R}^n$  be a model set such that the corresponding bounded set  $\mathcal{B}$  has a non-empty interior. Then  $\Lambda$  is a mathematical quasicrystal.*

*Conversely if  $\Lambda$  is a mathematical quasicrystal, there exists a finite set  $F$  and a model set  $\Lambda_0$  such that  $\Lambda \subset \Lambda_0 + F$ . Moreover this model set  $\Lambda_0$  corresponds to a bounded set  $\mathcal{B}_0$  with a non-empty interior.*

From [Theorem 2](#), the model set is a subset of mathematical quasicrystals. Nevertheless, cut-and-project method has three benefits in the study of quasicrystals. The first one is that it provides an explicit constructive approach to generate quasicrystals.

The second one is on explaining why only finitely many spectral points are observed in the experimental diffraction pattern of a quasicrystal. Using the cut-and-project scheme, we can obtain Poisson's summation formula for quasicrystals.

**Theorem 3.** (Poisson's summation formula for quasicrystal) *Let  $\Lambda \in \mathbb{R}^n$  be a model set such that the corresponding bounded set  $\mathcal{B}$  has a non-empty interior. And let  $\mu = \sum_{\lambda \in \Lambda} \delta_\lambda$  be the Dirac masses over  $\Lambda$ . Then the Fourier transform, in the distribution sense,  $\hat{\mu}$  of  $\mu$  is given by*

$$(10) \quad \hat{\mu} = \sum_{d \in \mathfrak{D}^*} \omega(p_2(d^*)) \delta_{p_1(d^*)},$$

where the weights  $\omega(p_2(d^*))$  are defined by  $\omega(y) = \frac{(2\pi)^n}{\text{vol}(\mathfrak{D})} \chi_{\mathcal{B}}(-y)$ ,  $y = p_2(d^*)$ ,  $\chi_{\mathcal{B}}$  is the indicator function of  $\mathcal{B}$ , and  $\text{vol}(\mathfrak{D})$  is the volume of fundamental domain of  $\mathfrak{D}$ .

The support of the measure  $\hat{\mu}$  is dense in  $\mathbb{R}^n$ , and the weight  $\omega(y)$  has a rapid decay at infinity. A general result of the above theorem is the Theorem 7 in [Meyer \[ibid.\]](#).

The third one is the connection between quasicrystals and almost periodic functions from the duality theory. For a set  $\Lambda \in \mathbb{R}^n$  and  $\varepsilon > 0$ , its  $\varepsilon$ -dual set  $\Lambda_\varepsilon^*$ , is defined by,  $\Lambda_\varepsilon^* = \{y : |e^{iy \cdot \lambda} - 1| \leq \varepsilon, \lambda \in \Lambda\}$ . It is a generalization of periodic lattice and its dual lattice.

**Definition 7.** *A mathematical quasicrystal  $\Lambda$  is a Delaunay set such that the  $\varepsilon$ -dual set  $\Lambda_\varepsilon^*$  is also a Delaunay sets whenever  $0 < \varepsilon \leq 1$ .*

Therefore, for any almost periodic function  $f$  whose spectrum lies in quasicrystal  $\Lambda$  and  $C > 0$ , we have

$$(11) \quad \sup_x |f(x - \tau) - f(x)| \leq C \varepsilon \sup_x |f(x)|,$$

where  $C$  and  $\varepsilon$ -almost periods  $\tau \in \Lambda_\varepsilon^*$  do not depend on  $f$ . [Definition 7](#) is equivalent to [Definition 5](#), which allows us build a connection between quasicrystals and almost periodic functions. More concretely, if  $\Lambda$  is a model set, then its  $\varepsilon$ -dual set  $\Lambda_\varepsilon^*$ ,  $0 < \varepsilon \leq 1$ , is also a model set (see Theorem 3 in [Meyer \[1995\]](#)).

Besides, self-similarity is an important property of quasicrystals. From this perspective, the Pisot (or Pisot-Vijayaraghavan) and Salem numbers play a key role in characterizing the property of quasicrystals.

**Definition 8** (Pisot number and Salem number). *A Pisot number  $\theta > 1$  is a real algebraic integer of degree  $n \geq 1$  if all its conjugates  $\theta_2, \dots, \theta_n$  satisfy  $|\theta_2| < 1, \dots, |\theta_n| < 1$ .*

*A Salem number  $\theta > 1$  is a real algebraic integer of degree  $n \geq 1$  if all its conjugates  $\theta_2, \dots, \theta_n$  satisfy  $|\theta_2| \leq 1, \dots, |\theta_n| \leq 1$  with, at least, equality somewhere.*

**Theorem 4** (Theorem 6 in [Meyer \[ibid.\]](#)). *If  $\Lambda$  is a quasicrystal,  $\theta > 1$  is a real number and  $\theta\Lambda \subset \Lambda$ , then  $\theta$  is either a Pisot number or is a Salem number.*

*Conversely, for each Pisot or Salem number  $\theta$ , there exists a quasicrystal  $\Lambda \subset \mathbb{R}^n$  such that  $\theta\Lambda \subset \Lambda$ .*

A physical quasicrystal is usually related to a Pisot number. For example, 5-fold symmetric quasicrystals refer to the golden number of  $\tau = (1 + \sqrt{5})/2$  which is a Pisot number.

### 3 Numerical Quasicrystals

In this section, we will review numerical mathematics of quasicrystals. In numerical computation, the implementation of algorithms is based on physical models. A class of useful physical models to describe the phase behaviour of quasicrystals is the phase-field

quasicrystal model. In particular, its free energy functional can be written as

$$(12) \quad F[\varphi(\mathbf{x})] = \lim_{R \rightarrow \infty} \frac{1}{B(0, R)} \int_{B(0, R)} \int_{\mathbb{R}^d} \frac{\gamma}{2} [\varphi(\mathbf{x}) G(\mathbf{x}, \mathbf{x}') \varphi(\mathbf{x}')] d\mathbf{x} d\mathbf{x}' \\ + \lim_{R \rightarrow \infty} \frac{1}{B(0, R)} \int_{B(0, R)} \left[ -\frac{\varepsilon}{2} \varphi^2(\mathbf{x}) - \frac{\alpha}{3} \varphi^3(\mathbf{x}) + \frac{1}{4} \varphi^4(\mathbf{x}) \right] d\mathbf{x},$$

where  $\varphi(\mathbf{x})$  is the order parameter to describe the order of structures.  $\gamma$ ,  $\varepsilon$  and  $\alpha$  are phenomenological parameters of the system. In this model, the polynomial term corresponds to the bulk free energy of the system, whereas the term involving  $G(\mathbf{x}, \mathbf{x}')$  is a two-body correlation potential, describing the free energy cost of inhomogeneity of the system. Different choices of  $G(\mathbf{x}, \mathbf{x}')$  result in the selection of different fundamental modes at particular length scales, thus promoting the formation of ordered structures. When the function  $G(\mathbf{x}, \mathbf{x}')$  is chosen such that two length scales with proper length ratios are selected, complicated ordered phases including quasicrystals can be stabilized Müller [1994], Lifshitz and Petrich [1997], Dotera, Oshiro, and Zihlerl [2014], and Jiang, Tong, and P. Zhang [2016]. The selection of two length scales in the potential function such that it has two equal-depth minima, 1 and  $q$ , which can be realized by differential term Lifshitz and Petrich [1997], a steplike function Barkan, Diamant, and Lifshitz [2011], or a Gaussian-type potential family Archer, Rucklidge, and Knobloch [2013] and Barkan, Engel, and Lifshitz [2014].

Theoretically, the ordered patterns, including periodic and quasiperiodic, are corresponding to local minima of the free energy functional. Seeking for the minima of (12) can directly use optimization methods, or gradient flow equations, such as Allen-Cahn, Cahn-Hilliard equations. However, there are multiple local minima of the energy functional 12 due to its the nonlinearity and nonconvexity. Thus the initial values are important to speed up the convergence, and critical to find the expected solutions. There is not an universal approach to choose initial values for a general nonlinear variation problem. In practice, the choice of initial values depends on specific problem. The group theory is a useful tool of screening initial values for symmetric structures: for periodic crystals with translational and rotational invariance, the space group is a perfect tool to chose initial values Jiang, Huang, and P. Zhang [2010], Xu, Jiang, P. Zhang, and Shi [2013], and Jiang, C. Wang, Huang, and P. Zhang [2013]; for quasicrystals, the point group can be used to screen initial values Jiang, Tong, P. Zhang, and Shi [2015] and Jiang, P. Zhang, and Shi [2017]. In general, the more complicated the phase is, the more sensitive to initial value the solution is.

Another important issue in numerical computations is how to discretize or represent ordered parameter  $\varphi(\mathbf{x})$ . For periodic structures, their study can be confined to one period. As a consequence, many traditional discretization methods can be applied to solve such



structures, e.g. Fourier spectral method, finite difference/element/volume method. However, these numerical methods are not directly applicable to quasicrystals due to their property of space-filling. In the following context, we review existing discretization schemes to decompose order parameter, especially the projection method.

**3.1 One-mode or multi-modes approximation method.** One-mode or multi-modes approximation method is a widely used semi-analytical approach in physics to find the approximation solution under some assumption [Chaikin and Lubensky \[2000\]](#). If the solution is symmetric, the approximation approach uses symmetric eigenfunctions to expand order parameter function  $\varphi(\mathbf{x})$ . Then the energy functional becomes a function. The optimization problem of minimizing a energy functional becomes minimizing a function with one or few variables.

In the model of (12), when the parameter  $\gamma \rightarrow +\infty$ , the interaction potential  $G(\mathbf{x}, \mathbf{x}')$  must be zero. Otherwise energy value goes to infinity. In this case, only fundamental Fourier modes that lie on rings of  $|\mathbf{k}| = 1$  or/and  $|\mathbf{k}| = q$  are nonzero. For example, in consideration of dodecagonal symmetric quasicrystals, order parameter  $\varphi(\mathbf{x})$  is represented by the linear combination of trigonometric functions satisfying two properties: the basis functions have 12-fold symmetry; and the frequencies lie on the two rings. Then the original energy functional (12) becomes an approximation function of only two variables [Lifshitz and Petrich \[1997\]](#) and [Jiang, Tong, P. Zhang, and Shi \[2015\]](#).

The approximation method can dramatically decrease the computational burden, and useful for qualitative analysis. However, it is only available to some limited cases.

**3.2 Crystalline approximant method.** Recently, a popular method to calculate a quasicrystal is a modified trigonometric spectral method which uses a large region with periodic boundary condition to approximate the quasicrystal. More precisely, this trigonometric spectral method can only be applied to periodic structures. Therefore, this method computes crystalline approximants rather than quasicrystals, for which we refer to it as crystalline approximant method [Jiang and P. Zhang \[2014\]](#).

For any  $d$ -dimensional periodic structure  $f(\mathbf{x})$ ,  $\mathbf{x} \in \mathbb{R}^d$ , the repeated structural unit is called a unit cell. A primitive unit cell, described by  $d$   $d$ -dimensional primitive vectors,  $\mathbf{a}_1, \dots, \mathbf{a}_d$ , has the smallest possible volume.  $f(\mathbf{x})$  has translational invariance on the lattice composed by the primitive vectors. Given the primitive vectors, the primitive reciprocal vectors in Fourier space,  $\mathbf{b}_1, \dots, \mathbf{b}_d$ ,  $\mathbf{b}_j \in \mathbb{R}^d$  satisfy  $\mathbf{a}_i \mathbf{b}_j = 2\pi \delta_{ij}$ ,  $1 \leq i, j \leq d$ . The primitive reciprocal vector, i.e., the spectra of  $f(\mathbf{x})$ , is specified by  $\Lambda = \{\mathbf{k} = \sum_{i=1}^n h_i \mathbf{b}_i, h_i \in \mathbb{Z}\}$ . One of the most important properties of the reciprocal primitive lattices is that trigonometric functions,  $\{e^{i(\mathbf{h}\mathbf{B})\cdot\mathbf{x}}\}_{\mathbf{h} \in \mathbb{Z}^d}$ ,  $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_d)$ , form a set of basis functions in  $L^2(\mathcal{V})$ ,  $\mathcal{V}$  is the primitive unit cell described by  $d$  vectors

$a_1, \dots, a_d$ . The periodic function  $f(\mathbf{x})$  on  $L^2(\mathcal{V})$  can be expanded as

$$(13) \quad f(\mathbf{x}) = \sum_{\mathbf{h} \in \mathbb{Z}^d} \hat{f}(\mathbf{h}) e^{i(\mathbf{h}\mathbf{B}) \cdot \mathbf{x}},$$

It is noted that, the spectra of a periodic structure is a linear combination of primitive reciprocal vectors on field  $\mathbb{Q}$  (actually on ring  $\mathbb{Z}$ ).

For a  $d$ -dimensional quasicrystal  $f(\mathbf{x})$ , its spectra  $\Lambda$  is a combination of primitive reciprocal vectors on field  $\mathbb{R}$  rather than  $\mathbb{Q}$ , i.e.,  $\Lambda_{QC} = \{\mathbf{k} = \sum_{j=1}^d p_j \mathbf{b}_j, p_j \in \mathbb{R}\}$ . The trigonometric spectral method (13) cannot directly apply to compute quasicrystals. A natural idea is to use a lattice of  $\mathbb{R}^d$  to approximate  $\Lambda_{QC}$  which is the Diophantine approximation. More concretely, the quasiperiodic function can be approximated by

$$(14) \quad f(\mathbf{x}) \approx \sum_{\mathbf{k} \in \Lambda_{QC}} \hat{f}([L\mathbf{k}]) e^{i \cdot [L\mathbf{k}] \cdot \mathbf{x} / L},$$

where  $\mathbf{x}$  belongs to the region  $L\mathcal{V}$ . To approximate the quasiperiodic function  $f(\mathbf{x})$  accurately, we need that  $[L\mathbf{k}]$  approximate  $L\mathbf{k}$  for all  $\mathbf{k} \in \Lambda_{QC}$  as close as possible. Without loss of generality, one can always use  $(1, 0, \dots, 0)$  as one of primitive reciprocal vectors. Therefore,  $L$  can be chosen as an integer. In view of numerical computability, the integer  $L$  should be as small as possible. However, since there exist irrational coefficients  $p_j$  in  $\Lambda_{QC}$ , from the Diophantine approximation theory,  $L$  increases nonlinearly and quickly as the desired approximation error becomes small. A concrete example is in the computation of a two-dimensional decagonal symmetric quasicrystal in the phase field quasicrystal model. Table 1 gives the least integer  $L$  when achieving desired Diophantine approximation error. It is easy to see that the computational amount increases quickly due to the

Table 1: For a two-dimensional decagonal symmetric quasicrystal, the Diophantine approximation error and the required least integer  $L$ . The computational region is  $[0, 2\pi L]^2$ . The first  $L = 126$  is the least integer in computing the crystalline approximant of two-dimensional decagonal symmetric quasicrystals based on the model of (12).

Error	0.1669	0.0918	0.0374	0.0299	...
$L$	126	204	3372	53654	...

increase of  $L$ .

The error of the crystalline approximant method comes from two parts: using a finite region to approximate the space-filling quasicrystal, and using a periodic function to approximate the quasicrystal in a finite domain. Numerical results have given an evidence

that the gap between the free energy of the quasicrystals and their corresponding approximants always exists [Jiang and P. Zhang \[2014\]](#). However, it still lacks a rigorous analysis about this phenomenon.

**3.3 Gaussian approximation method.** Gaussian approximation method is based on the assumption that the density function is a linear combination of Gaussian functions centered on a priori determined lattice or quasilattice  $\Lambda^*$  [Sachdev and Nelson \[1985\]](#). Under this assumption, the density function is represented as

$$(15) \quad \varphi_G(\mathbf{x}) = \sum_{\mathbf{x}_n \in \Lambda^*} G_\sigma(\mathbf{x} - \mathbf{x}_n),$$

where  $G_\sigma(\mathbf{x}) = (\pi\sigma^2)^{-3/2} \exp(-|\mathbf{x}|^2/\sigma^2)$  is the Gaussian function with width  $\sigma$ , and  $\Lambda^*$  is the given lattice or quasilattice. Note that if  $\sigma \rightarrow \infty$ , Gaussian approximation method will obliterate structured phases, which means an uniform state. Using the expression (15), the free energy functional  $F[\varphi(\mathbf{x}); \Lambda^*]$  becomes a function  $F(\sigma; \Lambda^*)$  of  $\sigma$ . It is much easier to minimize  $F(\sigma; \Lambda^*)$  than  $F[\varphi; \Lambda^*]$ . Furthermore, in many density functional frameworks, the Gaussian functions in (15) are assumed to be non-overlap. This provides a numerical approximation advantage to evaluate the integral terms in  $F$ , i.e., for any continuous function  $f$ , we have

$$(16) \quad \int_V f(\phi_G) d\mathbf{x} \approx \sum_{\mathbf{x}_n \in \Lambda^*} \int_V f(G_\sigma(\mathbf{x} - \mathbf{x}_n)) d\mathbf{x}.$$

When we describe the lattice or quasilattice by a measure

$$\gamma(\mathbf{x}) = \sum_{\mathbf{x}_n \in \Lambda^*} \delta(\mathbf{x} - \mathbf{x}_n),$$

the [Equation \(15\)](#) can be re-written in the Fourier space as (actually by the Poisson's summation formula of quasicrystals of [Theorem 3](#))

$$(17) \quad \hat{\varphi}_G(\mathbf{k}_n) = \widehat{(\gamma * G_\sigma)}(\mathbf{k}_n) = w(\mathbf{k}_n) \exp\left(-\frac{|\mathbf{k}_n|^2 \varepsilon^2}{4}\right), \quad \mathbf{k}_n \in \Lambda,$$

where  $w(\mathbf{k}_n)$ , the weight factor, is the spectral coefficients of the given structure. For a periodic lattice structure  $\Lambda^*$ , the weight factor  $w(\mathbf{k}_n) \equiv 1$  for  $\mathbf{k}_n \in \Lambda$ . Nevertheless, the weight factor of a quasicrystal structure is much complicated. The main method to obtain the weight factor of quasicrystals is employing the cut-and-project method, see [Theorem 3](#). In general, we should carefully choose the cut windows as different choices of windows result in different quasicrystals [Smith \[1990\]](#) and [MCarley and Ashcroft \[1994\]](#).

The success of Gaussian approximation method in density functional framework should attribute to the nature of physical problems. From our experience, this method is suitable for solid quasicrystals and the width  $\sigma$  should not be too large to overlap. It provides a simple route to approximate order parameter  $\varphi(\mathbf{x})$ . The performance might be further improved by extending the one parameter Gaussian in (15) to a summation of several general Gaussians. However, mathematically, it is hard to expect high accuracy of the Gaussian approximation method.

**3.4 Projection Method.** The projection method is inspired by a picture of diffraction pattern of quasicrystals. We observed that the diffraction point of a  $d$ -dimensional quasicrystal cannot be represented by a linear combination of  $d$   $d$ -dimensional vectors with integer coefficients like periodic lattice. Then we used  $n$  ( $n > d$ ) vectors to represent the diffraction pattern. From another viewpoint, the  $n$  vectors can be lifted up to  $n$  dimension to expand  $\mathbb{Z}^n$ . As a consequence, we project the  $\mathbb{Z}^n$  to  $\mathbb{R}^d$  space to obtain the diffraction pattern. In particular, consider the information of primitive lattice, we can represent the diffraction pattern of a  $d$ -dimensional quasicrystal as

$$(18) \quad \Lambda_{QC} = \{\mathbf{k} = \mathbf{P}\mathbf{B}\mathbf{h}, \mathbf{h} \in \mathbb{Z}^n, \mathbf{P} \in \mathbb{R}^{d \times n}, \mathbf{B} \in \mathbb{R}^{n \times n}\},$$

where  $\mathbf{P}$  is the projection matrix of rank  $d$ , and  $\mathbf{B} = (\mathbf{b}_1, \dots, \mathbf{b}_n)$  is the  $n$ -dimensional reciprocal primitive lattice. The projection matrix depends on the specific structure. Consider the rotational symmetry, the crystallographic restriction in two and three dimensions can be generalized to arbitrary dimensions [Steurer and Deloudi \[2009\]](#). For example, 5-, 8-, 10-, and 12-fold symmetric quasicrystals, the minimal dimension of embedded space is four. While 7-, 9-, and 18-fold symmetric quasicrystals must be restricted to six-dimensional space or above. A uniform computational formula for the minimal extended dimension of rotational symmetry can be obtained from an additive Euler totient function [Hiller \[1985\]](#). If we consider periodic lattice, the projection matrix is a  $d$ -order identity matrix.

Using the representation of  $\Lambda_{QC}$ , we proposed an expansion of a  $d$ -dimensional quasicrystal

$$(19) \quad f(\mathbf{x}) \approx \sum_{\mathbf{k} \in \Lambda_{QC}} \hat{f}(\mathbf{k}) e^{i\mathbf{k} \cdot \mathbf{x}}.$$

The next task is to compute coefficients  $\hat{f}(\mathbf{k})$ . In order to ensure the convergence of the above Fourier series, the coefficients should have decay property. In the model set, each spectrum point has a rapidly decay weight factor due to the cut window (see [Theorem 3](#)). Accordingly, in projection method, we assume that coefficients  $\{\hat{f}(\mathbf{k})\}_{\mathbf{k} \in \Lambda_{QC}} \in \ell^2(\mathbb{Z}^n)$ . In practice computation, we calculate it by the  $n$ -dimensional  $L^2$ -inner product,  $\hat{\phi}(\mathbf{h}) =$

$\langle \tilde{\phi}(\tilde{\mathbf{x}}), e^{-i(\mathbf{B}\mathbf{h})^T \tilde{\mathbf{x}}} \rangle$ , with  $\tilde{\mathbf{x}}$  being in the supercube of  $\{\sum_{i=1}^n s_i \mathbf{a}_i \in \mathbb{R}^n, 0 \leq s_i \leq 1\}$ . Here  $\mathbf{a}_i, i = 1, \dots, n$ , are the primitive lattice vectors forming the primitive lattice  $A$  of the  $n$ -dimensional periodic structure. Specific value of  $\hat{f}(\mathbf{k})$  is obtained by solving the physical models, such as the energy functional of (12).

With the help of  $n$ -dimensional reciprocal space, projection method can calculate the spectrum of quasicrystals directly. In the projection method, the physical space variable  $\mathbf{x}$  always belongs to  $d$ -dimensional space. Therefore an energy functional including  $d$ -dimensional quasicrystals is not required to be lifted up to  $n$  dimension. Since quasicrystals are space-filling structures, the energy values must be infinity. Instead, the energy density is considered. With the  $\mathcal{L}^2$ -inner product (3), the energy density of a quasicrystal can be calculated by the following lemma.

**Lemma 1** (Lemma in Jiang and P. Zhang [2014]). *For a  $d$ -dimensional quasiperiodic function  $f(\mathbf{x})$ , under the expansion (19), we have*

$$(20) \quad \lim_{R \rightarrow \infty} \frac{1}{B(0, R)} \int_{B(0, R)} \varphi(\mathbf{x}) d\mathbf{x} = \hat{\varphi}(\mathbf{k}) \Big|_{\mathbf{k}=0}.$$

Compared with the crystalline approximant method, the projection method overcomes the restriction of Diophantine approximation, and is able to compute quasicrystals rather than crystalline approximants. In projection method, the decay rate of coefficients  $\hat{f}(\mathbf{k})$  is dependent on the smoothness of the quasiperiodic function. To increase numerical precision, we just need add more trigonometric functions to expand the density function.

In practice, computational domain is an important variable in calculating ordered structures. The appropriate computational box is important to determine the final morphology of solutions, especially for complicated phases Jiang, C. Wang, Huang, and P. Zhang [2013]. In physics, an equilibrium periodic structure is related to the minimum of the energy functional of order parameter  $\varphi(\mathbf{x})$  and computational domain. Therefore, the optimization problem of minimizing energy functional (12) is extended to

$$(21) \quad \min_{\varphi, \{\mathbf{b}_1, \dots, \mathbf{b}_d\}} F[\varphi(\mathbf{x}), \{\mathbf{b}_1, \dots, \mathbf{b}_d\}],$$

where  $\mathbf{b}_1, \dots, \mathbf{b}_d$  is the reciprocal primitive lattice vectors.

For a  $d$ -dimensional periodic phase, one can always choose a proper coordinate system such that the freedom of computational domain is  $d(d+1)/2$  P. Zhang and X. Zhang [2008]. Through the dual relationship,  $\mathbf{a}_i \cdot \mathbf{b}_j = 2\pi\delta_{ij}$ , optimizing computational region in reciprocal space is equivalent to doing in real space for periodic structures. However, it does not hold for quasicrystals. In projection method, the freedom of computational region of a  $d$ -dimensional quasicrystals is  $d(d+1)/2 + d(n-d)$ . If the rotational symmetry is determined in advance, the degree of freedom can be greatly reduced.

We have successfully applied the projection method to study the emergence and thermodynamic stability of quasicrystals based on the class of physical models (12). With the choice of two-scale correlation potential  $G(\mathbf{x}, \mathbf{x}') = (\nabla^2 + 1)^2(\nabla^2 + q^2)^2\delta(\mathbf{x}, \mathbf{x}')$ , we have found the two-dimensional 8-, 10-, and 12-fold symmetric quasicrystals, and obtained corresponding phase diagram including periodic crystals and quasicrystals (see Figure 1). More details can be found in Jiang, Tong, P. Zhang, and Shi [2015]. The phase

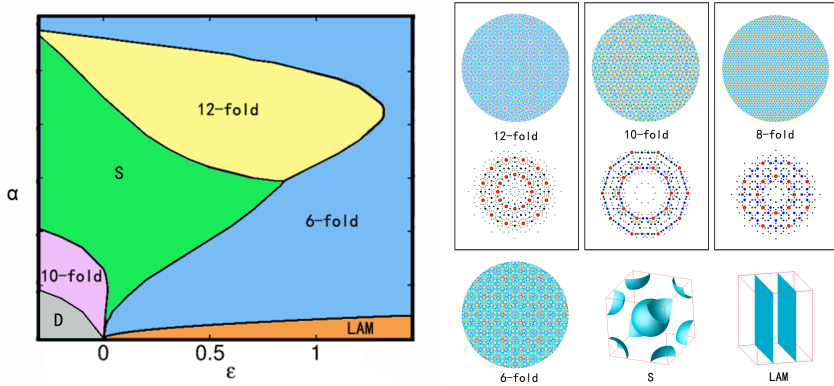


Figure 1: Phases and phase diagram of model (12) with  $G(\mathbf{x}, \mathbf{x}') = (\nabla^2 + 1)^2(\nabla^2 + q^2)^2\delta(\mathbf{x}, \mathbf{x}')$  using the projection method when  $\gamma = 100$ . The structural parameter  $q$  equals to  $2\cos(\pi/12)$ ,  $2\cos(\pi/5)$ , and  $2\cos(\pi/8)$  for 12-, 10-, and 8-fold symmetric quasicrystals, respectively. For periodic crystals,  $q = 2\cos(\pi/5)$ . The spectrum points of quasicrystals are also given. Beside ordered structures,  $D$  denotes the disordered phase. 8-fold symmetric quasicrystal is always metastable in the range of parameters of this phase diagram.

behaviour of quasicrystals in multi-component systems has been studied in Jiang, Tong, and P. Zhang [2016]. Three-dimensional icosahedral quasicrystals have been investigated in Jiang, P. Zhang, and Shi [2017] with two-scale Gaussian-polynomials type potential function.

## 4 Conclusion and Future Perspectives

Generally speaking, mathematicians investigate the quasicrystals from the abstract concepts which might not be closely related to the quasicrystals in real world. The discovered quasicrystals always depend on the underlying physical systems. Even for quasicrystals in the same category, they might have various morphologies due to the diverse building

blocks of materials. In the study of a specific system, numerical mathematics could build a bridge between mathematical and physical quasicrystals by solving a physical model of the system. During the computation of quasicrystals, numerical mathematics might provide a new possibility to understand the natural law of the quasicrystals. Meanwhile, the new generated problems in computing quasicrystals could promote the development of mathematical and physical fields. For our purpose in this article, we will give a perspective of related problems on a connection between physics and mathematics from numerical mathematics.

There are three mathematical formalisms to describe quasicrystals including point sets, dual sets, and functions. The point set represents the position of units or atoms, the dual set is for spectra, and the function is the density distribution of a quasicrystal. To study quasicrystals, we expect to integrate three formalisms under one umbrella. As far as we know, the constructive method of cut-and-project scheme fulfills the above requirements. Using this scheme, we can obtain the model set and its  $\varepsilon$ -dual set that are subsets of mathematical quasicrystals. As discussed in [Section 2.4](#), the model sets obtained by the cut-and-project scheme can explain many properties of quasicrystals. For example, cut-and-project method can produce aperiodic tilings with an appropriate choice of cut window. Besides, using Poisson's summation formula based on models sets and their  $\varepsilon$ -dual sets, the observation of only finite Bragg points in experiments can be also explained. However, the scheme still has limitations for specific physical systems. The Gaussian approximation method from the cut-and-project scheme can be used to represent quasicrystals when solving physical models. However, the numerical precision cannot be guaranteed.

In practice, the projection method demonstrates the best performance in solving physical models for quasicrystals. In this method, the quasiperiodic function is expressed as trigonometric function sum whose frequencies belong to a projection of higher-dimensional periodic lattice of (18). The decay rate of the Fourier coefficients depends on the smoothness of the quasiperiodic function. This explains that only finite spectrum points can be observed in computation and experiments. But the explanation is different from the cut-and-project method. Until now, rigorous analysis of the projection method still lacks in pure and numerical mathematics.

Many problems of quasicrystals have been solved, at least partially, with the efforts of physicists and mathematicians in the past few decades. However, the theory of quasicrystals, as well as aperiodic order, is a fast developing field. There are still lots of open problems in many areas of mathematics, physics, and numerical mathematics. Solving these problems requires the close collaboration of researchers from different areas.

An unsolved fundamental problem, in physics or material science, is why quasicrystals can emerge and be stable. In physics, we need to understand intrinsic mechanism of generating physical quasicrystals and give precise models. Current models are mainly based on energy functionals in which the important interaction potential comes from physical

understanding [Lifshitz and Petrich \[1997\]](#) or multiscale modeling [MCarley and Ashcroft \[1994\]](#). Most of these models are phenomenological, in which the relation between model parameters and the nature of the specific systems is uncertain. Establishing a physical model which is able to cover the details of a concrete system is still requiring more research.

The second open problem is mainly the computational challenge in numerical mathematics. Once we have appropriate physical models, efficient and convergent numerical methods should be developed to solve concrete models. Certainly, analysing these computational algorithms appeals to known and unknown mathematical theory. Actually, there is still lack of rigorous numerical analysis for the existing numerical methods, including projection method, crystalline approximant method, and Gaussian approximation approach. Moreover, using more properties of quasicrystals, more efficient and high accurate numerical methods could be developed.

The third open problem is whether there exist more non-periodic structures between periodic crystals and disordered phases. The positive answer is amazing. The intermediate structures may contain other aperiodic structures or beyond aperiodic structures. Solving this issue requires the multipartly efforts, including experimental discovery, physical modeling, mathematical theories and numerical methods. In mathematics, if we are lucky enough, the undiscovered non-periodic structures or part of them might belong to mathematical quasicrystals (see [Definition 5](#) or [7](#)). Otherwise, it would promote new research topics of mathematics. In physics, we expect the well defined models which can describe the detailed information of concrete systems including these new non-periodic structures. To establish the connection between the abstract mathematical theory and concrete physical systems, correspondingly, new numerical methods should be developed to solve these models. Last but not the least, these unknown non-periodic structures would give us new challenges and opportunities in mathematics, physics, numerical mathematics, and other fields.

**Acknowledgments.** The authors are grateful to Dr. Yongqiang Cai for useful discussions.

## References

- A. J. Archer, A. M. Rucklidge, and E. Knobloch (Oct. 2013). “[Quasicrystalline order and a crystal-liquid state in a soft-core fluid](#)”. *Phys. Rev. Lett.* 111 (16), p. 165501 (cit. on p. 3617).
- M. Baake and U. Grimm (2011). “[Kinematic diffraction from a mathematical viewpoint](#)”. *Z. Krist.* 226, pp. 711–725 (cit. on p. 3611).



- M. Baake and U. Grimm (2012). “[Mathematical diffraction of aperiodic structures](#)”. *Chem. Soc. Rev.* 41 (20), pp. 6821–6843 (cit. on p. 3611).
- (2013). *[Aperiodic order. Vol. 1: A mathematical invitation, Vol. 149 of encyclopedia mathematics and its applications](#)*. Cambridge University Press (cit. on pp. 3611, 3612).
- K. Barkan, H. Diamant, and R. Lifshitz (2011). “[Stability of quasicrystals composed of soft isotropic particles](#)”. *Phys. Rev. B* 83.17, p. 172201 (cit. on p. 3617).
- K. Barkan, M. Engel, and R. Lifshitz (Aug. 2014). “[Controlled self-assembly of periodic and aperiodic cluster crystals](#)”. *Phys. Rev. Lett.* 113 (9), p. 098304 (cit. on p. 3617).
- L. Bindi, P. J. Steinhardt, N. Yao, and P. J. Lu (2009). “[Natural quasicrystals](#)”. *Science* 324.5932, p. 1306 (cit. on p. 3610).
- H. Bohr (1925). “[Zur theorie der fast periodischen funktionen. I. Eine verallgemeinerung der theorie der fourierreihen](#)”. *Acta Math.* 45, pp. 29–127 (cit. on p. 3612).
- N. G. de Bruijn (1981a). “[Algebraic theory of Penrose’s non-periodic tilings of the plane, I](#)”. *Kon. Nederl. Akad. Wetensch. Proc. Ser. A* 43.84, pp. 39–52 (cit. on pp. 3611, 3612, 3615).
- (1981b). “[Algebraic theory of Penrose’s non-periodic tilings of the plane, II](#)”. *Kon. Nederl. Akad. Wetensch. Proc. Ser. A* 43.84, pp. 53–66 (cit. on pp. 3611, 3612, 3615).
- P. M. Chaikin and T. C. Lubensky (2000). *[Principles of condensed matter physics](#)*. Cambridge university press (cit. on p. 3618).
- C. Corduneanu (1968). *[Almost periodic functions, with the collaboration of N. Gheorghiu and V. Barbu, Translated from the Romanian by Gitta Bernstein and Eugene Tomer, Interscience Tracts in Pure and Applied Mathematics, No. 22. Inter-science Publishers](#)* (cit. on p. 3613).
- T. Dotera, T. Oshiro, and P. Ziherl (2014). “[Mosaic two-lengthscale quasicrystals](#)”. *Nature* 506, pp. 208–211 (cit. on p. 3617).
- F. Dyson (2009). “[Birds and frogs](#)”. *Notices of AMS* 56, pp. 212–223 (cit. on p. 3611).
- F. Gähler and J. Rhyner (1986). “[Equivalence of the generalised grid and projection methods for the construction of quasiperiodic tilings](#)”. *J. Phys. A: Math. Gen.* 19.2, p. 267 (cit. on p. 3615).
- B. Grünbaum and G. C. Shephard (1987). *[Tilings and patterns](#)*. Freeman and Company, New York (cit. on p. 3612).
- H. Hiller (1985). “[The crystallographic restriction in higher dimensions](#)”. *Acta Cryst. A* 41.6, pp. 541–544 (cit. on p. 3621).
- X. Hou and J. You (2012). “[Almost reducibility and non-perturbative reducibility of quasi-periodic linear systems](#)”. *Invent. Math.* 190, pp. 209–260 (cit. on p. 3611).
- K. Jiang, Y. Huang, and P. Zhang (2010). “[Spectral method for exploring patterns of diblock copolymers](#)”. *J. Comp. Phys.* 229.20, pp. 7796–7805 (cit. on p. 3617).

- K. Jiang, J. Tong, and P. Zhang (2016). “Stability of soft quasicrystals in a coupled-mode Swift-Hohenberg model for three-component systems”. *Commun. Comput. Phys.* 19, pp. 559–581 (cit. on pp. 3617, 3623).
- K. Jiang, J. Tong, P. Zhang, and A.-C. Shi (Oct. 2015). “Stability of two-dimensional soft quasicrystals in systems with two length scales”. *Phys. Rev. E* 92 (4), p. 042159 (cit. on pp. 3617, 3618, 3623).
- K. Jiang, C. Wang, Y. Huang, and P. Zhang (2013). “Discovery of new metastable patterns in diblock copolymers”. *Commun. Comput. Phys.* 14.2, pp. 443–460 (cit. on pp. 3617, 3622).
- K. Jiang and P. Zhang (2014). “Numerical methods for quasicrystals”. *J. Comp. Phys.* 256, pp. 428–440 (cit. on pp. 3618, 3620, 3622).
- K. Jiang, P. Zhang, and A.-C. Shi (2017). “Stability of icosahedral quasicrystals in a simple model with two-length scales”. *J. Phys.: Condens. Matter* 29.12, p. 124003 (cit. on pp. 3617, 3623).
- Y. Katznelson (2004). *An introduction to harmonic analysis, 3th edition*. Cambridge University Press (cit. on p. 3612).
- J. C. Lagarias (1996). “Meyer’s concept of quasicrystal and quasiregular sets”. *Commun. Math. Phys.* 179.2, pp. 365–376 (cit. on p. 3614).
- N. Lev and A. Oleviskii (2015). “Quasicrystals and Poisson’s summation formula”. *Invent. Math.* 2.200, pp. 585–606 (cit. on p. 3611).
- D. Levine and P. J. Steinhardt (1986). “Quasicrystals. I. Definition and structure”. *Phys. Rev. B* 34.2, p. 596 (cit. on p. 3611).
- R. Lifshitz (2003). “Quasicrystals: A matter of definition”. *Found. Phys.* 33.12, pp. 1703–1711 (cit. on p. 3611).
- R. Lifshitz and D. M. Petrich (Aug. 1997). “Theoretical model for Faraday waves with multiple-frequency forcing”. *Phys. Rev. Lett.* 79 (7), pp. 1261–1264 (cit. on pp. 3617, 3618, 3625).
- J. S. McCarley and N. W. Ashcroft (June 1994). “Hard-sphere quasicrystals”. *Phys. Rev. B* 49 (22), pp. 15600–15606 (cit. on pp. 3620, 3625).
- N. D. Mermin (1991). “Quasi crystallography is better in Fourier space”. In: *Quasicrystals: The State of the Art*. World Scientific, pp. 133–183 (cit. on p. 3611).
- Y. Meyer (1972). *Algebraic numbers and harmonic analysis*. North-Holland, Amsterdam (cit. on pp. 3611, 3614).
- (1995). “Quasicrystals, Diophantine approximation and algebraic numbers”. In: *Beyond quasicrystals*. Springer, pp. 3–16 (cit. on pp. 3615, 3616).
  - (2012). “Quasicrystals, almost periodic patterns, mean-periodic functions and irregular sampling”. *Afr. Diaspora J. Math.* 13.1, pp. 1–45 (cit. on p. 3611).
- H. W. Müller (1994). “Model equations for two-dimensional quasipatterns”. *Phys. Rev. E* 49.2, pp. 1273–1277 (cit. on p. 3617).

- R. Penrose (1974). “The role of aesthetics in pure and applied mathematical research”. *Bull. Inst. Math. Appl.* 10, pp. 266–271 (cit. on p. 3611).
- “Report of the executive committee for 1991” (1992). *Acta Crystal. A* 48, pp. 922–946 (cit. on p. 3610).
- S. Sachdev and D. R. Nelson (Oct. 1985). “Order in metallic glasses and icosahedral crystals”. *Phys. Rev. B* 32 (7), pp. 4592–4606 (cit. on p. 3620).
- R. Salem (1963). *Algebraic numbers and Fourier analysis*. Boston, Heath (cit. on p. 3611).
- M. Senechal (1995). *Quasicrystals and geometry*. Cambridge University Press (cit. on pp. 3611, 3612).
- D. Shechtman, I. Blech, D. Gratias, and J. W. Cahn (Nov. 1984). “Metallic phase with long-range orientational order and no translational symmetry”. *Phys. Rev. Lett.* 53 (20), pp. 1951–1953 (cit. on p. 3610).
- A. P. Smith (July 1990). “Electrostatic energy of a one-component quasicrystal”. *Phys. Rev. B* 42 (2), pp. 1189–1199 (cit. on p. 3620).
- W. Steurer (2004). “Twenty years of structure research on quasicrystals. Part I. Pentagonal, octagonal, decagonal and dodecagonal quasicrystals”. *Z. Krist.* 219.7/2004, pp. 391–446 (cit. on p. 3610).
- W. Steurer and S. Deloudi (2009). *Crystallography of quasicrystals: concepts, methods and structures*. Vol. 126. Springer Verlag (cit. on pp. 3611, 3621).
- A. P. Tsai (2008). “Icosahedral clusters, icosahedral order and stability of quasicrystals—view of metallurgy”. *Sci. Technol. Adv. Mat.* 9, p. 013008 (cit. on p. 3610).
- N. Wang, H. Chen, and K. H. Kuo (Aug. 1987). “Two-dimensional quasicrystal with eight-fold rotational symmetry”. *Phys. Rev. Lett.* 59 (9), pp. 1010–1013 (cit. on p. 3610).
- W. Xu, K. Jiang, P. Zhang, and A.-C. Shi (2013). “A strategy to explore stable and metastable ordered phases of block copolymers”. *J. Phys. Chem. B* 117.17. PMID: 23551204, pp. 5296–5305 (cit. on p. 3617).
- X. Zeng, G. Ungar, Y. Liu, V. Percec, A. E. Dulcey, and J. K. Hobbs (2004). “Supramolecular dendritic liquid quasicrystals”. *Nature* 428.6979, pp. 157–160 (cit. on p. 3610).
- P. Zhang and X. Zhang (2008). “An efficient numerical method of Landau-Brazovskii model”. *J. Comp. Phys.* 227.11, pp. 5859–5870 (cit. on p. 3622).

Received 2017-12-05.

KAI JIANG (蒋凯)

SCHOOL OF MATHEMATICS AND COMPUTATIONAL SCIENCE

XIANGTAN UNIVERSITY

P.R. CHINA, 411105

[kaijiang@xtu.edu.cn](mailto:kaijiang@xtu.edu.cn)

PINGWEN ZHANG (张平文)

SCHOOL OF MATHEMATICAL SCIENCES

PEKING UNIVERSITY

P.R. CHINA, 100871

[pzhang@pku.edu.cn](mailto:pzhang@pku.edu.cn)



# MATHEMATICAL ANALYSIS AND NUMERICAL METHODS FOR MULTISCALE KINETIC EQUATIONS WITH UNCERTAINTIES

SHI JIN

## Abstract

Kinetic modeling and computation face the challenges of multiple scales and uncertainties. Developing efficient multiscale computational methods, and quantifying uncertainties arising in their collision kernels or scattering coefficients, initial or boundary data, forcing terms, geometry, etc. have important engineering and industrial applications. In this article we will report our recent progress in the study of multiscale kinetic equations with uncertainties modelled by random inputs. We first study the mathematical properties of uncertain kinetic equations, including their regularity and long-time behavior in the random space, and sensitivity of their solutions with respect to the input and scaling parameters. Using the hypocoercivity of kinetic operators, we provide a general framework to study these mathematical properties for general class of linear and nonlinear kinetic equations in various asymptotic regimes. We then approximate these equations in random space by the stochastic Galerkin methods, study the numerical accuracy and long-time behavior of the methods, and furthermore, make the methods “stochastically asymptotic preserving”, in order to handle the multiple scales efficiently.

## 1 Introduction

Kinetic equations describe the probability density function of a gas or system comprised of a large number of particles. In multiscale modeling hierarchy, they serve as the bridge between atomistic and continuum models. On one hand, since they model the collective dynamics of particles, thus are more efficient than molecular dynamics; on the other

---

This author’s research was supported by NSF grants DMS-1522184 and DMS-1107291: RNMS KI-Net, NSFC grant No. 91330203, and the Office of the Vice Chancellor for Research and Graduate Education at the University of Wisconsin-Madison with funding from the Wisconsin Alumni Research Foundation.

*MSC2010:* primary 82B40; secondary 25Q20, 82D75, 65M70.

*Keywords:* Kinetic equations, multiscales, asymptotic-preserving, uncertainty quantification, sensitivity analysis, stochastic Galerkin, hypocoercivity.

hand, they provide more accurate solutions when the macroscopic fluid mechanics laws of Navier-Stokes and Fourier become inadequate. The most fundamental kinetic equation is the Boltzmann equation, an integro-differential equation describing particle transport with binary collisions [Chapman and Cowling \[1991\]](#) and [Cercignani \[1988\]](#). Now kinetic theory has seen expanding applications from rarefied gas dynamics [Cercignani \[2000\]](#), radiative transfer [Chandrasekhar \[1960\]](#), medical imaging [Arridge \[1999\]](#), plasma physics [Degond and Deluzet \[2017a\]](#), to microfabrication technology [Markowich, Ringhofer, and Schmeiser \[1990\]](#) and [Jüngel \[2009\]](#), biological and even social sciences [Naldi, Pareschi, and Toscani \[2010\]](#).

There are three main computational challenges in kinetic modeling and simulation: *Dimension curse*, *multiple scales*, and *uncertainty*.

A kinetic equation solves the particle density distribution  $f(t, x, v)$ , which depends on time  $t \in \mathbb{R}^+$ , space  $x \in \mathbb{R}^d$ , and particle velocity  $v \in \mathbb{R}^d$ . Typically,  $d = 3$ , therefore one has to solve a six dimensional differential-integral equation plus time.

Kinetic equations often have multiple scales, characterized by the Knudsen number  $\varepsilon$ , the ratio of particle mean free path over a typical length scale, which can vary spatially dramatically. In these problems, multiscale and multi physics modelings are essential. For example, in the space shuttle reentry problem, along the vehicle trajectory, one encounters free streaming, rarefied gas (described by the Boltzmann equation), transition to the macroscopic hydrodynamic (described by the Euler or Navier-Stokes equations) regimes. In this process the mean free path changes from  $O(1)$  meters to  $O(10^{-8})$  meters [Rivell \[2006\]](#). In plasma physics, one has to match the plasma and sheath where the quasineutral (which allows macroscopic modeling) and non-quasineutral (which needs kinetic modeling) models need to be coupled [Franklin and Ockendon \[1970\]](#). These multiscale and multi-physics problems pose tremendous numerical challenges, with stiff collision terms, strong (electric or magnetic) fields, fast advection speed, and long-time behavior that require prohibitively small time step and mesh size in order to obtain reliable computational results.

Another challenge, which has been ignored in the community, is the issue of *uncertainties* in kinetic models. In reality, there are many sources of uncertainties that can arise in these equations, such as collision kernels, scattering coefficients, initial or boundary data, geometry, source or forcing terms [Bird \[1994\]](#), [Berman, Haverkort, and Woerdman \[1986\]](#), and [Koura and Matsumoto \[1991\]](#). Understanding the impact of these uncertainties is crucial to the simulations of the complex kinetic systems in order to validate and improve these models.

To characterize the uncertainty, we assume that certain quantities depend on a random vector  $\mathbf{z} \in \mathbb{R}^n$  in a properly defined probability space  $(\Sigma, \mathcal{Q}, \mathbb{P})$ , whose event space is  $\Sigma$  and is equipped with  $\sigma$ -algebra  $\mathcal{Q}$  and probability measure  $\mathbb{P}$ . We also assume

the components of  $\mathbf{z}$  are mutually independent random variables with known probability  $\omega(\mathbf{z}) : I_{\mathbf{z}} \rightarrow \mathbb{R}^+$ , obtained already through some dimension reduction technique, e.g., Karhunen-Loève (KL) expansion [Loève \[1977\]](#), and do not pursue further the issue of random input parameterization.

Although uncertainty quantification (UQ) has been a popular field in scientific and engineering computing in the last two decades, UQ for kinetic equations has been largely an open area until recently. In this article we will present some of our recent results in UQ for multiscale kinetic equations. We will use *hypocoercivity* of kinetic operators to study the regularity and long-time behavior in the random space, as well as sensitivity of the solutions with respect to the random input parameters. Our results are fairly general, covering most important linear and nonlinear kinetic equations, including the Boltzmann, Landau, semi-classical relaxation models, and the Vlasov-Poisson-Fokker-Planck equations. We then introduce the stochastic Galerkin method for random kinetic equations, and study their numerical accuracy and long-time behavior, and formulate them as *stochastic asymptotic-preserving* methods for multiscale kinetic equations with uncertainties, which allows one to solve these problems with *all* numerical parameters—including the degree of orthogonal polynomials used in the polynomial chaos expansions—*independent of the Kundsén number*.

## 2 Basic mathematical theory for uncertain kinetic equations

**2.1 The linear transport equation with isotropic scattering.** We first introduce the linear transport equation in one dimensional slab geometry:

$$(2-1) \quad \varepsilon \partial_t f + v \partial_x f = \frac{\sigma}{\varepsilon} \mathcal{L} f - \varepsilon \sigma^a f + \varepsilon S, \quad t > 0, x \in [0, 1], v \in [-1, 1], z \in I_z,$$

$$(2-2) \quad \mathcal{L} f(t, x, v, z) = \frac{1}{2} \int_{-1}^1 f(t, x, v', z) dv' - f(t, x, v, z),$$

with the initial condition

$$(2-3) \quad f(0, x, v, z) = f^0(x, v, z).$$

This equation arises in neutron transport, radiative transfer, etc. and describes particles (for example neutrons) transport in a background media (for example nuclei).  $\mathcal{L}$  is the collision operator,  $v = \Omega \cdot e_x = \cos \theta$  where  $\theta$  is the angle between the moving direction and  $x$ -axis.  $\sigma(x, z)$ ,  $\sigma^a(x, z)$  are total and absorption cross-sections respectively.  $S(x, z)$  is the source term. For  $\sigma(x, z)$ , we assume

$$(2-4) \quad \sigma(x, z) \geq \sigma_{\min} > 0.$$



The equation is scaled in long time with strong scattering.

Denote

$$(2-5) \quad [\phi] = \frac{1}{2} \int_{-1}^1 \phi(v) dv$$

as the average of a velocity dependent function  $\phi$ .

Define in the Hilbert space  $L^2([-1, 1]; \phi^{-1} dv)$  the inner product and norm

$$(2-6) \quad \langle f, g \rangle_\phi = \int_{-1}^1 f(v)g(v)\phi^{-1} dv, \quad \|f\|_\phi^2 = \langle f, f \rangle_\phi.$$

The linear operator  $\mathfrak{L}$  satisfies the following *coercivity* properties Bardos, Santos, and Sentis [1984]:  $\mathfrak{L}$  is non-positive self-adjoint in  $L^2([-1, 1]; \phi^{-1} dv)$ , i.e., there is a positive constant  $s_m$  such that

$$(2-7) \quad \langle f, \mathfrak{L}f \rangle_\phi \leq -2s_m \|f\|_\phi^2, \quad \forall f \in \mathfrak{N}(\mathfrak{L})^\perp;$$

with  $\mathfrak{N}(\mathfrak{L}) = \text{span} \{ \phi \mid \phi = [\phi] \}$  the null space of  $\mathfrak{L}$ .

Let  $\rho = [f]$ . For each fixed  $z$ , the classical diffusion limit theory of linear transport equation Larsen and Keller [1974], Bensoussan, Lions, and Papanicolaou [1979], and Bardos, Santos, and Sentis [1984] gives that, as  $\varepsilon \rightarrow 0$ ,  $\rho$  solves the following diffusion equation:

$$(2-8) \quad \partial_t \rho = \partial_x \left( \frac{1}{3} \sigma(x, z)^{-1} \partial_x \rho \right) - \sigma^a(x, z) \rho + S(x, z).$$

To study the regularity and long-time behavior in the random space of the linear transport Equation (2-1)-(2-3), we use the Hilbert space of the random variable

$$(2-9) \quad H(I_z; \omega dz) = \left\{ f \mid I_z \rightarrow \mathbb{R}^+, \int_{I_z} f^2(z) \omega(z) dz < +\infty \right\},$$

equipped with the inner product and norm defined as

$$(2-10) \quad \langle f, g \rangle_\omega = \int_{I_z} fg \omega(z) dz, \quad \|f\|_\omega^2 = \langle f, f \rangle_\omega.$$

We also define the  $k$ th order differential operator with respect to  $z$  as

$$(2-11) \quad D^k f(t, x, v, z) := \partial_z^k f(t, x, v, z),$$

and the Sobolev norm in  $z$  as

$$(2-12) \quad \|f(t, x, v, \cdot)\|_{H^k}^2 := \sum_{\alpha \leq k} \|D^\alpha f(t, x, v, \cdot)\|_\omega^2.$$

Finally, we introduce norms in space and velocity as follows,

$$(2-13) \quad \|f(t, \cdot, \cdot, \cdot)\|_{\Gamma}^2 := \int_Q \|f(t, x, v, \cdot)\|_{\omega}^2 dx dv, \quad t \geq 0,$$

$$(2-14) \quad \|f(t, \cdot, \cdot, \cdot)\|_{\Gamma^k}^2 := \int_Q \|f(t, x, v, \cdot)\|_{H^k}^2 dx dv, \quad t \geq 0,$$

where  $Q = [0, 1] \times [-1, 1]$  denotes the domain in the phase space. For simplicity of notations, we will suppress the dependence of  $t$  and just use  $\|f\|_{\Gamma}$ ,  $\|f\|_{\Gamma^k}$  in the following results, which were established in [Jin, J.-G. Liu, and Ma \[2017\]](#).

**Theorem 2.1 (Uniform regularity).** *If for some integer  $m \geq 0$ ,*

$$(2-15) \quad \|D^k \sigma(z)\|_{L^\infty} \leq C_\sigma, \quad \|D^k f_0\|_{\Gamma} \leq C_0, \quad k = 0, \dots, m,$$

*then the solution  $f$  to the linear transport [Equation \(2-1\)–\(2-3\)](#), with  $\sigma^a = S = 0$  and periodic boundary condition in  $x$ , satisfies,*

$$(2-16) \quad \|D^k f\|_{\Gamma} \leq C, \quad k = 0, \dots, m, \quad \forall t > 0,$$

*where  $C_\sigma$ ,  $C_0$  and  $C$  are constants independent of  $\varepsilon$ .*

The above theorem shows that, under some smoothness assumption on  $\sigma$ , the regularity of the initial data is preserved in time and the Sobolev norm of the solution is bounded uniformly in  $\varepsilon$ .

**Theorem 2.2 ( $\varepsilon^2$ -estimate on  $[f] - f$ ).** *With all the assumptions in [Theorem 2.1](#) and furthermore,  $\sigma \in W^{k, \infty} = \{\sigma \in L^\infty([0, 1] \times I_z) \mid D^j \sigma \in L^\infty([0, 1] \times I_z) \text{ for all } j \leq k\}$ . For a given time  $T > 0$ , the following regularity result of  $[f] - f$  holds:*

$$(2-17) \quad \|D^k([f] - f)\|_{\Gamma}^2 \leq e^{-\sigma_{\min} t / 2\varepsilon^2} \|D^k([f_0] - f_0)\|_{\Gamma}^2 + C' \varepsilon^2$$

*for any  $t \in (0, T]$  and  $0 \leq k \leq m$ , where  $C'$  and  $C$  are constants independent of  $\varepsilon$ .*

The first term on the right hand side of (2-17) is the behavior of the initial layer, which is damped exponentially in  $t/\varepsilon^2$ . After the initial layer, the high order derivatives in  $z$  of the difference between  $f$  and its local equilibrium  $[f]$  is of  $O(\varepsilon)$ .

Such results have been generalized to linear anisotropic collision operators in [L. Liu \[n.d.\]](#). For general linear collision operators conserving mass, the hypocoercivity framework of [Dolbeault, Mouhot, and Schmeiser \[2015\]](#) was first used by [Li and Wang \[n.d.\]](#) to prove regularity in the random space with sharp constants.

**2.2 General collisional nonlinear kinetic equations with random uncertainties.** Consider the initial value problem for kinetic equations of the form

$$(2-18) \quad \begin{cases} \partial_t f + \frac{1}{\varepsilon^\alpha} v \cdot \nabla_x f = \frac{1}{\varepsilon^{1+\alpha}} \mathcal{Q}(f), \\ f(0, x, v, z) = f_{in}(x, v, z), \quad x \in \Omega \subset \mathbb{T}^d, v \in \mathbb{R}^d, z \in I_z \subset \mathbb{R}. \end{cases}$$

The operator  $\mathcal{Q}$  models the collisional interactions of particles, which is either binary or between particles and a surrounding medium.  $\alpha = 1$  is referred to the *incompressible Navier-Stokes* scaling, while  $\alpha = 0$  corresponds to the Euler (or acoustic) scaling. The periodic boundary conditions for the spatial domain  $\Omega = \mathbb{T}^d$  is assumed here for theoretical purpose. In the sequel  $\mathcal{L}$  is used for both the linear collision operator and the linearized collision operator for nonlinear equations. Consider the linearized equation

$$(2-19) \quad \partial_t g + \frac{1}{\varepsilon^\alpha} v \cdot \nabla_x g = \frac{1}{\varepsilon^{1+\alpha}} \mathcal{L}(g),$$

Since  $\mathcal{L}$  is not fully dissipative, as summarized in [S. Daus, Jüngel, Mouhot, and Zamponi \[2016\]](#) and [Dolbeault, Mouhot, and Schmeiser \[2015\]](#), the idea is to use the hypocoercivity of the linearized kinetic operator

$$\mathcal{G} = \frac{1}{\varepsilon^{1+\alpha}} \mathcal{L} - \frac{1}{\varepsilon^\alpha} \mathcal{T},$$

where  $\mathcal{T} = v \cdot \nabla_x$  is the streaming operator, using the dissipative properties of  $\mathcal{L}$  and the conservative properties of  $\mathcal{T}$ . The aim is to find a Lyapunov type functional  $\eta[h]$  which is equivalent to the square of the norm of a Banach space, for example

$$H_{x,v}^1 = \left\{ f \mid \int_{\Omega \times \mathbb{R}^d} \sum_{|i|+|j| \leq 1} \|\partial_{x_i} \partial_{v_j} g\|_{L_{x,v}^2}^2 dx dv < \infty \right\},$$

such that

$$\kappa_1 \|g\|_{H_{x,v}^1} \leq \eta[g] \leq \kappa_2 \|g\|_{H_{x,v}^1}, \quad \text{for } g \in H_{x,v}^1,$$

which leads to

$$\frac{d}{dt} \eta[g(t)] \leq -\kappa \|g(t)\|_{H_{x,v}^1}, \quad t > 0,$$

with constants  $\kappa_1, \kappa_2, \kappa > 0$ . Then one concludes the exponential convergence of  $g$  in  $H_{x,v}^1$ . The obvious choice of  $\eta[g] = c_1 \|g\|_{L_{x,v}^2}^2 + c_2 \|\nabla_x g\|_{L_{x,v}^2}^2 + c_3 \|\nabla_v g\|_{L_{x,v}^2}^2$  does not work, since the collision operator is not coercive. The key idea, first seen in [Villani \[2009\]](#) and implemented in [Mouhot and Neumann \[2006\]](#), is to add the “mixing term”  $c \langle \nabla_x g, \nabla_v g \rangle_{L_{x,v}^2}$  to the definition of  $\eta[g]$ , that is

$$\frac{d}{dt} \langle \nabla_x g, \nabla_v g \rangle_{L_{x,v}^2} = -\|\nabla_x g\|_{L_{x,v}^2}^2 + 2 \langle \nabla_x \mathcal{L}(g), \nabla_v g \rangle_{L_{x,v}^2}.$$

Mouhot and Neumann [ibid.] discusses the linearized equation  $\partial_t g + v \cdot \nabla_x g = \mathfrak{L}(g)$  and proves that if the linear operator  $\mathfrak{L}$  satisfies some assumptions, then  $\mathfrak{L} - v \cdot \nabla_x$  generates a strongly continuous evolution semi-group  $e^{t\mathfrak{G}}$  on  $H_{x,v}^s$ , which satisfies

$$(2-20) \quad \|e^{t\mathfrak{G}}(\mathbb{I} - \Pi_{\mathfrak{G}})\|_{H_{x,v}^s} \leq C \exp[-\tau t],$$

for some explicit constants  $C$ ,  $\tau > 0$  depending only on the constants determined by the equation itself. Here  $\Pi_{\mathfrak{G}}$  is the orthogonal projection in  $L_v^2$  onto the null space of  $\mathfrak{L}$ . This result shows that apart from 0, the spectrum of  $\mathfrak{G}$  is included in

$$\{\xi \in \mathbb{C} : \operatorname{Re}(\xi) \leq -\tau\}.$$

For nonlinear kinetic equations, the main idea is to use *the perturbative setting* Guo [2006] and Strain and Guo [2008]. Equations defined in (2-18) admit a unique global equilibrium in the torus, denoted by  $\mathfrak{M}$  which is independent of  $t, x$ . Now consider the linearization around this equilibrium and perturbation of the solution of the form

$$(2-21) \quad f = \mathfrak{M} + \varepsilon M h$$

with  $\mathfrak{M}$  being the global equilibrium (or global) Maxwellian, and  $M = \sqrt{\mathfrak{M}}$ . Then  $h$  satisfies

$$(2-22) \quad \partial_t h + \frac{1}{\varepsilon^\alpha} v \cdot \nabla_x h = \frac{1}{\varepsilon^{1+\alpha}} \mathfrak{L}(h) + \frac{1}{\varepsilon^\alpha} \mathfrak{F}(h, h).$$

$\mathfrak{L}$  is the linearized (around  $\mathfrak{M}$ ) collision operator acting on  $L_v^2 = \{f \mid \int_{\mathbb{R}^d} f^2 dv < \infty\}$ , with the kernel denoted by  $N(\mathfrak{L}) = \operatorname{span}\{\psi_1, \dots, \psi_d\}$ .  $\{\psi_i\}_{1 \leq i \leq d}$  is an orthonormal family of polynomials in  $v$  corresponding to the manifold of local equilibria for the linearized kinetic models. The orthogonal projection on  $N(\mathfrak{L})$  in  $L_v^2$  is defined by

$$(2-23) \quad \Pi_{\mathfrak{L}}(h) = \sum_{i=1}^n \left( \int_{\mathbb{R}^d} h \psi_i dv \right) \psi_i,$$

where  $\Pi_{\mathfrak{L}}$  is the projection on the 'fluid part' and  $\mathbb{I} - \Pi_{\mathfrak{L}}$  is the projection on the kinetic part, with  $\mathbb{I}$  the identity operator. The global equilibrium is then

$$(2-24) \quad \mathfrak{M} = \Pi_{\mathfrak{G}}(h) = \sum_{i=1}^n \left( \int_{\mathbb{T}^d \times \mathbb{R}^d} h \psi_i dx dv \right) \psi_i,$$

which is independent of  $x$  and  $t$  and is the orthogonal projection on  $N(\mathfrak{G}) = N(\mathfrak{L})$  in  $L_{x,v}^2 = \{f \mid \int_{\Omega \times \mathbb{R}^d} f^2 dx dv < \infty\}$ .

Since the linear part  $\frac{1}{\varepsilon^{1+\alpha}} \mathfrak{L}$  part has one extra factor of  $\frac{1}{\varepsilon}$  than the nonlinear part  $\frac{1}{\varepsilon^\alpha} \mathfrak{T}$ , one hopes to use the hypocoercivity from the linear part to control the nonlinear part in order to come up with the desired decay estimate. This is only possible for initial data close to  $\mathfrak{M}$ , as in (2-21). In addition, one needs some assumptions on these operators, which can be checked for a number of important collision kernels, such as the Boltzmann, Landau, and semi-classical relaxation models [Briant \[2015\]](#).

**Assumption on the linear operator  $\mathfrak{L}$ .**  $\mathfrak{L}$  has the local coercivity property: There exists  $\lambda > 0$  such that  $\forall h \in L_v^2$ ,

$$(2-25) \quad \langle \mathfrak{L}(h), h \rangle_{L_v^2} \leq -\lambda \|h^\perp\|_{\Lambda_v}^2,$$

where

$$h^\perp = h - \Pi_{\mathfrak{L}}(h)$$

stands for the microscopic part of  $h$ , which satisfies  $h^\perp \in N(\mathfrak{L})^\perp$  in  $L_v^2$ . Here  $\Lambda_v$ -norm is collision operator specific. For the Boltzmann collision operator, it is given in (2-33).

To extend to higher-order Sobolev spaces, let us first introduce some notations of multi-indices and Sobolev norms. For two multi-indices  $j$  and  $l$  in  $\mathbb{N}^d$ , define

$$\partial_l^j = \partial / \partial v_j \partial / \partial x_l.$$

For  $i \in \{1, \dots, d\}$ , denote by  $c_i(j)$  the value of the  $i$ -th coordinate of  $j$  and by  $|j|$  the  $l^1$  norm of the multi-index, that is,  $|j| = \sum_{i=1}^d c_i(j)$ . Define the multi-index  $\delta_{i_0}$  by:  $c_i(\delta_{i_0}) = 1$  if  $i = i_0$  and 0 otherwise. We use the notation

$$\partial_z^\alpha h = \partial^\alpha h.$$

Denote  $\|\cdot\|_\Lambda := \|\cdot\|_{\Lambda_v} \| \cdot \|_{L_x^2}$ . The Sobolev norms on  $H_{x,v}^s$  and  $H_\Lambda^s$  are defined by

$$\|h\|_{H_{x,v}^s}^2 = \sum_{|j|+|l|\leq s} \|\partial_l^j h\|_{L_{x,v}^2}^2, \quad \|h\|_{H_\Lambda^s}^2 = \sum_{|j|+|l|\leq s} \|\partial_l^j h\|_\Lambda^2.$$

Define the sum of Sobolev norms of the  $z$  derivatives by

$$\begin{aligned} \|h\|_{H_{x,v}^{s,r}}^2 &= \sum_{|m|\leq r} \|\partial^m h\|_{H_{x,v}^s}^2 \\ \|h\|_{H_\Lambda^{s,r}}^2 &= \sum_{|m|\leq r} \|\partial^m h\|_{H_\Lambda^s}^2 \\ \|h\|_{H_x^{s,r} L_v^2}^2 &= \sum_{|m|\leq r} \|\partial^m h\|_{H_x^s L_v^2}^2 \end{aligned}$$

Note that these norms are all functions of  $z$ . Define the norms in the  $(x, v, z)$  space

$$\begin{aligned} \|h(x, v, \cdot)\|_{H_z^s}^2 &= \int_{I_z} \|h\|_{H_{x,v}^s}^2 \pi(z) dz \\ \|h(x, v, \cdot)\|_{H_{x,v}^s H_z^r}^2 &= \int_{I_z} \|h\|_{H_{x,v}^{s,r}}^2 \pi(z) dz \end{aligned}$$

in addition to the sup norm in  $z$  variable,

$$\|h\|_{H_{x,v}^s L_z^\infty} = \sup_{z \in I_z} \|h\|_{H_{x,v}^s}.$$

**Assumptions on the nonlinear term  $\mathfrak{F}$ :**  $\mathfrak{F} : L_v^2 \times L_v^2 \rightarrow L_v^2$  is a bilinear symmetric operator such that for all multi-indexes  $j$  and  $l$  such that  $|j| + |l| \leq s$ ,  $s \geq 0$ ,  $m \geq 0$ ,

$$(2-26) \quad \left| \langle \partial_l^m \mathfrak{F}(h, h), f \rangle_{L_{x,v}^2} \right| \leq \begin{cases} \mathfrak{G}_{x,v,z}^{s,m}(h, h) \|f\|_\Lambda, & \text{if } j \neq 0, \\ \mathfrak{G}_{x,z}^{s,m}(h, h) \|f\|_\Lambda, & \text{if } j = 0. \end{cases}$$

Sum up  $m = 0, \dots, r$ , then  $\exists s_0 \in \mathbb{N}$ ,  $\forall s \geq s_0$ , there exists a  $z$ -independent  $C_{\mathfrak{F}} > 0$  such that for all  $z$ ,

$$\begin{aligned} \sum_{|m| \leq r} (\mathfrak{G}_{x,v,z}^{s,m}(h, h))^2 &\leq C_{\mathfrak{F}} \|h\|_{H_{x,v}^{s,r}}^2 \|h\|_{H_\Lambda^{s,r}}^2 \\ \sum_{|m| \leq r} (\mathfrak{G}_{x,z}^{s,m}(h, h))^2 &\leq C_{\mathfrak{F}} \|h\|_{H_{x,z}^{s,r} L_v^2}^2 \|h\|_{H_\Lambda^{s,r}}^2 \end{aligned}$$

With uncertainty in the equation, following the deterministic framework in [Briant \[ibid.\]](#), we define a Lyapunov type functional

$$(2-27) \quad \begin{aligned} \|\cdot\|_{\mathfrak{H}_{\varepsilon_\perp}^s}^2 &= \sum_{|j|+|l| \leq s, |j| \geq 1} b_{j,l}^{(s)} \|\partial_l^j (\mathbb{I} - \Pi_{\mathfrak{L}}) \cdot\|_{L_{x,v}^2}^2 + \sum_{|l| \leq s} \alpha_l^{(s)} \|\partial_l^0 \cdot\|_{L_{x,v}^2}^2 \\ &+ \sum_{|l| \leq s, i, c_i(l) > 0} \varepsilon a_{i,l}^{(s)} \langle \partial_{l-\delta_i}^{\delta_i} \cdot, \partial_l^0 \cdot \rangle_{L_{x,v}^2}, \end{aligned}$$

and the corresponding Sobolev norms

$$\|h\|_{\mathfrak{H}_{\varepsilon_\perp}^{s,r}}^2 = \sum_{|m| \leq r} \|\partial^m h\|_{\mathfrak{H}_{\varepsilon_\perp}^s}^2, \quad \|h\|_{\mathfrak{H}_{\varepsilon_\perp}^{s,r} L_z^\infty} = \sup_{z \in I_z} \|h\|_{\mathfrak{H}_{\varepsilon_\perp}^{s,r}}.$$

The following theorem is from [L. Liu and Jin \[2017\]](#):

**Theorem 2.3.** For all  $s \geq s_0$ ,  $\exists (b_{j,l}^{(s)}), (\alpha_l^{(s)}), (a_{i,l}^{(s)}) > 0$  and  $0 \leq \varepsilon_d \leq 1$ , such that for all  $0 \leq \varepsilon \leq \varepsilon_d$ ,

$$(1) \quad \|\cdot\|_{\mathcal{H}_{\varepsilon_\perp}^s} \sim \|\cdot\|_{H_{x,v}^s};$$

(2) Assume  $\|h_{in}\|_{H_{x,v}^s L_z^\infty} \leq C_I$ , then if  $h_\varepsilon$  is a solution of (2-22) in  $H_{x,v}^s$  for all  $z$ , we have

$$(2-28) \quad \|h_\varepsilon\|_{H_{x,v}^{s,r} L_z^\infty} \leq C_I e^{-\tau_s t}, \quad \|h_\varepsilon\|_{H_{x,v}^s H_z^r} \leq C_I e^{-\tau_s t}, \quad \text{for } \alpha = 1;$$

$$(2-29) \quad \|h_\varepsilon\|_{H_{x,v}^{s,r} L_z^\infty} \leq C_I e^{-\varepsilon \tau_s t}, \quad \|h_\varepsilon\|_{H_{x,v}^s H_z^r} \leq C_I e^{-\varepsilon \tau_s t}, \quad \text{for } \alpha = 0,$$

where  $C_I, \tau_s$  are positive constants independent of  $\varepsilon$ .

**Remark 2.4.** Theorem 2.3 provides the regularity of  $h$  (thus  $f$ ) in the random space, which preserves the regularity of the initial data in time. Furthermore, it shows that the uncertainty from the initial datum will eventually diminish and the solution will exponentially decay to the deterministic global equilibrium in the long time, with a decay rate of  $\mathcal{O}(e^{-t})$  under the incompressible Navier-Stokes scaling and  $\mathcal{O}(e^{-\varepsilon t})$  under the acoustic scaling.

**2.3 The Boltzmann equation with uncertainties.** As an example of the general theory in subSection 2.2, we consider the Boltzmann equation with uncertain initial data and uncertain collision kernel:

$$(2-30) \quad \begin{cases} \partial_t f + \frac{1}{\varepsilon^\alpha} v \cdot \nabla_x f = \frac{1}{\varepsilon^{1+\alpha}} \mathcal{Q}(f, f), \\ f(0, x, v, z) = f^0(x, v, z), \end{cases} \quad x \in \Omega \subset \mathbb{T}^d, v \in \mathbb{R}^d, z \in I_z.$$

The collision operator is

$$\mathcal{Q}(f, f) = \int_{\mathbb{R}^d \times \mathbb{S}^{d-1}} B(|v - v_*|, \cos \theta, z) (f' f'_* - f f_*) dv_* d\sigma.$$

We adopt notations  $f' = f(v')$ ,  $f_* = f(v_*)$  and  $f'_* = f(v'_*)$ , where

$$v' = (v + v_*)/2 + (|v - v_*|/2)\sigma, \quad v'_* = (v + v_*)/2 - (|v - v_*|/2)\sigma$$

are the post-collisional velocities of particles with pre-collisional velocities  $v$  and  $v_*$ .  $\theta \in [0, \pi]$  is the deviation angle between  $v' - v'_*$  and  $v - v_*$ . The global equilibrium distribution is given by the Maxwellian distribution

$$(2-31) \quad \mathfrak{M}(\rho_\infty, u_\infty, T_\infty) = \frac{\rho_\infty}{(2\pi T_\infty)^{N/2}} \exp\left(-\frac{|u_\infty - v|^2}{2T_\infty}\right),$$

where  $\rho_\infty$ ,  $u_\infty$ ,  $T_\infty$  are the density, mean velocity and temperature of the gas

$$\begin{aligned}\rho_\infty &= \int_{\Omega \times \mathbb{R}^d} f(v) dx dv, & u_\infty &= \frac{1}{\rho_\infty} \int_{\Omega \times \mathbb{R}^d} v f(v) dx dv, \\ T_\infty &= \frac{1}{N\rho_\infty} \int_{\Omega \times \mathbb{R}^d} |u_\infty - v|^2 f(v) dx dv,\end{aligned}$$

which are all determined by the initial datum due to the conservation properties. We will consider hard potentials with  $B$  satisfying Grad's angular cutoff, that is,

$$B(|v - v_*|, \cos \theta, z) = \phi(|v - v_*|) b(\cos \theta, z), \quad \phi(\xi) = C_\phi \xi^\gamma, \text{ with } \gamma \in [0, 1], \quad (2-32)$$

$$\forall \eta \in [-1, 1], |b(\eta, z)| \leq C_b, |\partial_\eta b(\eta, z)| \leq C_b, |\partial_z^k b(\eta, z)| \leq C_b^*, \forall 0 \leq k \leq r.$$

where  $b$  is non-negative and not identically equal to 0. Recall that  $h$  solves (2-22), with the linearized collision operator given by

$$\mathfrak{L}(h) = M^{-1} [\mathbb{Q}(Mh, \mathfrak{M}) + \mathbb{Q}(\mathfrak{M}, Mh)],$$

while the bilinear part is given by

$$\begin{aligned}\mathfrak{F}(h, h) &= 2M^{-1} \mathbb{Q}(Mh, Mh) \\ &= \int_{\mathbb{R}^d \times \mathbb{S}^{d-1}} \phi(|v - v_*|) b(\cos \theta, z) M_* (h'_* h' - h_* h) dv_* d\sigma.\end{aligned}$$

The the coercivity norm used in (2-25) is

$$(2-33) \quad \|h\|_\Lambda = \|h(1 + |v|)^{\gamma/2}\|_{L^2}.$$

The coercivity argument of  $\mathfrak{L}$  is proved in Mouhot [2006]:

$$(2-34) \quad -\langle h, \mathfrak{L}(h) \rangle_{L_v^2} \geq \lambda \|h^\perp\|_{\Lambda_v^2}.$$

Explicit spectral gap estimates for the linearized Boltzmann and Landau operators with hard potentials have been obtained in Mouhot and Baranger [2005] and extended to estimates given in Mouhot [2006]. Proofs of  $\mathfrak{L}$  satisfying Equation (2-25) and  $\mathfrak{F}$  satisfying (2-26), even for random collision kernel satisfying conditions given in (2-32), were given in L. Liu and Jin [2017]. Thus Theorem 2.3 holds for the Boltzmann equation with random initial data and collision kernel. Similar results can be extended to Landau equation and semi-classical relaxation model, see L. Liu and Jin [ibid.].



**2.4 The Vlasov-Poisson-Fokker-Planck system.** One kinetic equation which does not fit the collisional framework presented in subSection 2.2 is the Vlasov-Poisson-Fokker-Planck (VPFP) system that arises in the kinetic modeling of the Brownian motion of a large system of particles in a surrounding bath Chandrasekhar [1943]. One application of such system is the electrostatic plasma, in which one considers the interactions between the electrons and a surrounding bath via the Coulomb force. With the electrical potential  $\phi(t, \mathbf{x}, \mathbf{z})$ , the equations read

$$(2-35) \quad \begin{cases} \partial_t f + \frac{1}{\delta} \partial_x f - \frac{1}{\varepsilon} \partial_x \phi \partial_v f = \frac{1}{\delta \varepsilon} \mathfrak{F} f, \\ -\partial_{xx} \phi = \rho - 1, \quad t > 0, \quad x \in \Omega \subset \mathbb{R}, \quad v \in \mathbb{R}, \quad z \in I_z, \end{cases}$$

with initial condition

$$(2-36) \quad f(0, x, v, z) = f^0(x, v, z).$$

Here,  $\mathfrak{F}$  is the Fokker-Planck operator describing the Brownian motion of the particles,

$$(2-37) \quad \mathfrak{F} f = \partial_v \left( \mathfrak{M} \nabla_v \left( \frac{f}{\mathfrak{M}} \right) \right),$$

where  $\mathfrak{M}$  is the *global equilibrium* or *global Maxwellian*,

$$(2-38) \quad \mathfrak{M} = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{|v|^2}{2}}.$$

$\delta$  is the reciprocal of the scaled thermal velocity,  $\varepsilon$  represents the scaled thermal mean free path. There are two different regimes for this system. One is the *high field regime*, where  $\delta = 1$ . As  $\varepsilon \rightarrow 0$ ,  $f$  goes to the local Maxwellian  $\mathfrak{M}_l = \frac{1}{(2\pi)^{\frac{d}{2}}} e^{-\frac{|v - \nabla_x \phi|^2}{2}}$ , and the VPFP system converges to a hyperbolic limit Arnold, Carrillo, Gamba, and C.-W. Shu [2001], Goudon, Nieto, Poupaud, and Soler [2005], and Nieto, Poupaud, and Soler [2001]:

$$(2-39) \quad \begin{cases} \partial_t \rho + \nabla_x \cdot (\rho \nabla_x \phi) = 0, \\ -\Delta_x \phi = \rho - 1. \end{cases}$$

Another regime is the *parabolic regime*, where  $\delta = \varepsilon$ . When  $\varepsilon \rightarrow 0$ ,  $f$  goes to the global Maxwellian  $\mathfrak{M}$ , and the VPFP system converges to a parabolic limit Poupaud and Soler [2000]:

$$(2-40) \quad \begin{cases} \partial_t \rho - \nabla_x \cdot (\nabla_x \rho - \rho \nabla_x \phi) = 0, \\ -\Delta_x \phi = \rho - 1. \end{cases}$$

Define the  $L^2$  space in the measure of

$$(2-41) \quad d\mu = d\mu(x, v, z) = \omega(z) dx dv dz.$$

With this measure, one has the corresponding Hilbert space with the following inner product and norms:

(2-42)

$$\langle f, g \rangle = \int_{\Omega} \int_{\mathbb{R}} \int_{I_z} fg d\mu(x, v, z), \quad \text{or,} \quad \langle \rho, j \rangle = \int_{\Omega} \int_{I_z} \rho j d\mu(x, z),$$

with norm

$$\|f\|^2 = \langle f, f \rangle.$$

In order to get the convergence rate of the solution to the global equilibrium, define

$$(2-43) \quad h = \frac{f - \mathfrak{m}}{\sqrt{\mathfrak{m}}}, \quad \sigma = \int_{\mathbb{R}} h \sqrt{M} dv, \quad u = \int_{\mathbb{R}} h v \sqrt{M} dv,$$

where  $h$  is the (microscopic) *fluctuation* around the equilibrium,  $\sigma$  is the (macroscopic) density fluctuation, and  $u$  is the (macroscopic) velocity fluctuation. Then the microscopic quantity  $h$  satisfies,

$$(2-44) \quad \varepsilon \delta \partial_t h + \beta v \partial_x h - \delta \partial_x \phi \partial_v h + \delta \frac{v}{2} \partial_x \phi h + \delta v \sqrt{M} \partial_x \phi = \mathfrak{L}^F h,$$

$$(2-45) \quad \partial_x^2 \phi = -\sigma,$$

while the macroscopic quantities  $\sigma$  and  $u$  satisfy

$$(2-46) \quad \delta \partial_t \sigma + \partial_x u = 0,$$

$$(2-47) \quad \varepsilon \delta \partial_t u + \varepsilon \partial_x \sigma + \varepsilon \int v^2 \sqrt{M} (1 - \Pi) \partial_x h dv + \delta \partial_x \phi \sigma + u + \delta \partial_x \phi = 0,$$

where  $\mathfrak{L}^F$  is the so-called linearized Fokker-Planck operator,

$$(2-48) \quad \mathfrak{L}^F h = \frac{1}{\sqrt{\mathfrak{m}}} \mathfrak{F} \left( \mathfrak{m} + \sqrt{\mathfrak{m}} h \right) = \frac{1}{\sqrt{\mathfrak{m}}} \partial_v \left( \mathfrak{m} \partial_v \left( \frac{h}{\sqrt{\mathfrak{m}}} \right) \right).$$

Introduce projection operator

$$(2-49) \quad \Pi h = \sigma \sqrt{\mathfrak{m}} + v u \sqrt{\mathfrak{m}}.$$

Furthermore, we also define the following norms and energies,

$$\begin{aligned} \|h\|_{L^2(v)}^2 &= \int_{\mathbb{R}} h^2 dv, \quad \|f\|_{H^m}^2 = \sum_{l=0}^m \|\partial_z^l f\|^2, \\ E_h^m &= \|h\|_{H^m}^2 + \|\partial_x h\|_{H^{m-1}}^2, \quad E_{\phi}^m = \|\partial_x \phi\|_{H^m}^2 + \|\partial_x^2 \phi\|_{H^{m-1}}^2. \end{aligned}$$

We need the following hypocoercivity properties proved in [Duan, Fornasier, and Toscani \[2010\]](#):

**Proposition 2.5.** *For  $\mathfrak{L}^F$  defined in (2-48),*

- (a)  $-\langle \mathfrak{L}^F h, h \rangle = -\langle L(1 - \Pi)h, (1 - \Pi)h \rangle + \|u\|^2;$
- (b)  $-\langle \mathfrak{L}^F (1 - \Pi)h, (1 - \Pi)h \rangle = \|\partial_v(1 - \Pi)h\|^2 + \frac{1}{4}\|v(1 - \Pi)h\|^2 - \frac{1}{2}\|(1 - \Pi)h\|^2;$
- (c)  $-\langle \mathfrak{L}^F (1 - \Pi)h, (1 - \Pi)h \rangle \geq \|(1 - \Pi)h\|^2;$
- (d) *There exists a constant  $\lambda_0 > 0$ , such that the following hypocoercivity holds,*

$$(2-50) \quad -\langle \mathfrak{L}^F h, h \rangle \geq \lambda_0 \|(1 - \Pi)h\|_v^2 + \|u\|^2,$$

*and the largest  $\lambda_0 = \frac{1}{7}$  in one dimension.*

The following results were obtained in [Jin and Y. Zhu \[n.d.\]](#).

**Theorem 2.6.** *For the high field regime ( $\delta = 1$ ), if*

$$(2-51) \quad E_h^m(0) + \frac{1}{\varepsilon^2} E_\phi^m(0) \leq \frac{C_0}{\varepsilon},$$

*then,*

$$(2-52) \quad E_h^m(t) \leq \frac{3}{\lambda_0} e^{-\frac{t}{\varepsilon^2}} \left( E_h^m(0) + \frac{1}{\varepsilon^2} E_\phi^m(0) \right), \quad E_\phi^m(t) \leq \frac{3}{\lambda_0} e^{-t} (\varepsilon^2 E_h^m(0) + E_\phi^m(0));$$

*For the parabolic regime ( $\delta = \varepsilon$ ), if*

$$(2-53) \quad E_h^m(0) + \frac{1}{\varepsilon^2} E_\phi^m(0) \leq \frac{C_0}{\varepsilon^2},$$

*then,*

$$(2-54) \quad E_h^m(t) \leq \frac{3}{\lambda_0} e^{-\frac{t}{\varepsilon}} \left( E_h^m(0) + \frac{1}{\varepsilon^2} E_\phi^m(0) \right), \quad E_\phi^m(t) \leq \frac{3}{\lambda_0} e^{-t} (\varepsilon^2 E_h^m(0) + E_\phi^m(0)).$$

Here  $C_0 = 2\lambda_0/(32BC_1^2\sqrt{\varepsilon})^2$ ,  $B = 48\sqrt{m} \binom{m}{[m/2]}$  is a constant only depending on  $m$ ,  $[m/2]$  is the smallest integer larger or equal to  $\frac{m}{2}$ , and  $C_1$  is the Sobolev constant in one dimension, and  $m \geq 1$ .

These results show that the solution will converge to the global Maxwellian  $\mathfrak{M}$ . Since  $\mathfrak{M}$  is independent of  $z$ , one sees that the impact of the randomness dies out exponentially in time, in both asymptotic regimes. One should also note the small initial data requirement:  $E_\phi^m = O(\varepsilon)$  for  $\delta = 1$ .

The above theorem also leads to the following regularity result for the solution to VPFP system:

**Theorem 2.7.** *Under the same condition given in [Theorem 2.6](#), for  $x \in [0, l]$ , one has*

$$(2-55) \quad \|f(t)\|_{H_x^m}^2 \leq \frac{3}{\lambda_0} E^m(0) + 2l^2,$$

where  $E^m(0) = E_h^m(0) + \frac{1}{\varepsilon^2} E_\phi^m$ .

This Theorem shows that the regularity of the initial data in the random space is preserved in time. Furthermore, the bound of the Sobolev norm of the solution is independent of the small parameter  $\varepsilon$ .

### 3 Stochastic Galerkin methods for random kinetic equations

In order to quantify the uncertainty of kinetic equation we will use polynomial chaos expansion based stochastic Galerkin (SG) method [Ghanem and Spanos \[1991\]](#) and [Xiu and Karniadakis \[2002\]](#). As is well known, the SG methods can achieve spectral accuracy if the solution has the regularity. This makes it very efficient if the dimension of the random space is not too high, compared with the classical Monte-Carlo method.

Due to its Galerkin formulation, mathematical analysis of the SG methods can be conducted more conveniently. Indeed many of the analytical methods well-established in kinetic theory can be easily adopted or extended to study the SG system of the random kinetic equations. For example, the study of regularity, and hypocoercivity based sensitivity analysis, as presented in [Section 2](#), can be used to analyze the SG methods. Furthermore, for multiscale kinetic equations, the SG methods allow one to extend the deterministic *Asymptotic-preserving* framework—a popular computational paradigm for multiscale kinetic and hyperbolic problems—to the random problem naturally. Finally, kinetic equations often contain small parameters such as the mean free path/time which asymptotically lead to macroscopic hyperbolic/diffusion equations. We are interested in developing the stochastic analogue of the asymptotic-preserving (AP) scheme, a scheme designed to capture the asymptotic limit at the discrete level. The SG method yields systems of deterministic equations that *resemble the deterministic kinetic equations*, although in vector forms. Thus it allows one to easily use the deterministic AP framework for the random problems, and allowing minimum “intrusion” to the legacy deterministic codes.

**3.1 The generalized polynomial chaos expansion based SG methods.** In the generalized polynomial chaos (gPC) expansion, one approximates the solution of a stochastic problem via an orthogonal polynomial series by seeking an expansion in the following form:

$$(3-1) \quad f(t, \mathbf{x}, v, z) \approx \sum_{|\mathbf{k}|=0}^{KM} f_{\mathbf{k}}(t, \mathbf{x}, v) \Phi_{\mathbf{k}}(z) := f^K(t, \mathbf{x}, v, z),$$

where  $\mathbf{k} = (k_1, \dots, k_n)$  is a multi-index with  $|\mathbf{k}| = k_1 + \dots + k_n$ .  $\{\Phi_{\mathbf{k}}(z)\}$  are from  $\mathbb{P}_K^n$ , the set of all  $n$ -variate polynomials of degree up to  $M$  and satisfy

$$\langle \Phi_{\mathbf{k}}, \Phi_{\mathbf{j}} \rangle_{\omega} = \int_{I_z} \Phi_{\mathbf{k}}(z) \Phi_{\mathbf{j}}(z) \omega(z) dz = \delta_{\mathbf{kj}}, \quad 0 \leq |\mathbf{k}|, |\mathbf{j}| \leq K.$$

Here  $\delta_{\mathbf{kj}}$  is the Kronecker delta function. The orthogonality with respect to  $\omega(z)$ , the probability density function of  $z$ , then defines the orthogonal polynomials. For example, the Gaussian distribution defines the Hermite polynomials; the uniform distribution defines the Legendre polynomials, etc.

Now inserting (3-1) into a general kinetic equation

$$(3-2) \quad \begin{cases} \partial_t f + v \cdot \nabla_{\mathbf{x}} f - \nabla_{\mathbf{x}} \phi \cdot \nabla_v f = \mathcal{Q}(f), & t > 0, \mathbf{x} \in \Omega, v \in \mathbb{R}^d, z \in I_z, \\ f(0, \mathbf{x}, v) = f^0(\mathbf{x}, v), & \mathbf{x} \in \Omega, v \in \mathbb{R}^d, z \in I_z. \end{cases}$$

Upon a standard Galerkin projection, one obtains for each  $0 \leq |\mathbf{k}| \leq M$ ,

$$(3-3) \quad \begin{cases} \partial_t f_{\mathbf{k}} + \mathbf{v} \cdot \nabla_{\mathbf{x}} f_{\mathbf{k}} - \sum_{|\mathbf{j}|=0}^K \nabla_{\mathbf{x}} \phi_{\mathbf{kj}} \cdot \nabla_v f_{\mathbf{j}} = \mathcal{Q}_{\mathbf{k}}(f^K), & t > 0, \mathbf{x} \in \Omega, \mathbf{v} \in \mathbb{R}^d, \\ f_{\mathbf{k}}(0, \mathbf{x}, \mathbf{v}) = f_{\mathbf{k}}^0(\mathbf{x}, \mathbf{v}), & \mathbf{x} \in \Omega, \mathbf{v} \in \mathbb{R}^d, \end{cases}$$

with

$$\begin{aligned} \mathcal{Q}_{\mathbf{k}}(f^K) &:= \int_{I_z} \mathcal{Q}(f^K)(t, \mathbf{x}, \mathbf{v}, z) \Phi_{\mathbf{k}}(z) \omega(z) dz \\ \phi_{\mathbf{kj}} &:= \int_{I_z} \phi(t, \mathbf{x}, z) \Phi_{\mathbf{k}}(z) \Phi_{\mathbf{j}}(z) \omega(z) dz \\ f_{\mathbf{k}}^0 &:= \int_{I_z} f^0(\mathbf{x}, \mathbf{v}, z) \Phi_{\mathbf{k}}(z) \omega(z) dz. \end{aligned}$$

We also assume that the potential  $\phi(t, \mathbf{x}, z)$  is given a priori for simplicity (the case that it is coupled to a Poisson equation can be treated similarly).

Therefore, one has a system of *deterministic* equations to solve and the unknowns are gPC coefficients  $f_{\mathbf{k}}$ , which are *independent of*  $z$ . Mostly importantly, the resulting SG system is just a vector analogue of its deterministic counterpart, thus allowing straightforward extension of the existing deterministic kinetic solvers. Once the coefficients  $f_{\mathbf{k}}$  are obtained through some numerical procedure, the statistical information such as the mean, covariance, standard deviation of the true solution  $f$  can be approximated as

$$\mathbb{E}[f] \approx f_0, \quad \text{Var}[f] \approx \sum_{|\mathbf{k}|=1}^K f_{\mathbf{k}}^2, \quad \text{Cov}[f] \approx \sum_{|\mathbf{i}|, |\mathbf{j}|=1}^K f_{\mathbf{i}} f_{\mathbf{j}}.$$

**3.2 Hypocoercivity estimate of the SG system.** The hypocoercivity theory presented in Section 2.2 can be used to study the properties of the SG methods. Here we take  $\phi = 0$ . Assume the random collision kernel has the assumptions given by (2-32). Consider the perturbative form

$$(3-4) \quad f_{\mathbf{k}} = \mathfrak{M} + \varepsilon M h_{\mathbf{k}},$$

where  $h_{\mathbf{k}}$  is the coefficient of the following gPC expansion

$$h(t, \mathbf{x}, v, z) \approx \sum_{|\mathbf{k}|=0}^M h_{\mathbf{k}}(t, \mathbf{x}, v) \Phi_{\mathbf{k}}(z) := h^K(t, \mathbf{x}, v, z).$$

Inserting ansatz (3-4) into (3-3) and conducting a standard Galerkin projection, one obtains the gPC-SG system for  $h_{\mathbf{k}}$ :

$$(3-5) \quad \begin{cases} \partial_t h_{\mathbf{k}} + \frac{1}{\varepsilon} v \cdot \nabla_x h_{\mathbf{k}} = \frac{1}{\varepsilon^2} \mathfrak{L}_{\mathbf{k}}(h^K) + \frac{1}{\varepsilon} \mathfrak{F}_{\mathbf{k}}(h^K, h^K), \\ h_{\mathbf{k}}(0, x, v) = h_{\mathbf{k}}^0(x, v), \quad x \in \Omega \subset \mathbb{T}^d, v \in \mathbb{R}^d, \end{cases}$$

for each  $1 \leq |\mathbf{k}| \leq K$ , with a periodic boundary condition and the initial data given by

$$h_{\mathbf{k}}^0 := \int_{I_z} h^0(x, v, z) \psi_{\mathbf{k}}(z) \pi(z) dz.$$

For the Boltzmann equation, the collision parts are given by

$$\begin{aligned} \mathfrak{L}_{\mathbf{k}}(h^K) &= \mathfrak{L}_{\mathbf{k}}^+(h^K) \\ &= \sum_{|\mathbf{i}|=1}^K \int_{\mathbb{R}^d \times \mathbb{S}^{d-1}} \widetilde{S}_{\mathbf{ki}} \phi(|v - v_*|) (h_{\mathbf{i}}(v') M(v'_*) + h_{\mathbf{i}}(v'_*) M(v')) M(v_*) dv_* d\sigma \\ &\quad - M(v) \sum_{|\mathbf{i}|=1}^K \int_{\mathbb{R}^d \times \mathbb{S}^{d-1}} \widetilde{S}_{\mathbf{ki}} \phi(|v - v_*|) h_{\mathbf{i}}(v_*) M(v_*) dv_* d\sigma - \sum_{|\mathbf{i}|=1}^K v_{\mathbf{ki}} h_{\mathbf{i}} \end{aligned}$$

$$\begin{aligned} \mathfrak{F}_{\mathbf{k}}(h^K, h^K)(t, x, v) &= \\ &= \sum_{|\mathbf{i}|, |\mathbf{j}|=1}^K \int_{\mathbb{R}^d \times \mathbb{S}^{d-1}} S_{\mathbf{kij}} \phi(|v - v_*|) M(v_*) (h_{\mathbf{i}}(v') h_{\mathbf{j}}(v'_*) - h_{\mathbf{i}}(v) h_{\mathbf{j}}(v_*)) dv_* d\sigma, \end{aligned}$$

with

$$\begin{aligned} \widetilde{S}_{\mathbf{ki}} &:= \int_{I_z} b(\cos \theta, z) \psi_{\mathbf{k}}(z) \psi_{\mathbf{i}}(z) \pi(z) dz \\ v_{\mathbf{ki}} &:= \int_{\mathbb{R}^d \times \mathbb{S}^{d-1}} \widetilde{S}_{\mathbf{ki}} \phi(|v - v_*|) \mathfrak{M}(v_*) dv_* d\sigma \\ S_{\mathbf{kij}} &:= \int_{I_z} b(\cos \theta, z) \psi_{\mathbf{k}}(z) \psi_{\mathbf{i}}(z) \psi_{\mathbf{j}}(z) \pi(z) dz \end{aligned}$$

For technical reasons, we assume  $z \in I_z$  is one dimensional and  $I_z$  has finite support  $|z| \leq C_z$  (which is the case, for example, for the uniform and Beta distribution). In [L. Liu and Jin \[2017\]](#) the following results are given:

**Theorem 3.1.** *Assume the collision kernel  $B$  satisfies (2-32) and is linear in  $z$ , with the form of*

$$(3-6) \quad b(\cos \theta, z) = b_0(\cos \theta) + b_1(\cos \theta)z,$$

with  $|\partial_z b| \leq O(\varepsilon)$ . We also assume the technical condition

$$(3-7) \quad \|\psi_k\|_{L^\infty} \leq Ck^p, \quad \forall k,$$

with a parameter  $p > 0$ . Let  $q > p + 2$ , define the energy  $E^K$  by

$$(3-8) \quad E^K(t) = E_{s,q}^K(t) = \sum_{k=1}^K \|k^q h_k\|_{H_{x,v}^s}^2,$$

with the initial data satisfying  $E^K(0) \leq \eta$ . Then for all  $s \geq s_0$ ,  $0 \leq \varepsilon_d \leq 1$ , such that for  $0 \leq \varepsilon \leq \varepsilon_d$ , if  $h^K$  is a gPC solution of (3-5) in  $H_{x,v}^s$ , we have the following:

(i) Under the incompressible Navier-Stokes scaling ( $\alpha = 1$ ),

$$E^K(t) \leq \eta e^{-\tau t}.$$

(ii) Under the acoustic scaling ( $\alpha = 0$ ),

$$E^K(t) \leq \eta e^{-\varepsilon \tau t},$$

where  $\eta, \tau$  are all positive constants that only depend on  $s$  and  $q$ , independent of  $K$  and  $z$ .

**Remark 3.2.** *The choice of energy  $E^K$  in (3-8) enables one to obtain the desired energy estimates with initial data independent of  $K$  [R. W. Shu and Jin \[2017\]](#).*

From here, one also concludes that,  $\|h^K\|_{H_{x,v}^s L_z^\infty}$  also decays exponentially in time, with the same rate as  $E^K(t)$ , namely

$$(3-9) \quad \|h^K\|_{H_{x,v}^s L_z^\infty} \leq \eta e^{-\tau t}$$

in the incompressible Navier-Stokes scaling, and

$$\|h^K\|_{H_{x,v}^s L_z^\infty} \leq \eta e^{-\varepsilon \tau t}$$

in the acoustic scaling.

For other kinetic models like the Landau equation, the proof is similar and we omit it here.

[L. Liu and Jin \[2017\]](#) also gives the following error estimates on the SG method for the uncertain Boltzmann equations.

**Theorem 3.3.** *Suppose the assumptions on the collision kernel and basis functions in [Theorem 3.1](#) are satisfied, and the initial data are the same in those in [Theorem 2.3](#), then*  
*(i) Under the incompressible Navier-Stokes scaling,*

$$(3-10) \quad \|h - h^K\|_{H_z^s} \leq C_e \frac{e^{-\lambda t}}{K^r},$$

*(ii) Under the acoustic scaling,*

$$(3-11) \quad \|h - h^K\|_{H_z^s} \leq C_e \frac{e^{-\varepsilon \lambda t}}{K^r},$$

with the constants  $C_e, \lambda > 0$  independent of  $K$  and  $\varepsilon$ .

The above results not only give the regularity of the SG solutions, which are the same as the initial data, but also show that the numerical fluctuation  $h^K$  converges with spectral accuracy to  $h$ , and the numerical error will also decay exponentially in time in the random space.

For more general solution (not the perturbative one given by (2-21) to the uncertain Boltzmann equation, one cannot obtain similar estimates. Specifically, for  $\alpha = 1$ , as  $\varepsilon \rightarrow 0$ , the moments of  $f$  is governed by the compressible Euler equations whose solution may develop shocks, thus the Sobolev norms used in this paper are not adequate. For  $\varepsilon = O(1)$ , [Hu and Jin \[2016\]](#) proved that, in the space homogeneous case, the regularity of the initial data in the random space is preserved in time. They also introduced a fast algorithm to compute the collision operator  $\mathbb{Q}_k$ . When the random variable is in higher dimension, sparse grids can be used, see [R. W. Shu, Hu, and Jin \[2017\]](#).



## 4 Stochastic asymptotic-preserving (sAP) schemes for multiscale random kinetic equations

When  $\varepsilon$  is small, numerically solving the kinetic equations is challenging since time and spatial discretizations need to resolve  $\varepsilon$ . Asymptotic-preserving (AP) schemes are those that mimic the asymptotic transitions from kinetic equations to their hydrodynamic/diffusion limits in the discrete setting [Jin \[1999, 2012\]](#). The AP strategy has been proved to be a powerful and robust technique to address multiscale problems in many kinetic problems. The main advantage of AP schemes is that they are very efficient even when  $\varepsilon$  is small, since they do not need to resolve the small scales numerically, and yet can still capture the macroscopic behavior governed by the limiting macroscopic equations. Indeed, it was proved, in the case of linear transport with a diffusive scaling, an AP scheme converges uniformly with respect to the scaling parameter [Golse, Jin, and Levermore \[1999\]](#). This is expected to be true for all AP schemes [Jin \[2012\]](#), although specific proofs are needed for specific problems. AP schemes avoid the difficulty of coupling a microscopic solver with a macroscopic one, as the micro solver *automatically* becomes a macro solver as  $\varepsilon \rightarrow 0$ . Interested readers may also consult earlier reviews in this subject [Jin \[2012\]](#), [Degond and Deluzet \[2017b\]](#), and [Hu, Jin, and Li \[2017\]](#).

Here we are interested in the scenario when the uncertainty (random inputs) and small scaling both present in a kinetic equation. Since the SG method makes the random kinetic equations into deterministic systems which are vector analogue of the original scalar deterministic kinetic equations, one can naturally utilize the deterministic AP machinery to solve the SG system to achieve the desired AP goals. To this aim, the notion of *stochastic asymptotic preserving (sAP)* was introduced in [Jin, Xiu, and X. Zhu \[2015\]](#). A scheme is sAP if an SG method for the random kinetic equation becomes an SG approximation for the limiting macroscopic, random (hydrodynamic or diffusion) equation as  $\varepsilon \rightarrow 0$ , with highest gPC degree, mesh size and time step all held fixed. Such schemes guarantee that even for  $\varepsilon \rightarrow 0$ , *all* numerical parameters, including the number of gPC modes, can be chosen only for accuracy requirement and *independent* of  $\varepsilon$ .

Next we use the linear transport [Equation \(2-1\)](#) as an example to derive an sAP scheme. It has the merit that rigorous convergence and sAP theory can be established, see [Jin, J.-G. Liu, and Ma \[2017\]](#).

**4.1 An sAP-SG method for the linear transport equation.** We assume the complete orthogonal polynomial basis in the Hilbert space  $H(I_z; \omega(z) dz)$  corresponding to the weight  $\omega(z)$  is  $\{\phi_i(z), i = 0, 1, \dots, \}$ , where  $\phi_i(z)$  is a polynomial of degree  $i$  and satisfies the orthonormal condition:

$$\langle \phi_i, \phi_j \rangle_\omega = \int \phi_i(z) \phi_j(z) \omega(z) dz = \delta_{ij}.$$

Here  $\phi_0(z) = 1$ , and  $\delta_{ij}$  is the Kronecker delta function. Since the solution  $f(t, \cdot, \cdot, \cdot)$  is defined in  $L^2([0, 1] \times [-1, 1] \times \mathbb{I}_z; d\mu)$ , one has the gPC expansion

$$f(t, x, v, z) = \sum_{i=0}^{\infty} f_i(t, x, v) \phi_i(z), \quad \hat{f} = (f_i)_{i=0}^{\infty} := (\bar{f}, \hat{f}_1).$$

The mean and variance of  $f$  can be obtained from the expansion coefficients as

$$\bar{f} = E(f) = \int_{I_z} f \omega(z) dz = f_0, \quad \text{var}(f) = |\hat{f}_1|^2.$$

Denote the SG solution by

$$(4-1) \quad f^K = \sum_{i=0}^K f_i \phi_i, \quad \hat{f}^K = (f_i)_{i=0}^M := (\bar{f}, \hat{f}_1^K),$$

from which one can extract the mean and variance of  $f^K$  from the expansion coefficients as

$$E(f^K) = \bar{f}, \quad \text{var}(f^K) = |\hat{f}_1^K|^2 \leq \text{var}(f).$$

Furthermore, we define

$$\sigma_{ij} = \langle \phi_i, \sigma \phi_j \rangle_{\omega}, \quad \Sigma = (\sigma_{ij})_{M+1, M+1}; \quad \sigma_{ij}^a = \langle \phi_i, \sigma^a \phi_j \rangle_{\omega}, \quad \Sigma^a = (\sigma_{ij}^a)_{M+1, M+1}$$

for  $0 \leq i, j \leq M$ . Let  $\text{Id}$  be the  $(M+1) \times (M+1)$  identity matrix.  $\Sigma, \Sigma^a$  are symmetric positive-definite matrices satisfying (Xiu [2010])

$$\Sigma \geq \sigma_{\min} \text{Id}.$$

If one applies the gPC ansatz (4-1) into the transport Equation (2-1), and conduct the Galerkin projection, one obtains

$$(4-2) \quad \varepsilon \partial_t \hat{f} + v \partial_x \hat{f} = -\frac{1}{\varepsilon} (I - [\cdot]) \Sigma \hat{f} - \varepsilon \Sigma^a \hat{f} - \hat{S},$$

where  $\hat{S}$  is defined similarly as (4-1).

We now use the micro-macro decomposition (Lemou and Mieussens [2008]):

$$(4-3) \quad \hat{f}(t, x, v, z) = \hat{\rho}(t, x, z) + \varepsilon \hat{g}(t, x, v, z),$$

where  $\hat{\rho} = [\hat{f}]$  and  $[\hat{g}] = 0$ , in (4-2) to get

$$(4-4a) \quad \partial_t \hat{\rho} + \partial_x [v \hat{g}] = -\Sigma^a \hat{\rho} + \hat{S},$$

$$(4-4b) \quad \partial_t \hat{g} + \frac{1}{\varepsilon} (I - [\cdot]) (v \partial_x \hat{g}) = -\frac{1}{\varepsilon^2} \Sigma \hat{g} - \Sigma^a \hat{g} - \frac{1}{\varepsilon^2} v \partial_x \hat{\rho},$$

with initial data

$$\hat{\rho}(0, x, z) = \hat{\rho}_0(x, z), \quad \hat{g}(0, x, v, z) = \hat{g}_0(x, v, z).$$

It is easy to see that system (4-4) formally has the diffusion limit as  $\varepsilon \rightarrow 0$ :

$$(4-5) \quad \partial_t \hat{\rho} = \partial_x (K \partial_x \hat{\rho}) - \Sigma^a \hat{\rho} + \hat{S},$$

where

$$(4-6) \quad K = \frac{1}{3} \Sigma^{-1}.$$

This is the sG approximation to the random diffusion Equation (2-8). Thus the gPC approximation is sAP in the sense of Jin, Xiu, and X. Zhu [2015].

Let  $f$  be the solution to the linear transport Equation (2-1)–(2-2). Use the  $K$ -th order projection operator  $P_M : P_K f = \sum_{i=0}^K f_i \phi_i(z)$ , the error arisen from the gPC-sG can be split into two parts  $r_K$  and  $e_K$ ,

$$(4-7) \quad f - f^K = f - P_K f + P_K f - f^K := r_K + e_K,$$

where  $r_K = f - P_K f$  is the projection error, and  $e_K = P_K M f - f^K$  is the SG error.

Here we summarize the results of Jin, J.-G. Liu, and Ma [2017].

**Lemma 4.1 (Projection error).** *Under all the assumption in Theorem 2.1 and Theorem 2.2, we have for  $t \in (0, T]$  and any integer  $k = 0, \dots, m$ ,*

$$(4-8) \quad \|r_K\|_{\Gamma} \leq \frac{C_1}{K^k}.$$

Moreover,

$$(4-9) \quad \|[r_K] - r_K\|_{\Gamma} \leq \frac{C_2}{K^k} \varepsilon,$$

where  $C_1$  and  $C_2$  are independent of  $\varepsilon$ .

**Lemma 4.2 (SG error).** *Under all the assumptions in Theorem 2.1 and Theorem 2.2, we have for  $t \in (0, T]$  and any integer  $k = 0, \dots, m$ ,*

$$(4-10) \quad \|e_M\|_{\Gamma} \leq \frac{C(T)}{M^k},$$

where  $C(T)$  is a constant independent of  $\varepsilon$ .

Combining the above lemmas gives the uniform (in  $\varepsilon$ ) convergence theorem:

**Theorem 4.3.** *If for some integer  $m \geq 0$ ,*

$$(4-11) \quad \|\sigma(z)\|_{H^k} \leq C_\sigma, \quad \|D^k f_0\|_\Gamma \leq C_0, \quad \|D^k(\partial_x f_0)\|_\Gamma \leq C_x, \quad k = 0, \dots, m,$$

*then the error of the sG method is*

$$(4-12) \quad \|f - f^K\|_\Gamma \leq \frac{C(T)}{K^k},$$

*where  $C(T)$  is a constant independent of  $\varepsilon$ .*

**Theorem 4.3** gives a uniformly in  $\varepsilon$  spectral convergence rate, thus one can choose  $K$  independent of  $\varepsilon$ , a very strong sAP property. Such a result is also obtained with the anisotropic scattering case, for the linear semiconductor Boltzmann equation ([Jin and L. Liu \[2017\]](#) and [L. Liu \[n.d.\]](#)).

**4.2 A full discretization.** By using the SG formulation, one obtains a vector version of the original deterministic transport equation. This enables one to use the deterministic AP methodology. Here, we adopt the micro-macro decomposition based AP scheme developed in [Lemou and Mieussens \[2008\]](#) for the gPC-sG system (4-4).

We take a uniform grid  $x_i = ih, i = 0, 1, \dots, N$ , where  $h = 1/N$  is the grid size, and time steps  $t^n = n\Delta t$ .  $\rho_i^n$  is the approximation of  $\rho$  at the grid point  $(x_i, t^n)$  while  $g_{i+\frac{1}{2}}^{n+1}$  is defined at a staggered grid  $x_{i+1/2} = (i + 1/2)h, i = 0, \dots, N - 1$ .

The fully discrete scheme for the gPC system (4-4) is

$$(4-13a) \quad \frac{\hat{\rho}_i^{n+1} - \hat{\rho}_i^n}{\Delta t} + \left[ v \frac{\hat{g}_{i+\frac{1}{2}}^{n+1} - \hat{g}_{i-\frac{1}{2}}^{n+1}}{\Delta x} \right] = -\Sigma_i^a \hat{\rho}_i^{n+1} + \hat{S}_i,$$

$$(4-13b) \quad \frac{\hat{g}_{i+\frac{1}{2}}^{n+1} - \hat{g}_{i+\frac{1}{2}}^n}{\Delta t} + \frac{1}{\varepsilon \Delta x} (I - [\cdot]) \left( v^+ (\hat{g}_{i+\frac{1}{2}}^n - \hat{g}_{i-\frac{1}{2}}^n) + v^- (\hat{g}_{i+\frac{3}{2}}^n - \hat{g}_{i+\frac{1}{2}}^n) \right) \\ = -\frac{1}{\varepsilon^2} \Sigma_i \hat{g}_{i+\frac{1}{2}}^{n+1} - \Sigma^a \hat{g}_{i+\frac{1}{2}}^{n+1} - \frac{1}{\varepsilon^2} v \frac{\hat{\rho}_{i+1}^n - \hat{\rho}_i^n}{\Delta x}.$$

It has the formal diffusion limit when  $\varepsilon \rightarrow 0$  given by

$$(4-14) \quad \frac{\hat{\rho}_i^{n+1} - \hat{\rho}_i^n}{\Delta t} - K \frac{\hat{\rho}_{i+1}^n - 2\hat{\rho}_i^n + \hat{\rho}_{i-1}^n}{\Delta x^2} = -\Sigma_i^a \hat{\rho}_i^{n+1} + \hat{S}_i,$$

where  $K = \frac{1}{3}\Sigma^{-1}$ . This is the fully discrete sG scheme for (4-5). Thus the fully discrete scheme is sAP.

One important property for an AP scheme is to have a stability condition independent of  $\varepsilon$ , so one can take  $\Delta t \gg O(\varepsilon)$ . The next theorem from [Jin, J.-G. Liu, and Ma \[2017\]](#) answers this question.

**Theorem 4.4.** *Assume  $\sigma^a = S = 0$ . If  $\Delta t$  satisfies the following CFL condition*

$$(4-15) \quad \Delta t \leq \frac{\sigma_{\min}}{3} \Delta x^2 + \frac{2\varepsilon}{3} \Delta x,$$

*then the sequences  $\hat{\rho}^n$  and  $\hat{g}^n$  defined by scheme (4-13) satisfy the energy estimate*

$$\Delta x \sum_{i=0}^{N-1} \left( (\hat{\rho}_i^n)^2 + \frac{\varepsilon^2}{2} \int_{-1}^1 \left( \hat{g}_{i+\frac{1}{2}}^n \right)^2 dv \right) \leq \Delta x \sum_{i=0}^{N-1} \left( (\hat{\rho}_i^0)^2 + \frac{\varepsilon^2}{2} \int_{-1}^1 \left( \hat{g}_{i+\frac{1}{2}}^0 \right)^2 dv \right)$$

*for every  $n$ , and hence the scheme (4-13) is stable.*

Since the right hand side of (4-15) has a lower bound when  $\varepsilon \rightarrow 0$  (and the lower bound being that of a stability condition of the discrete diffusion [Equation \(4-14\)](#)), the scheme is asymptotically stable and  $\Delta t$  remains finite even if  $\varepsilon \rightarrow 0$ .

A discontinuous Galerkin method based sAP scheme for the same problem was developed in [Chen, L. Liu, and Mu \[2017\]](#), where uniform stability and rigorous sAP property were also proven.

sAP schemes were also developed recently for other multiscale kinetic equations, for example the radiative heat transfer equations [Jin and Lu \[2017\]](#), and the disperse two-phase kinetic-fluid model [Jin and R. Shu \[2017\]](#).

**4.3 Numerical examples.** We now show one example from [Jin, J.-G. Liu, and Ma \[2017\]](#) to illustrate the sAP properties of the scheme. The random variable  $z$  is one-dimensional and obeys uniform distribution.

Consider the linear transport [Equation \(2-1\)](#) with  $\sigma^a = S = 0$  and random coefficient  $\sigma(z) = 2 + z$ , subject to zero initial condition  $f(0, x, v, z) = 0$  and boundary condition

$$f(t, 0, v, z) = 1, \quad v \geq 0; \quad f(t, 1, v, z) = 0, \quad v \leq 0.$$

When  $\varepsilon \rightarrow 0$ , the limiting random diffusion equation is

$$(4-16) \quad \partial_t \rho = \frac{1}{3\sigma(z)} \partial_{xx} \rho,$$

with initial and boundary conditions:

$$\rho(0, x, z) = 0, \quad \rho(t, 0, z) = 1, \quad \rho(t, 1, z) = 0.$$

The analytical solution for (4-16) with the given initial and boundary conditions is

$$(4-17) \quad \rho(t, x, z) = 1 - \operatorname{erf} \left( x / \sqrt{\frac{4}{3\sigma(z)} t} \right).$$

When  $\varepsilon$  is small, we use this as the reference solution, as it is accurate with an error of  $O(\varepsilon^2)$ . For other implementation details, see Jin, J.-G. Liu, and Ma [ibid.].

In Figure 1, we plot the errors in mean and standard deviation of the SG numerical solutions at  $t = 0.01$  with different gPC orders  $M$ . Three sets of results are included: solutions with  $\Delta x = 0.04$  (squares),  $\Delta x = 0.02$  (circles),  $\Delta x = 0.01$  (stars). We always use  $\Delta t = 0.0002/3$ . One observes that the errors become smaller with finer mesh. One can see that the solutions decay rapidly in  $M$  and then saturate where spatial discretization error dominates. It is then obvious that the errors due to gPC expansion can be neglected at order  $M = 4$  even for  $\varepsilon = 10^{-8}$ . From this simple example, we can see that using the properly designed sAP scheme, the time, spatial, and random domain discretizations can be chosen independently of the small parameter  $\varepsilon$ .

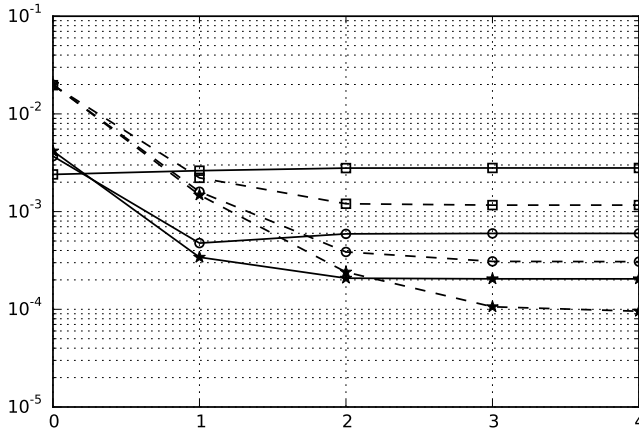


Figure 1: Errors of the mean (solid line) and standard deviation (dash line) of  $\rho$  with respect to the gPC order  $M$  at  $\varepsilon = 10^{-8}$ :  $\Delta x = 0.04$  (squares),  $\Delta x = 0.02$  (circles),  $\Delta x = 0.01$  (stars).  $\Delta t = 0.0002/3$ .

In Figure 2, we examine the difference between the solution at  $t = 0.01$  obtained by the 4th-order gPC method with  $\Delta x = 0.01$ ,  $\Delta t = \Delta x^2/12$  and the limiting analytical

solution (4-17). As expected, we observe the differences become smaller as  $\varepsilon$  is smaller in a quadratic fashion, before the numerical errors become dominant. This shows the sAP scheme works uniformly for different  $\varepsilon$ .

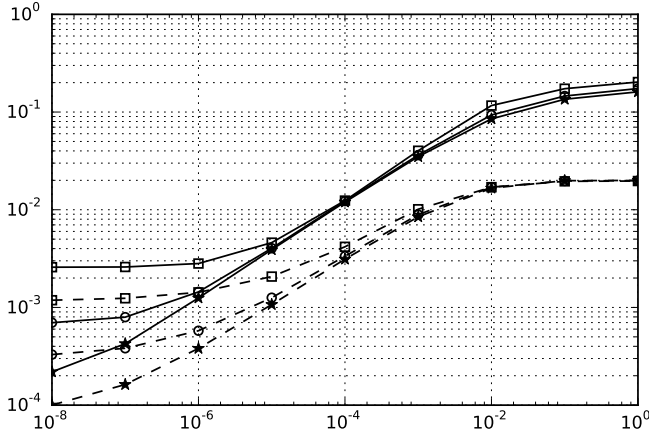


Figure 2: Differences in the mean (solid line) and standard deviation (dash line) of  $\rho$  with respect to  $\varepsilon^2$ , between the limiting analytical solution (4-17) and the 4th-order gPC solution with  $\Delta x = 0.04$  (squares),  $\Delta x = 0.02$  (circles) and  $\Delta x = 0.01$  (stars).

## 5 Conclusion and open problems

In this article we have presented some of our recent development of uncertainty quantification (UQ) for multiscale kinetic equations. The uncertainties for such equations typically come from collision/scattering kernels, boundary data, initial data, forcing terms, among others. Using hypocoercivity theory of kinetic operators, we proved the regularity, sensitivity, and long-time behavior in the random space in a general framework, and then adopted the generalized polynomial chaos based stochastic Galerkin (gPC-SG) method to handle the random inputs which can be proved spectrally accurate, under some regularity assumption on the initial data and random coefficients. When one needs to compute multiple scales, the SG method is constructed to possess the stochastic Asymptotic-Preserving (sAP) property, which allows all numerical parameters, including the gPC order, to be

chosen independently of the small parameter, hence is highly efficient when the scaling parameter, the Knudsen number, becomes small.

UQ for kinetic equations is a fairly recent research field, and many interesting problems remain open. We list a few such problems here:

- Whole space problem. Our hypocoercivity theory is developed for periodic spatial domain, which gives exponential decay towards the deterministic global Maxwellian. For the whole space problem, one cannot use the same abstract framework presented in subSection 2.2. For deterministic problems one can obtain only algebraic decay Duan and Strain [2011], Guo [2004], and Strain [2012]. It will be interesting to establish a corresponding theory for the uncertain Boltzmann equation.
- Boundary value problems. The uncertainty could also arise from boundary data. For the Maxwellian boundary condition, one can use the SG framework Hu and Jin [2016]. However, for small  $\varepsilon$ , the sensitivity analysis for random boundary input remain unexplored, even for the linear transport equation in the diffusive regime.
- Landau damping. While one can use hypocoercivity for collisional operator or Fokker-Planck operator, for Vlasov type equation (such as the Vlasov-Poisson equations) from collisionless plasma, the system does not have any dissipation, yet one still observes the asymptotical decay of a perturbation around a stationary homogeneous solution and the vanishing of electric field, a phenomenon called the Landau damping Landau [1946]. It will be interesting to invest the impact of uncertainty on Landau damping, although a rigorous nonlinear mathematical theory is very challenging Mouhot and Villani [2011].
- High dimensional random space. When the dimension of the random parameter  $z$  is moderate, sparse grids have been introduced R. W. Shu, Hu, and Jin [2017] and Hu, Jin, and R. Shu [n.d.] using wavelet approximations. Since wavelet basis does not have high order accuracy, it remains to construct sparse grids with high (or spectral) order of accuracy in the random space. When the random dimension is much higher, new methods need to be introduced to reduce the dimension.
- Study of sampling based methods such as collocation and multi-level Monte-Carlo methods. In practice, sampling based non-intrusive methods are attractive since they are based on the deterministic, or legacy codes. So far there has been no analysis done for the stochastic collocation methods for random kinetic equations. Moreover, multi-level Monte-Carlo method could significantly reduce the cost of sampling based methods Giles [2015]. Its application to kinetic equations with uncertainty remains to be investigated.



Despite at its infancy, due to the good regularity and asymptotic behavior in the random space for kinetic equations with uncertain random inputs, the UQ for kinetic equations is a promising research direction that calls for more development in their mathematical theory, efficient numerical methods, and applications. Moreover, since the random parameters in uncertain kinetic equations share some properties of the velocity variable for a kinetic equation, the ideas from kinetic theory can be very useful for UQ [Cho, Venturi, and Karniadakis \[2016\]](#), and vice versa, thus the marriage of the two fields can be very fruitful.

## References

- A. Arnold, J. A. Carrillo, I. Gamba, and C.-W. Shu (2001). “Low and high field scaling limits for the Vlasov- and Wigner-Poisson-Fokker-Planck systems”. *Transp. Theory Stat. Phys.* 30, pp. 121–153 (cit. on p. [3640](#)).
- Simon R Arridge (1999). “Optical tomography in medical imaging”. *Inverse problems* 15.2, R41 (cit. on p. [3630](#)).
- C. Bardos, R. Santos, and R. Sentis (1984). “[Diffusion approximation and computation of the critical size](#)”. *Trans. Amer. Math. Soc.* 284.2, pp. 617–649. MR: [743736](#) (cit. on p. [3632](#)).
- A. Bensoussan, J.-L. Lions, and G. C. Papanicolaou (1979). “[Boundary layers and homogenization of transport processes](#)”. *Publ. Res. Inst. Math. Sci.* 15.1, pp. 53–157. MR: [533346](#) (cit. on p. [3632](#)).
- P. R. Berman, J. E. M. Haverkort, and J. P. Woerdman (1986). “Collision kernels and transport coefficients”. *Phys. Rev. A* 34, pp. 4647–4656 (cit. on p. [3630](#)).
- G. A. Bird (1994). *Molecular Gas Dynamics and the Direct Simulation of Gas Flows*. Clarendon Press, Oxford (cit. on p. [3630](#)).
- Marc Briant (2015). “From the Boltzmann equation to the incompressible Navier–Stokes equations on the torus: A quantitative error estimate”. *Journal of Differential Equations* 259, pp. 6072–6141 (cit. on pp. [3636](#), [3637](#)).
- C. Cercignani (1988). *The Boltzmann Equation and Its Applications*. Springer-Verlag, New York (cit. on p. [3630](#)).
- (2000). *Rarefied Gas Dynamics: From Basic Concepts to Actual Calculations*. Cambridge University Press, Cambridge (cit. on p. [3630](#)).
- S. Chandrasekhar (1943). “[Stochastic problems in physics and astronomy](#)”. *Rev. Modern Phys.* 15, pp. 1–89 (cit. on p. [3640](#)).
- (1960). *Radiative Transfer*. Dover Publications (cit. on p. [3630](#)).
- S. Chapman and T. G. Cowling (1991). *The Mathematical Theory of Non-Uniform Gases*. third. Cambridge University Press, Cambridge (cit. on p. [3630](#)).

- Z. Chen, L. Liu, and L. Mu (2017). “DG-IMEX stochastic asymptotic-preserving schemes for linear transport equation with random inputs and diffusive scalings”. *J. Sci. Comput.* 73, pp. 566–592 (cit. on p. 3652).
- H. Cho, D. Venturi, and G. E. Karniadakis (2016). “Numerical methods for high-dimensional probability density function equations”. *J. Comput. Phys.* 305, pp. 817–837 (cit. on p. 3656).
- Pierre Degond and Fabrice Deluzet (2017a). “Asymptotic-Preserving methods and multiscale models for plasma physics”. *Journal of Computational Physics* 336, pp. 429–457 (cit. on p. 3630).
- (2017b). “Asymptotic-Preserving methods and multiscale models for plasma physics”. *Journal of Computational Physics* 336, pp. 429–457 (cit. on p. 3648).
- J. Dolbeault, C. Mouhot, and C. Schmeiser (2015). “Hypocoercivity for linear kinetic equations conserving mass”. *Trans. Amer. Math. Soc.* 367.6, pp. 3807–3828 (cit. on pp. 3633, 3634).
- R. Duan, M. Fornasier, and G. Toscani (2010). “A kinetic flocking model with diffusion”. *Commun. Math. Phys.* 300.1, pp. 95–145 (cit. on p. 3642).
- Renjun Duan and Robert M Strain (2011). “Optimal Time Decay of the Vlasov–Poisson–Boltzmann System in  $\mathbb{R}^3$ ”. *Archive for rational mechanics and analysis* 199.1, pp. 291–328 (cit. on p. 3655).
- RN Franklin and JR Ockendon (1970). “Asymptotic matching of plasma and sheath in an active low pressure discharge”. *Journal of plasma physics* 4.2, pp. 371–385 (cit. on p. 3630).
- R. G. Ghanem and P. D. Spanos (1991). *Stochastic Finite Elements: A Spectral Approach*. New York: Springer-Verlag (cit. on p. 3643).
- M. B. Giles (2015). “Multilevel Monte Carlo methods”. *Acta Numerica* 24, p. 259 (cit. on p. 3655).
- F. Golse, S. Jin, and C. D. Levermore (1999). “The convergence of numerical transfer schemes in diffusive regimes. I. Discrete-ordinate method”. *SIAM J. Numer. Anal.* 36.5, pp. 1333–1369. MR: 1706766 (cit. on p. 3648).
- T. Goudon, J. Nieto, F. Poupaud, and J. Soler (2005). “Multidimensional high-field limit of the electrostatic Vlasov–Poisson–Fokker–Planck system”. *J. Differ. Equ* 213.2, pp. 418–442 (cit. on p. 3640).
- Yan Guo (2004). “The Boltzmann equation in the whole space”. *Indiana University mathematics journal*, pp. 1081–1094 (cit. on p. 3655).
- (2006). “Boltzmann diffusive limit beyond the Navier-Stokes approximation”. *Communications on Pure and Applied Mathematics* 59.5, pp. 626–687 (cit. on p. 3635).
- J. Hu and S. Jin (2016). “A stochastic Galerkin method for the Boltzmann equation with uncertainty”. *J. Comput. Phys.* 315, pp. 150–168 (cit. on pp. 3647, 3655).

- J. Hu, S. Jin, and Q. Li (2017). “Asymptotic-preserving schemes for multiscale hyperbolic and kinetic equations”. In: *Handbook of Numerical Methods for Hyperbolic Problems*. Ed. by R. Abgrall and C.-W. Shu. Vol. 18. North-Holland. Chap. 5, pp. 103–129 (cit. on p. 3648).
- J. Hu, S. Jin, and R. Shu (n.d.). “A stochastic Galerkin method for the Fokker-Planck-Landau equation with random uncertainties”. To appear in *Proc. 16th Int’l Conf. on Hyperbolic Problems* (cit. on p. 3655).
- S. Jin (1999). “Efficient asymptotic-preserving (AP) schemes for some multiscale kinetic equations”. *SIAM J. Sci. Comput.* 21, pp. 441–454 (cit. on p. 3648).
- (2012). “Asymptotic preserving (AP) schemes for multiscale kinetic and hyperbolic equations: a review”. *Riv. Mat. Univ. Parma* 3, pp. 177–216 (cit. on p. 3648).
- S. Jin, J.-G. Liu, and Z. Ma (2017). “Uniform spectral convergence of the stochastic Galerkin method for the linear transport equations with random inputs in diffusive regime and a micro-macro decomposition based asymptotic preserving method”. *Research in Math. Sci.* 4 (15) (cit. on pp. 3633, 3648, 3650, 3652, 3653).
- S. Jin and L. Liu (2017). “An asymptotic-preserving stochastic Galerkin method for the semiconductor Boltzmann equation with random inputs and diffusive scalings”. *Multi-scale Model. Simul.* 15.1, pp. 157–183. MR: 3597157 (cit. on p. 3651).
- S. Jin and H. Lu (2017). “An asymptotic-preserving stochastic Galerkin method for the radiative heat transfer equations with random inputs and diffusive scalings”. *J. Comput. Phys.* 334, pp. 182–206 (cit. on p. 3652).
- S. Jin and R. Shu (2017). “A stochastic asymptotic-preserving scheme for a kinetic-fluid model for disperse two-phase flows with uncertainty”. *J. Comput. Phys.* 335, pp. 905–924 (cit. on p. 3652).
- S. Jin, D. Xiu, and X. Zhu (2015). “Asymptotic-preserving methods for hyperbolic and transport equations with random inputs and diffusive scalings”. *J. Comput. Phys.* 289, pp. 35–52 (cit. on pp. 3648, 3650).
- S. Jin and Y. Zhu (n.d.). “Hypocoercivity and Uniform Regularity for the Vlasov-Poisson-Fokker-Planck System with Uncertainty and Multiple Scales”. To appear in *SIAM J. Math. Anal.* (cit. on p. 3642).
- A. Jüngel (2009). *Transport Equations for Semiconductors*. Vol. 773. Lecture Notes in Physics. Berlin: Springer (cit. on p. 3630).
- K. Koura and H. Matsumoto (1991). “Variable soft sphere molecular model for inverse-power-law or Lennard-Jones potential”. *Phys. Fluids A* 3, pp. 2459–2465 (cit. on p. 3630).
- Lev Davidovich Landau (1946). “On the vibrations of the electronic plasma”. *Zh. Eksp. Teor. Fiz.* 10, p. 25 (cit. on p. 3655).
- E. W. Larsen and J. B. Keller (1974). “Asymptotic solution of neutron transport problems for small mean free paths”. *J. Math. Phys.* 15, pp. 75–81. MR: 0339741 (cit. on p. 3632).

- M. Lemou and L. Mieussens (2008). “A new asymptotic preserving scheme based on micro-macro formulation for linear kinetic equations in the diffusion limit”. *SIAM J. Sci. Comput.* 31.1, pp. 334–368. MR: [2460781](#) (cit. on pp. [3649](#), [3651](#)).
- Q. Li and L. Wang (n.d.). “Uniform regularity for linear kinetic equations with random input base d on hypocoercivity”. To appear in *SIAM/ASA J. UQ.* arXiv: [1612.01219](#) (cit. on p. [3633](#)).
- L. Liu (n.d.). “Uniform Spectral Convergence of the Stochastic Galerkin method for the Linear Semiconductor Boltzmann Equation with Random Inputs and Diffusive Scaling” (cit. on pp. [3633](#), [3651](#)).
- L. Liu and S. Jin (2017). “Hypocoercivity based Sensitivity Analysis and Spectral Convergence of the Stochastic Galerkin Approximation to Collisional Kinetic Equations with Multiple Scales and Random Inputs”. To appear in *Multiscale Model. Simult.* (cit. on pp. [3637](#), [3639](#), [3646](#), [3647](#)).
- M. Loève (1977). *Probability Theory*. fourth. Springer-Verlag, New York (cit. on p. [3631](#)).
- P. A. Markowich, C. Ringhofer, and C. Schmeiser (1990). *Semiconductor Equations*. New York: Springer Verlag Wien (cit. on p. [3630](#)).
- Clement Mouhot (2006). “Explicit coercivity estimates for the linearized Boltzmann and Landau operators”. *Comm. Partial Differential Equations* 31, 7–9, pp. 1321–1348 (cit. on p. [3639](#)).
- Clement Mouhot and Celine Baranger (2005). “Explicit spectral gap estimates for the linearized Boltzmann and Land au operators with hard potentials”. *Rev. Mat. Iberoamericana* 21, pp. 819–841 (cit. on p. [3639](#)).
- Clement Mouhot and Lukas Neumann (2006). “Quantitative perturbative study of convergence to equilibrium for col lisional kinetic models in the torus”. *Nonlinearity* 19, pp. 969–998 (cit. on pp. [3634](#), [3635](#)).
- Clément Mouhot and Cédric Villani (2011). “On landau damping”. *Acta mathematica* 207.1, pp. 29–201 (cit. on p. [3655](#)).
- G. Naldi, L. Pareschi, and G. Toscani, eds. (2010). *Mathematical Modeling of Collective Behavior in Socio-Economic and Life Sciences*. Birkhauser Basel (cit. on p. [3630](#)).
- J. Nieto, F. Poupaud, and J. Soler (2001). “High-field limit for the Vlasov-Poisson-Fokker-Planck system”. *Arch. Ration. Mech. Anal.* 158.1, pp. 29–59 (cit. on p. [3640](#)).
- F. Poupaud and J. Soler (2000). “Parabolic limit and stability of the Vlasov-Fokker-Planck system”. *Math. Models Methods Appl. Sci.* 10.7, pp. 1027–1045. MR: [1780148](#) (cit. on p. [3640](#)).
- T. Rivell (2006). “Notes on earth atmospheric entry for Mars sample return missions”. *NASA/TP-2006-213486* (cit. on p. [3630](#)).
- Esther S. Daus, Ansgar Jüngel, Clement Mouhot, and Nicola Zamponi (2016). “Hypocoercivity for a linearized multispecies Boltzmann system”. *SIAM J. MATH. ANAL.* 48, 1, pp. 538–568 (cit. on p. [3634](#)).

- R. W. Shu, J. Hu, and S. Jin (2017). “A Stochastic Galerkin Method for the Boltzmann Equation with multi-dimensional random inputs using sparse wavelet bases”. *Num. Math.: Theory, Methods and Applications (NMTMA)* 10, pp. 465–488 (cit. on pp. [3647](#), [3655](#)).
- R. W. Shu and S. Jin (2017). “Uniform regularity in the random space and spectral accuracy of the stochastic Galerkin method for a kinetic-fluid two-phase flow model with random initial inputs in the light particle regime”. To appear in *Math. Model Num. Anal.* (cit. on p. [3647](#)).
- Robert M Strain (2012). “OPTIMAL TIME DECAY OF THE NON CUT-OFF BOLTZMANN EQUATION IN THE WHOLE SPACE.” *Kinetic & Related Models* 5.3 (cit. on p. [3655](#)).
- Robert M Strain and Yan Guo (2008). “Exponential decay for soft potentials near Maxwellian”. *Archive for Rational Mechanics and Analysis* 187.2, pp. 287–339 (cit. on p. [3635](#)).
- C. Villani (2009). “Hypocoercivity”. *Mem. Amer. Math. Soc.* (Cit. on p. [3634](#)).
- D. Xiu (2010). *Numerical methods for stochastic computations*. Princeton, New Jersey: Princeton University Press (cit. on p. [3649](#)).
- D. Xiu and G. E. Karniadakis (2002). “The Wiener-Askey polynomial chaos for stochastic differential equations”. *SIAM J. Sci. Comput.* 24, pp. 619–644 (cit. on p. [3643](#)).

Received 2017-11-28.

SHI JIN  
DEPARTMENT OF MATHEMATICS  
UNIVERSITY OF WISCONSIN-MADISON  
MADISON, WI 53706  
USA

and

INSTITUTE OF NATURAL SCIENCES  
DEPARTMENT OF MATHEMATICS  
MOE-LSEC AND SHL-MAC  
SHANGHAI JIAO TONG UNIVERSITY  
SHANGHAI 200240  
CHINA

[jin@math.wisc.edu](mailto:jin@math.wisc.edu)  
[sjin@wisc.edu](mailto:sjin@wisc.edu)

# ON THE CONVERGENCE OF NUMERICAL SCHEMES FOR HYPERBOLIC SYSTEMS OF CONSERVATION LAWS

SIDDHARTHA MISHRA

## Abstract

A large variety of efficient numerical methods, of the finite volume, finite difference and DG type, have been developed for approximating hyperbolic systems of conservation laws. However, very few rigorous convergence results for these methods are available. We survey the state of the art on this crucial question of numerical analysis by summarizing classical results of convergence to entropy solutions for scalar conservation laws. Very recent results on convergence of ensemble Monte Carlo methods to the measure-valued and statistical solutions of multi-dimensional systems of conservation laws are also presented.

## 1 Introduction

Hyperbolic systems of conservation laws are nonlinear partial differential equations that arise in a large number of models in physics and engineering. These PDEs are of the generic form,

$$(1-1a) \quad \partial_t u + \nabla_x \cdot f(u) = 0$$

$$(1-1b) \quad u(x, 0) = \bar{u}(x).$$

Here, the unknown  $u = u(x, t) : \mathbb{R}^d \times \mathbb{R}_+ \rightarrow \mathbb{R}^N$  is the vector of *conserved variables* and  $f = (f^1, \dots, f^d) : \mathbb{R}^N \rightarrow \mathbb{R}^{N \times d}$  is the *flux function*. We denote  $\mathbb{R}_+ := [0, \infty)$ . The system is termed *hyperbolic* if the flux Jacobian matrix (along normal directions) has real eigenvalues [Dafermos \[2010\]](#).

Examples for (1-1a) include the compressible Euler equations of gas dynamics, the shallow water equations of oceanography, the magneto-hydrodynamics (MHD) equations of plasma physics and the equations governing nonlinear elastodynamics [Dafermos \[ibid.\]](#).

It is well known that solutions of (1-1) can form discontinuities such as *shock waves*, even for smooth initial data  $\bar{u}$ . Hence, solutions of systems of conservation laws (1-1) are sought in the sense of distributions. These *weak solutions* are not necessarily unique. They need to be augmented with additional admissibility criteria, often termed *entropy conditions*, to single out the physically relevant solution. *Entropy solutions* are widely regarded as the standard solution paradigm for systems of conservation laws Dafermos [2010].

Global well-posedness (existence, uniqueness and continuous dependence on initial data) of entropy solutions of scalar conservation laws ( $N = 1$  in (1-1)), was established in the pioneering work of Kružkov [1970]. For one-dimensional systems ( $d = 1$ ,  $N > 1$  in (1-1)), global existence, under the assumption of small initial total variation, was shown by Glimm in Glimm [1965] and by Bianchini and Bressan in Bianchini and Bressan [2005]. Uniqueness and stability of entropy solutions for one-dimensional systems has also been shown; see Bressan [2000] and references therein.

Although existence results have been obtained for some very specific examples of multi-dimensional systems (see Benzoni-Gavage and Serre [2007] and references therein), there are *no global well-posedness results* for any generic class of multi-dimensional systems. In fact, De Lellis, Székelyhidi et al. have recently been able to construct *infinitely many* entropy solutions for prototypical multi-dimensional systems such as the Euler equations for polytropic gas dynamics (see De Lellis and Székelyhidi [2009] and Chiodaroli, De Lellis, and Kreml [2015] and references therein).

It is not possible to obtain explicit solution formulas for (1-1), except in some very special cases. Consequently, numerical methods are necessary to simulate the solutions of systems of conservation laws. A large variety of efficient numerical methods have been developed over the last three to four decades to approximate solutions of (1-1). These include finite volume, conservative finite difference, discontinuous Galerkin finite element and spectral (viscosity) methods. Detailed accounts of these methods can be read from standard textbooks such as Godlewski and Raviart [1991], LeVeque [2002], Kröner [1997], and Toro [1999].

The fundamental question, in the context of numerical analysis of systems of conservation laws, is whether numerical schemes approximating (1-1) converge to an appropriate solution of (1-1) on mesh refinement? Surprisingly and in spite of the remarkable success of numerical methods in approximating solutions of systems of conservation laws, this fundamental question has only been answered in a few cases. The main objective to this article is to summarize some results on the convergence of numerical methods for (1-1).

The question of convergence has been answered fully in one particular context, namely the convergence of the so-called *monotone* finite volume (difference) schemes to entropy solutions of (multi-dimensional) scalar conservation laws Godlewski and Raviart [1991].

We begin by revisiting these classical results and showing that they can be generalized to some high-resolution schemes and to some arbitrarily high-order schemes, provided that the underlying reconstruction procedures satisfy a sign property and the numerical method is consistent with a discrete version of the entropy inequality.

On the other hand, the question of convergence of numerical methods for systems of conservation laws is largely unanswered, particularly for several space dimensions. Although it was widely believed that well designed numerical methods converge to entropy solutions, even for multi-dimensional systems of conservation laws, numerical experiments, such as those reported recently in [Fjordholm, Käppeli, Mishra, and Tadmor \[2017\]](#) and [Fjordholm, Mishra, and Tadmor \[2016\]](#), have revealed that even state of the art entropy stable numerical methods may not converge to any function as the mesh is refined. Rather, structures at finer and finer scales appear and impede convergence. Consequently, it was argued in [Fjordholm, Käppeli, Mishra, and Tadmor \[2017\]](#) and [Fjordholm, Mishra, and Tadmor \[2016\]](#) that entropy solutions may not be an appropriate paradigm to establish convergence of numerical methods approximating (1-1).

It was suggested in [Fjordholm, Käppeli, Mishra, and Tadmor \[2017\]](#) and [Fjordholm, Mishra, and Tadmor \[2016\]](#) that *entropy measure-valued solutions* are a promising candidate for an appropriate solution framework for systems of conservation laws. Measure-valued solutions are *Young measures* i.e., space-time parameterized probability measures, and were introduced by DiPerna in [DiPerna \[1985\]](#). In recent papers [Fjordholm, Käppeli, Mishra, and Tadmor \[2017\]](#) and [Fjordholm, Mishra, and Tadmor \[2016\]](#), the authors showed that a Monte-Carlo algorithm, based on underlying entropy stable finite volume schemes, converges to an entropy measure-valued solution of (1-1) on mesh refinement. This provided the first rigorous convergence result for numerical approximations of generic multi-dimensional systems of conservation laws.

Unfortunately, entropy measure-valued solutions are not necessarily unique as they lack information about multi-point correlations. More recently, a novel solution concept termed as statistical solutions has been introduced in [Fjordholm, Lanthaler, and Mishra \[2017\]](#). Statistical solutions are time-parameterized probability measures on  $L^p(\mathbb{R}^N)$  that are constrained in terms of an infinite family of equations evolving moments of the probability measure. The concept of statistical solutions amounts to providing information about all possible multi-point correlations for a measure-valued solution. In a forthcoming paper [Fjordholm, Lye, and Mishra \[2017b\]](#), the authors will show that under certain reasonable assumptions on the underlying numerical scheme, a Monte-Carlo algorithm, similar to the one proposed in [Fjordholm, Käppeli, Mishra, and Tadmor \[2017\]](#), converges in an appropriate topology to a statistical solution of (1-1),

We will survey all the afore-mentioned results in this article. We structure the rest of the paper as follows ; in [Section 2](#), we provide a brief introduction to numerical schemes



for (1-1) in one space dimension. The convergence to entropy solutions for scalar conservation laws is presented in Section 3. In sections 4 and 5, we present convergence of a Monte-Carlo algorithm to an entropy measure-valued solution and to a statistical solution, respectively.

## 2 Preliminaries

**2.1 Entropy solutions.** For simplicity of the notation and the exposition, we will focus on the one-dimensional version of (1-1) for the remainder of the paper. The conservation law reads as,

$$(2-1) \quad \begin{aligned} u_t + (f(u))_x &= 0, \quad (x, t) \in (\mathbb{R}, \mathbb{R}_+) \\ u(x, 0) &= \bar{u}(x), \quad x \in \mathbb{R} \end{aligned}$$

Here, the unknown  $u = u(x, t) : \mathbb{R} \times \mathbb{R}_+ \rightarrow \mathbb{R}^N$  is the vector of *conserved variables* and  $f : \mathbb{R}^N \rightarrow \mathbb{R}^N$  is the *flux function*.

**Definition 2.1.** A function  $u \in L^\infty(\mathbb{R} \times \mathbb{R}_+, \mathbb{R}^N)$  is a *weak solution* of (2-1) if it satisfies (2-1) in the sense of distributions:

$$(2-2) \quad \int_{\mathbb{R}_+} \int_{\mathbb{R}} \partial_t \varphi(x, t) u(x, t) + \partial_x \varphi(x, t) f(u(x, t)) dx dt + \int_{\mathbb{R}^d} \varphi(x, 0) u_0(x) dx = 0$$

for all test functions  $\varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+)$ . □

Weak solutions are not necessarily unique. We need to specify additional admissibility conditions in order to select physically meaningful weak solutions. These take the form of entropy conditions Dafermos [2010], given in terms of *entropy pairs*.

**Definition 2.2.** A pair of functions  $(\eta, q)$  with  $\eta : \mathbb{R}^N \rightarrow \mathbb{R}$ ,  $q : \mathbb{R}^N \rightarrow \mathbb{R}$  is called an *entropy pair* if  $\eta$  is convex and  $q$  satisfies the compatibility condition  $\nabla_u q(u) = \nabla_u \eta(u) \nabla_u f(u)$ , for all  $u \in \mathbb{R}^N$ . □

**Definition 2.3.** A weak solution  $u$  of (2-1) is an *entropy solution* if the entropy inequality

$$\partial_t \eta(u) + \partial_x q(u) \leq 0 \quad \text{in } \mathfrak{D}'(\mathbb{R} \times \mathbb{R}_+)$$

is satisfied for all entropy pairs  $(\eta, q)$ , that is, if

$$(2-3) \quad \int_{\mathbb{R}_+} \int_{\mathbb{R}} \partial_t \varphi(x, t) \eta(u(x, t)) + \partial_x \varphi(x, t) q(u(x, t)) dx dt + \int_{\mathbb{R}} \varphi(x, 0) \eta(u_0(x)) dx \geq 0$$

for all nonnegative test functions  $0 \leq \varphi \in C_c^1(\mathbb{R} \times \mathbb{R}_+)$ . □

**2.2 Numerical methods.** For simplicity, we fix a uniform mesh size  $\Delta x > 0$  and divide the computational domain into cells  $C_j := [x_{j-1/2}, x_{j+1/2}]$ ,  $\forall j \in \mathbb{Z}$ , with  $x_{j+1/2} = (j + 1/2)\Delta x$ . On this uniform grid, we will approximate (2-1) with the following sets of numerical schemes,

**2.2.1 Finite volume methods.** In this class of numerical methods, one approximates cell averages of the form:

$$(2-4) \quad U_j(t) = \frac{1}{\Delta x} \int_{C_j} u(x, t) dx$$

The cell averages satisfy a discrete form of the conservation law (2-1) resulting in the *semi-discrete* form of the finite volume method,

$$(2-5) \quad \frac{d}{dt} U_j(t) + \frac{1}{\Delta x} (F_{j+1/2}(t) - F_{j-1/2}(t)) = 0, \quad \forall j \in \mathbb{Z}.$$

Here, the *numerical flux function* is given by

$$F_{j+1/2}(t) := F(U_{j-r+1}, \dots, U_j, \dots, U_{j+r})$$

for some  $r \geq 1$ . The numerical flux function is assumed to be consistent i.e.,  $F(U, \dots, U) = f(U)$  and (locally) Lipschitz continuous in all its arguments [Godlewski and Raviart \[1991\]](#).

By now, there is an elaborate algorithmic procedure to determine suitable numerical fluxes in (2-5). A popular choice [LeVeque \[2002\]](#) is to set (suppressing the time dependence of all quantities),

$$(2-6) \quad F_{j+1/2} := F(U_{j+1/2}^-, U_{j+1/2}^+),$$

with  $U_{j+1/2}^\pm$  being the trace values at the point  $x = x_{j+1/2}$  of piecewise polynomial reconstructions of the cell averages  $U_j$ . To be more specific, we construct polynomials belonging to

$$(2-7) \quad \mathbb{P}_l^{\Delta x} := \{P^{\Delta x} \in L^\infty(\mathbb{R}) : P_j^{\Delta x} = P^{\Delta x}|_{C_j} \text{ is a polynomial of degree } l\},$$

for some integer  $l \geq 0$ . Then, we set

$$U_{j+1/2}^- = P_j^{\Delta x}(x_{j+1/2}), \quad U_{j+1/2}^+ = P_{j+1}^{\Delta x}(x_{j+1/2})$$

Different polynomial reconstruction procedures can be employed in this step. Popular choices include the use of TVD and TVB limiters [Godlewski and Raviart \[1991\]](#) and [LeVeque \[2002\]](#) which restrict the overall formal accuracy to second order. One can also use

the Essentially non-oscillatory (ENO) [Harten, Engquist, Osher, and Chakravarthy \[1987\]](#) and Weighted essentially non-oscillatory (WENO) procedures for obtaining arbitrary high-order of accuracy.

One can use a whole family of approximate Riemann solvers to calculate the numerical flux  $F$  in (2-6) [Toro \[1999\]](#). A different strategy would be to employ entropy stable (sum of entropy conservative and numerical diffusion) fluxes as advocated in [Fjordholm, Mishra, and Tadmor \[2012\]](#).

The resulting system of ODEs (2-5) is initialized with the cell averages,

$$U_j(0) := \frac{1}{\Delta x} \int_{C_j} \bar{u}(x) dx$$

A variant of this scheme is obtained by setting the initial values of (2-5) to point values i.e, requiring,  $U_j(0) = \bar{u}(x_j)$ . The resulting scheme is a form of conservative finite difference schemes [LeVeque \[2002\]](#).

**2.2.2 Discontinuous Galerkin Method.** The discontinuous Galerkin (DG) method [Cockburn and Shu \[1989\]](#) is a finite element method for discretizing (1-1) based on test spaces of piecewise polynomial functions. In one space dimension, the discontinuous Galerkin approximation of (2-1) consists of finding a function  $U^{\Delta x} \in \mathbb{P}_l^{\Delta x}$  that satisfies for every test function  $w \in \mathbb{P}_l^{\Delta x}$ , the following integral identity, (2-8)

$$\begin{aligned} \sum_j \int_{C_j} (\partial_t U^{\Delta x}(x, t) w(x) - f(U^{\Delta x}(x, t)) \partial_x w(x)) dx \\ + \sum_j \left( F(U^{\Delta x}(x_{j+1/2}^-), U^{\Delta x}(x_{j+1/2}^+)) w(x_{j+1/2}^-) \right. \\ \left. - F(U^{\Delta x}(x_{j-1/2}^-), U^{\Delta x}(x_{j-1/2}^+)) w(x_{j-1/2}^+) \right) = 0. \end{aligned}$$

Here  $w(x_{j\pm 1/2}^\pm)$  denotes taking left and right limits of a piecewise smooth function. The numerical flux  $F$  can be similar to the one considered in definition of the finite volume scheme (2-5) through (2-6).

**2.2.3 Time stepping.** The semi-discrete forms of the finite volume method (2-5) and the discontinuous Galerkin method (2-8) both result in a (large) non-linear system of ODEs. This system is usually solved using explicit high-order Runge-Kutta methods. A particularly attractive choice is that of *strong stability preserving* (SSP) Runge-Kutta methods [Gottlieb, Shu, and Tadmor \[2001\]](#) that retain the non-oscillatory properties of the spatial discretization. A less used but viable alternative is the family of *time discontinuous*

Galerkin methods [Johnson and Szepessy \[1987\]](#) and [Hiltebrand and Mishra \[2014\]](#) that are a finite element method in time. These methods are preferable for problems with multiple time scales.

**2.2.4 Multi-dimensional problems.** It is straightforward to extend (arbitrary high-order) finite volume methods like (2-5) for multi-dimensional problems (1-1) on domains that can be discretized with Cartesian or (block)-structured grids. On the other hands, it is very difficult to use such grids for domains with complex geometries. Unstructured grids, such as triangles in two dimensions and tetrahedra in three dimensions are more feasible for such domains. Although high-order finite volume schemes can be defined on such grids [Kröner \[1997\]](#), discontinuous Galerkin methods such as (2-8) are more suited for these class of problems [Cockburn and Shu \[1989\]](#).

### 3 Convergence to entropy solutions

**3.1 Scalar conservation laws.** We start with the case of scalar conservation laws in one space dimension i.e, the unknown  $u : \mathbb{R} \times \mathbb{R}_+ \rightarrow \mathbb{R}$  in (2-1). The convergence of numerical methods in scalar case has been well studied [Godlewski and Raviart \[1991\]](#). The most common approach for proving convergence of a numerical scheme, say a semi-discrete finite volume scheme such as (2-5), is to show that the approximate solutions, generated by the numerical scheme, are of finite total variation. Given the compact embedding of  $BV(\mathbb{R}^d)$  in  $L^1(\mathbb{R}^d)$ , one can show compactness for the approximations and establish pointwise convergence. The fact that the limit is a weak solution (2-2) and satisfies the entropy condition (2-3) can be verified using a Lax-Wendroff argument [Godlewski and Raviart \[ibid.\]](#). This approach works well for the convergence of *monotone schemes* for scalar conservation laws [Godlewski and Raviart \[ibid.\]](#). A variant of this argument can also be used to show that high-resolution schemes can be of finite total variation as in the theorem below,

**Theorem 3.1.** *Let  $\{U_j(t)\}_{j \in \mathbb{Z}}$  be approximations generated by a semi-discrete finite volume scheme of the form (2-5), with flux (2-6). Assume that*

(i.) *The numerical flux  $F$  in (2-5), (2-6) is monotone non-decreasing in its first argument and monotone non-increasing in the second argument, i.e,*

$$(3-1) \quad F(a_1, b) \leq F(a_2, b), \text{ if } a_1 \leq a_2, \quad \forall b \in \mathbb{R}$$

$$(3-2) \quad F(a, b_1) \geq F(a, b_2), \text{ if } b_1 \leq b_2, \quad \forall a \in \mathbb{R}.$$

(ii.) *The reconstruction satisfies the sign property i.e,*

$$(3-3) \quad \text{sign}(U_{j+1/2}^+(t) - U_{j+1/2}^-(t))\text{sign}(U_{j+1}(t) - U_j(t)) \geq 0, \quad \forall j.$$

(iii.) *The reconstruction clips local extrema i.e,*

$$(3-4) \quad U_{j+1/2}^-(t) = U_{j-1/2}^+(t) = U_j(t)$$

$$\text{if } \text{sign}(U_{j+1}(t) - U_j(t)) \neq \text{sign}(U_j(t) - U_{j-1}(t)).$$

Then, the scheme (2-5) is total variation diminishing (TVD) i.e,

$$(3-5) \quad \frac{d}{dt} \sum_j |U_{j+1}(t) - U_j(t)| \leq 0.$$

*Proof.* We suppress the time dependence of all quantities below for notational convenience. Multiplying both sides of (2-5) with  $(\text{sign}(U_j - U_{j-1}) - \text{sign}(U_j - U_{j+1}))$ , summing over  $j$  and arranging terms yields,

$$(3-6) \quad \Delta x \frac{d}{dt} \sum_j |U_{j+1}(t) - U_j(t)| = \sum_j [(\text{sign}(U_{j+1} - U_j) - \text{sign}(U_j - U_{j-1}))] \\ \left( F(U_{j+1/2}^-, U_{j+1/2}^+) - F(U_{j-1/2}^-, U_{j-1/2}^+) \right).$$

The only non-zero contributions to the sum on the right hand side is for those indices  $j \in \mathbb{Z}$  for which  $\text{sign}(U_{j+1}(t) - U_j(t)) \neq \text{sign}(U_j(t) - U_{j-1}(t))$ . For definiteness, we select an index  $j$  for which  $U_j \leq U_{j+1}$  and  $U_j \leq U_{j-1}$ . Expanding the flux difference, we obtain

$$F(U_{j+1/2}^-, U_{j+1/2}^+) - F(U_{j-1/2}^-, U_{j-1/2}^+) = \underbrace{F(U_{j+1/2}^-, U_{j+1/2}^+) - F(U_{j+1/2}^-, U_{j+1/2}^-)}_{T_1} \\ + \underbrace{F(U_{j+1/2}^-, U_{j+1/2}^-) - F(U_{j-1/2}^+, U_{j-1/2}^+)}_{T_2} \\ + \underbrace{F(U_{j-1/2}^+, U_{j-1/2}^+) - F(U_{j-1/2}^-, U_{j-1/2}^-)}_{T_3}$$

As  $U_j < U_{j+1}$ , the sign property (3-3) implies that  $U_{j+1/2}^- \leq U_{j+1/2}^+$ . Hence, by property (3-1) of the numerical flux  $F$ , we obtain that  $T_1 \leq 0$ . An identical argument shows that  $T_3 \leq 0$ . Given that the numerical flux is consistent, we see from the clipping at local extremum (3-4) that

$$T_2 = f(U_{j+1/2}^-) - f(U_{j-1/2}^+) = 0.$$

Hence, we obtain the summand in (3-6) corresponding to the index  $j$  is non-positive and this property holds for every  $j$  at which the summand is non-zero. This implies the TVD property (3-5).  $\square$

Once the TVD property is shown, one obtains pointwise a.e. convergence of the approximations generated by the semi-discrete scheme (2-5). A Lax-Wendroff theorem showing that the limit is a weak solution can be readily verified [Godlewski and Raviart \[1991\]](#).

The above [Theorem 3.1](#) illustrates some of the key requirements for a numerical scheme to converge. The flux needs certain monotonicity properties that are satisfied by several popular numerical fluxes for scalar conservation laws such as Godunov, Engquist-Osher, Rusanov and Lax-Friedrichs fluxes [LeVeque \[2002\]](#). On the other hand, the reconstruction procedure has to satisfy a sign property (3-3). This property was introduced independently in the context of the design of entropy stable schemes in [Fjordholm, Mishra, and Tadmor \[2012\]](#). Its role in providing a TVD bound on the scheme is novel. Among piecewise linear reconstruction procedures, the well-known *minmod* limiter enforces the sign property [Fjordholm, Mishra, and Tadmor \[2012\]](#) and [Fjordholm \[2013\]](#). On the other hand, other well-known limiters such as MC and Superbee do not satisfy this property. The ENO reconstruction procedure remarkably satisfies the sign property for all polynomial orders [Fjordholm, Mishra, and Tadmor \[2013\]](#).

However, the sign property does not suffice for a TVD bound. As expected, one has to switch off the reconstruction at local extrema (3-4). Hence and consistent with classical results of Harten [Godlewski and Raviart \[1991\]](#), one loses order of accuracy at local extrema for a TVD scheme to first order.

The preceding discussion shows that one requires some degree of numerical oscillations near local extrema in order to obtain a (formally) high-order accurate finite volume scheme. Thus, total variation bounds may not be an appropriate framework for showing convergence of arbitrary high-order schemes. Instead, compensated compactness techniques [Tartar \[1979\]](#) provide an alternative framework. A powerful illustration of this technique was presented in [Fjordholm \[2013\]](#) where the author showed that a class of arbitrary high-order entropy stable schemes converge to the entropy solution of a scalar conservation law with convex flux. We reproduce theorem 3.3 from reference [Fjordholm \[ibid.\]](#) below,

**Theorem 3.2.** [*Theorem 3.3 of [Fjordholm \[ibid.\]](#)*] *Let the flux  $f$  in (2-1) be strictly convex and the scalar conservation law be equipped with a strictly convex entropy function  $\eta$ . Assume that the approximations  $\{U_j(t)\}_{j \in \mathbb{Z}}$  generated by the semi-discrete finite volume scheme (2-5) for any  $\Delta x > 0$  and  $t \in [0, T]$  satisfies*

*i. A discrete entropy inequality of the form,*

$$(3-7) \quad \frac{d}{dt} \eta(U_j(t)) + \frac{1}{\Delta x} (Q_{j+1/2}(t) - Q_{j-1/2}(t)) \leq 0,$$

*for a numerical entropy flux  $Q_{j+1/2}(t) := Q(U_{j-r+1}, \dots, U_j, \dots, U_{j+r})$  for some  $r \geq 1$ . The numerical flux function is assumed to be consistent with the*

entropy flux  $q$  in (2-3) i.e,  $Q(U, \dots, U) = g(U)$  and (locally) Lipschitz continuous in all its arguments

ii.  $L^\infty$  bound i.e  $\exists M \in \mathbb{R}$

$$(3-8) \quad |U_j(t)| \leq M, \quad \forall t, \forall j \in \mathbb{Z}$$

iii. Compact support, i.e  $\exists J \in \mathbb{N}$  such that for all  $j$  with  $|j| \geq J$ ,  $U_j(t) \equiv 0$ .

iv. Weak BV bound i.e,

$$(3-9) \quad \int_0^T \sum_j |U_{j+1} - U_j(t)|^p dt \leq C, \quad \text{for some } p \in [2, \infty)$$

Then, define  $u^{\Delta x} \in L^1(\mathbb{R} \times \mathbb{R}_+)$  as  $u^{\Delta x}(x, t) = U_j(t)$  for all  $x \in C_j$ . The sequence of approximations  $u^{\Delta x}$  (upto a subsequence) converge point wise almost everywhere to a function  $u \in L^1(\mathbb{R} \times \mathbb{R}_+)$  as  $\Delta x \rightarrow 0$  and  $u$  is the unique entropy solution of (2-1).

The proof of this convergence theorem is based on the compensated compactness principle, more precisely on the Murat Lemma [Murat \[1981\]](#). It is easy to see that the weak BV property (3-9) is significantly weaker than the standard TVD bound and implies a rate of blow up for the total variation as  $\Delta x \rightarrow 0$ . The TeCNO schemes of [Fjordholm, Mishra, and Tadmor \[2012\]](#) have been shown to satisfy the discrete entropy inequality (3-7) and the weak BV bound (3-9), under an additional hypothesis [Fjordholm \[2013\]](#). Hence, they provide non-trivial examples of (formally) arbitrarily high-order schemes that converge to the entropy solution of the scalar version of (2-5) on mesh refinement. Although, this result is restricted to one-dimensional problems, it is conceivable that it can extended to several space dimensions using the technique of H-measures [Coclite, Mishra, and Risebro \[2010\]](#).

**3.2 One-dimensional systems.** Convergence to entropy solutions for (first-order) numerical schemes approximating one-dimensional  $2 \times 2$  systems i.e, (2-5) with  $N = 2$  can be established using variants of the compensated compactness technique [Ding, Chen, and Luo \[1989\]](#). For  $N \geq 3$  in (2-5), there are convergence results for the numerical methods such as the Glimm's random choice method [Glimm \[1965\]](#) or front tracking methods [Holden and Risebro \[2011\]](#), based on very delicate estimates on the approximations in BV. However, extending these techniques to standard first-order schemes such as the Godunov's scheme and the Lax-Friedrichs scheme is much harder. These convergence theorems require that the initial data should be of infinitesimally small total variation. Although not yet available, one expects that the techniques of [Bianchini and Bressan \[2005\]](#) etc can be modified to prove convergence of atleast first-order schemes to entropy solutions of the one-dimensional system (2-1).

**3.3 Multi-dimensional systems.** It was widely believed that well-designed numerical methods for approximating (1-1) in several space dimensions also converge to the entropy solution on mesh refinement. The lack of rigorous convergence proofs was blamed to the paucity of available theoretical tools. However, more recent investigations into the limit of popular numerical approximation frameworks for (1-1) have revealed some surprises in terms of a lack of convergence of numerical methods. A good example to illustrate this phenomenon was provided in [Fjordholm, Käppeli, Mishra, and Tadmor \[2017\]](#) and [Fjordholm, Mishra, and Tadmor \[2016\]](#) and we reproduce it here for the sake of completeness.

We consider the compressible Euler equations in two space dimensions,

$$(3-10) \quad \frac{\partial}{\partial t} \begin{pmatrix} \rho \\ \rho u \\ \rho v \\ E \end{pmatrix} + \frac{\partial}{\partial x_1} \begin{pmatrix} \rho u \\ \rho u^2 + p \\ \rho uv \\ (E + p)u \end{pmatrix} + \frac{\partial}{\partial x_2} \begin{pmatrix} \rho v \\ \rho uv \\ \rho v^2 + p \\ (E + p)v \end{pmatrix} = 0.$$

Here, the density  $\rho$ , velocity field  $(u, v)$ , pressure  $p$  and total energy  $E$  are related by the equation of state

$$E = \frac{p}{\gamma - 1} + \frac{\rho(u^2 + v^2)}{2},$$

with adiabatic constant  $\gamma = 1.4$ . We consider (3-10) in the computational domain  $x \in [0, 1]^2$  with periodic boundary conditions and with initial data:

$$(3-11) \quad p(x) = \begin{cases} 20 & \text{if } r < 0.1 \\ 1 & \text{otherwise,} \end{cases} \quad \rho(x) = \begin{cases} 2 & \text{if } r < I(\omega; x) \\ 1 & \text{otherwise,} \end{cases} \quad u = v = 0,$$

where  $r := |x - (0.5, 0.5)|$  denotes the distance to the center of the domain. The radial density interface  $I(\omega; x) = 0.25 + \varepsilon Y(\omega; \varphi(x))$  is perturbed with

$$(3-12) \quad Y(\omega; \varphi) = \sum_{n=1}^K a^n(\omega) \cos(\varphi + b^n(\omega)),$$

where  $\varphi(x) = \arccos((x_1 - 1/2)/r)$  and  $a_j^n = a_j^n(\omega) \in [0, 1]$  and  $b_j^n = b_j^n(\omega) \in [-\pi, \pi]$ ,  $i = 1, 2$ ,  $n = 1, \dots, K$  are uniformly randomly chosen numbers. The coefficients  $a_j^n$  have been normalized such that  $\sum_{n=1}^K a_j^n = 1$  to guarantee that  $|I_j(\omega; x) - J_j| \leq \varepsilon$  for  $j = 1, 2$ . We set  $K = 10$ . We fix  $\varepsilon = 10^{-3}$  and a single realization of random numbers  $a^n, b^n$ . The numerical scheme is a high-resolution finite volume scheme, based on an approximate Riemann solver of the HLLC type, a non-oscillatory MC limiter based piecewise linear reconstruction, in combination with a second-order, strong stability preserving Runge-Kutta time stepping routine. It is implemented within a massively parallel astrophysics code [Käppeli, Whitehouse, Scheidegger, Pen, and Liebendörfer \[2011\]](#).



This initial data is known as the Richtmeyer-Meshkov problem and consists of an initial (large) jump in the density and pressure across a slightly perturbed interface. We compute the solution at different mesh resolutions ranging from  $128^2$  to  $1024^2$ . The computed density at time  $t = 4$  is plotted in Figure 1. As shown in the figure, the solution is quite complex: the initial shock waves generated from the explosion have exited and reentered the domain (on account of periodic boundary conditions) and are interacting with an unstable interface. Furthermore, the reentered shock creates a complex pattern of small scale eddies on hitting the interface. These structures are formed at finer and finer scales as the mesh is refined.

The appearance of structures at finer and finer scales under mesh refinement may inhibit convergence of the scheme. To test this proposition, we compute the *Cauchy rates* i.e., the difference in the approximate solutions on two successive resolutions:

$$(3-13) \quad \mathfrak{E}^N = \|\rho^{2N} - \rho^N\|_{L^1([0,1]^2)}.$$

Here,  $N$  represents the number of mesh points in each direction. The results are shown in Figure 2 and demonstrate that the numerical approximation does not form a Cauchy sequence, let alone converge, as the mesh is refined. Similar results are also obtained with other  $L^p$  norms. This lack of convergence is not an artifact of the scheme discussed here; as reported in Fjordholm, Käppeli, Mishra, and Tadmor [2017], very similar results have been obtained with other state of the art schemes, such as the high-order TeCNO schemes Fjordholm, Mishra, and Tadmor [2012] and WENO schemes.

## 4 Convergence to measure-valued solutions

The last numerical example clearly indicated that one cannot expect that approximations, generated by state of the art numerical schemes, will converge to an entropy solution of the multi-dimensional system (1-1). An alternative solution paradigm is required in order to characterize the limits of numerical schemes for (1-1). A promising candidate for this paradigm is that of *entropy measure-valued solutions* of the system (1-1). We follow Fjordholm, Käppeli, Mishra, and Tadmor [2017] and Fjordholm, Mishra, and Tadmor [2016] and present a concise description of this solution concept in the following.

**4.1 Young measures.** Young measures were introduced in the context of PDEs by Tartar in Tartar [1979] in order to represent weak\* limits of  $L^\infty$  bounded sequences of highly oscillatory functions. A *Young measure* from  $D \subset \mathbb{R}^k$  to  $\mathbb{R}^N$  is a function which maps  $z \in D$  to a probability measure on  $\mathbb{R}^N$ . More precisely, a Young measure is a weak\* measurable map  $\nu : D \rightarrow \mathcal{P}(\mathbb{R}^N)$ , meaning that

the mapping  $z \mapsto \langle \nu_z, g \rangle$  is Borel measurable for every  $g \in C_0(\mathbb{R}^N)$ .

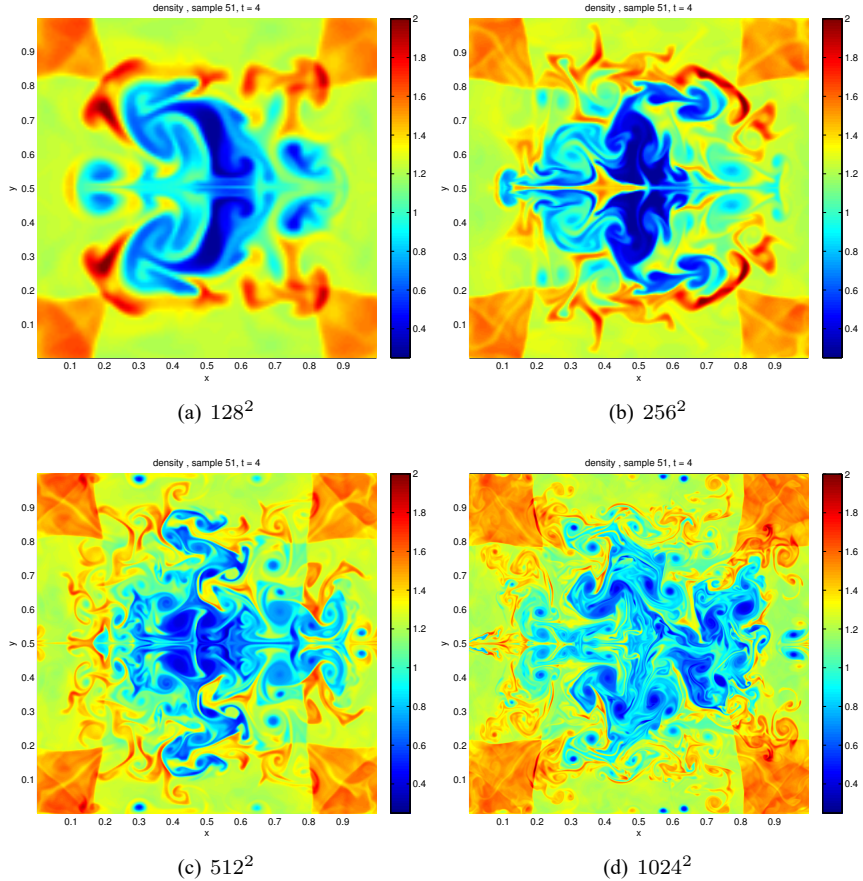


Figure 1: Approximate density at time  $t = 4$  for a single sample, computed with the high-resolution finite volume scheme of Käppeli, Whitehouse, Scheidegger, Pen, and Liebendörfer [2011], for the Richtmyer-Meshkov problem (3-11) for different grid resolutions.

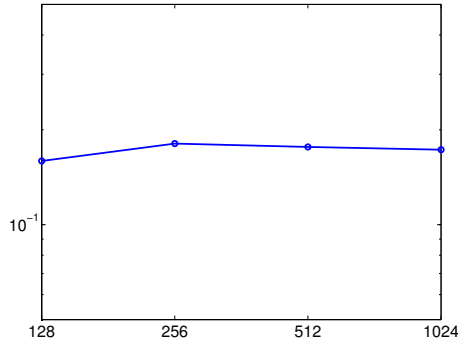


Figure 2: Cauchy rates (3-13) for the density (y-axis) in a single sample of the Richtmyer-Meshkov problem (3-11) at time  $t = 4$ , with respect to different grid resolutions (x-axis).

The set of all Young measures from  $D$  into  $\mathbb{R}^N$  is denoted by  $\mathbf{Y}(D, \mathbb{R}^N)$ .

The *fundamental theorem of Young measures* was first introduced by Tartar for  $L^\infty$ -bounded sequences Tartar [1979] and then generalized by Schonbek [1982] and Ball [1989] for sequences of measurable functions. A further generalization was presented in a recent paper Fjordholm, Käppeli, Mishra, and Tadmor [2017]: every sequence  $\nu^n \in \mathbf{Y}(D, \mathbb{R}^N)$  which does not “leak mass at infinity” has a weak\* convergent subsequence in the following sense (see theorem 3.3 of Fjordholm, Mishra, and Tadmor [2016]):

**Theorem 4.1.** [Theorem 3.3 of Fjordholm, Mishra, and Tadmor [ibid.]] Let  $\nu^n \in \mathbf{Y}(D, \mathbb{R}^N)$  for  $n \in \mathbb{N}$  be a sequence of Young measures. Then there exists a subsequence  $\nu^m$  which converges weak\* to a nonnegative measure-valued function  $\nu : D \rightarrow \mathcal{M}_+(\mathbb{R}^N)$  in the sense that

- (i)  $\langle \nu_z^m, g \rangle \xrightarrow{*} \langle \nu, g \rangle$  in  $L^\infty(D)$  for all  $g \in C_0(\mathbb{R}^N)$ ,
- (ii) Suppose further there is a nonnegative function  $\kappa \in C(\mathbb{R}^N)$  with  $\lim_{|\xi| \rightarrow \infty} \kappa(\xi) = \infty$  such that

$$(4-1) \quad \sup_n \int_D \langle \nu_z^n, \kappa \rangle dz < \infty.$$

Then  $\|\nu_z\|_{\mathcal{M}(\mathbb{R}^N)} = 1$  for a.e.  $z \in D$ ,

whence  $\nu \in \mathbf{Y}(D, \mathbb{R}^N)$

**4.2 Measure-valued solutions.** As mentioned earlier, entropy measure-valued solutions for nonlinear systems of conservation laws were introduced by DiPerna in DiPerna

[1985]. Here, we follow the presentation of a recent paper [Fjordholm, Käppeli, Mishra, and Tadmor \[2017\]](#).

**Definition 4.2.** Let  $\sigma \in \mathbf{Y}(\mathbb{R}^d, \mathbb{R}^N)$  be uniformly bounded given initial data. A family of Young measures  $\nu_t \in \mathbf{Y}(\mathbb{R}^d, \mathbb{R}^N)$  is a *measure-valued solution* (MV solution) of (1-1a) with data  $\sigma$  if

$$(4-2) \quad \int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \left( \langle \nu_{x,t}, \xi \rangle \partial_t \varphi + \langle \nu_{x,t}, f(\xi) \rangle \cdot \nabla \varphi \right) dx dt + \int_{\mathbb{R}^d} \varphi(x, 0) \langle \sigma_x, \xi \rangle dx = 0$$

for all  $\varphi \in C_c^\infty(\Omega)$ . □

Note that we allow for uncertainty in the initial data by considering a general initial Young measure  $\sigma$ , rather than restricting attention to atomic initial data  $\sigma = \delta_{u_0}$ .

As in the case of weak solutions, we need to impose additional admissibility criteria to enforce uniqueness of the measure-valued solution (4-2). This brings us to the following entropy inequalities.

**Definition 4.3.** A measure-valued solution  $\nu$  is an *entropy measure-valued (EMV) solution* of (1-1a) if  $\nu$  satisfies the following entropy inequality for all entropy pairs  $(\eta, q)$ :

$$(4-3) \quad \int_{\mathbb{R}_+} \int_{\mathbb{R}^d} \left( \langle \nu_{x,t}, \eta(\xi) \rangle \partial_t \varphi(x, t) + \langle \nu_{x,t}, q(\xi) \rangle \cdot \nabla_x \varphi(x, t) \right) dx dt + \\ + \int_{\mathbb{R}^d} \varphi(x, 0) \langle \sigma_x, \eta \rangle dx \geq 0$$

for all nonnegative test functions  $0 \leq \varphi \in C_c^1(\mathbb{R}^d \times \mathbb{R}_+)$ . □

**4.3 The FKMT algorithm for computing entropy measure-valued solutions.** Although the concept of entropy measure-valued solutions was introduced by DiPerna, the computation (algorithmic realization) of measure-valued solutions is quite challenging. In a recent paper [Fjordholm, Käppeli, Mishra, and Tadmor \[ibid.\]](#), the authors designed the following Monte-Carlo ensemble algorithm, termed here as the FKMT algorithm, to approximate entropy-measure valued solutions. For notational simplicity, we restrict the following discussion to the one dimensional case (2-1).

**Algorithm 4.4.** Let  $\Delta = \Delta x$  denote the grid size parameter and let  $M \in \mathbb{N}$ . Let  $\sigma \in \mathbf{Y}(\mathbb{R}^d, \mathbb{R}^N)$  be the initial Young measure.

**Step 1:** For some probability space  $(\Omega, \mathcal{X}, \mathbb{P})$ , draw  $M$  independent and identically distributed random fields  $u_0^{\Delta, 1}, \dots, u_0^{\Delta, M} : \Omega \times \mathbb{R}^d \rightarrow \mathbb{R}^N$ , all with the same probability law  $\sigma$ .

**Step 2:** For each  $1 \leq k \leq M$  and for a fixed  $\omega \in \Omega$ , use the finite volume scheme (2-5) to numerically approximate the conservation law (2-1) with initial data  $u_0^{\Delta,k}(\omega)$ . Denote  $u^{\Delta,k}(\omega; \cdot, t) = \mathbf{S}_t^\Delta u_0^{\Delta,k}(\omega; \cdot)$ . with  $\mathbf{S}_t^\Delta$  being the data to solution operator associated with the scheme (2-5).

**Step 3:** Define the approximate measure-valued solution,

$$(4-4) \quad \nu_{x,t}^{\Delta,M} := \frac{1}{M} \sum_{k=1}^M \delta_{u^{\Delta,k}(\omega;x,t)}.$$

□

In Fjordholm, Käppeli, Mishra, and Tadmor [2017], this ensemble Monte-Carlo algorithm was shown to converge to an entropy-measure valued solution of (2-1) in the following theorem,

**Theorem 4.5.** [Theorem 6 of Fjordholm, Käppeli, Mishra, and Tadmor [ibid.]] Denote  $U_j^k(t)$  as the approximate solution generated at time  $t$  by the semi-discrete finite volume scheme (2-5) for initial data  $u_0^k(\omega)$  for  $1 \leq k \leq M$ , corresponding to the  $k$ -th sample of the Algorithm 4.4. Assume that the numerical approximations satisfy the following,

i. Uniform  $L^\infty$  bound: i.e  $\exists M \in \mathbb{R}$

$$(4-5) \quad |U_j^k(t)| \leq M, \quad \forall t, \forall j \in \mathbb{Z}, \quad \forall 1 \leq k \leq M, \quad a.e \omega \in \Omega$$

ii. A discrete entropy inequality of the form,

$$(4-6) \quad \frac{d}{dt} \eta(U_j^k(t)) + \frac{1}{\Delta x} \left( Q_{j+1/2}^k(t) - Q_{j-1/2}^k(t) \right) \leq 0, \quad \forall 1 \leq k \leq M, \quad a.e \omega \in \Omega$$

for a numerical entropy flux  $Q_{j+1/2}^k(t) := Q(U_{j-r+1}^k, \dots, U_j^k, \dots, U_{j+r}^k)$  for some  $r \geq 1$ . The numerical flux function is assumed to be consistent with the entropy flux  $q$  in (2-3) i.e,  $Q(U, \dots, U) = q(U)$  and (locally) Lipschitz continuous in all its arguments

iii. Weak BV bound i.e,

$$(4-7) \quad \int_0^T \sum_j |U_{j+1}^k(t) - U_j^k(t)|^p dt \leq C, \quad \text{for some } p \in [2, \infty),$$

for all  $1 \leq k \leq M$  and a.e  $\omega \in \Omega$ .

Then, the approximate measure-valued solutions  $v_{x,t}^{\Delta,M}$ , generated by [Algorithm 4.4](#), converges weak-\* (in the sense of [Theorem 4.1](#)) as  $(\Delta, M) \rightarrow (0, \infty)$ , upto a subsequence, to a Young measure  $v_{x,t}$ , which is an entropy measure-valued solution of (2-1).

**Remark 4.6.** The TeCNO schemes of [Fjordholm, Mishra, and Tadmor \[2012\]](#) and the space-time DG schemes of [Hiltebrand and Mishra \[2014\]](#) are shown to verify the discrete entropy inequality (4-6) and the weak BV-bound (4-7), even for (formally) arbitrary orders of accuracy. The  $L^\infty$  bound is a technical assumption that can be relaxed by introducing the concept of *generalized measure-valued solutions*, that also account for possible concentrations, as in [Fjordholm, Mishra, and Tadmor \[2016\]](#). In that case, the discrete entropy inequality provides bounds in  $L^p$  and this suffices to show convergence of Algorithm 4.4.  $\square$

**Remark 4.7.** One has to treat the case of Dirac initial Young measure i.e.,  $\sigma_x = \delta_{\bar{u}(x)}$  for some  $\bar{u} \in L^1(\mathbb{R}^d)$  by using a perturbation of the Monte Carlo algorithm, see Algorithm 4.3 of [Fjordholm, Mishra, and Tadmor \[ibid.\]](#).  $\square$

**Remark 4.8.** The extension of the [Algorithm 4.4](#) and the convergence [Theorem 4.5](#) to several space dimensions is straightforward, see Theorem 7 of [Fjordholm, Käppeli, Mishra, and Tadmor \[2017\]](#).  $\square$

The weak-\* convergence for Young measures amounts to requiring that *one-point statistical quantities of interest* converge as the mesh is refined. In particular, mean, variance and one-point probability density functions (pdfs) are shown to converge. Numerical experiments reported in [Fjordholm, Käppeli, Mishra, and Tadmor \[2017\]](#) and [Fjordholm, Mishra, and Tadmor \[2016\]](#) illustrate this convergence rather well. Another interesting numerical observation from [Fjordholm, Käppeli, Mishra, and Tadmor \[2017\]](#) (see figure 18) is the realization that even if the initial data is a  $L^1$  function i.e., a Young measure concentrated on a single atom, the computed measure-valued solution may not be atomic. This further reinforces the contention that one cannot search for limits of numerical methods approximating (1-1) within the class of functions but has to rely on weaker notions of solutions, such as measure-valued solutions.

However, entropy measure-valued solutions are not necessarily unique. This is even true for the simplest case of a one-dimensional Burgers' equation provided that the initial data is a non-atomic Young measure, see example 9.1 of [Fjordholm, Mishra, and Tadmor \[2016\]](#) and several other counter-examples presented in [Schochet \[1989\]](#). On the other hand, the entropy measure-valued solution, computed by [Algorithm 4.4](#) has been observed to stable in numerical experiments, see [Fjordholm, Käppeli, Mishra, and Tadmor \[2017\]](#) and [Fjordholm, Mishra, and Tadmor \[2016\]](#). This stability holds with respect to perturbations of the initial Young measure data and with respect to the variation of the underlying numerical scheme. Clearly, the numerical algorithm provides a selection

principle that chooses a particular entropy measure-valued solution that is stable. Can we further constrain entropy measure-valued solutions to characterize this stable solution ?

## 5 Convergence to Statistical Solutions

A possible answer to the above question was proposed in a recent paper [Fjordholm, Lanthaler, and Mishra \[2017\]](#). Therein, the authors identified the principal reason for the non-uniqueness of measure-valued solutions being the lack of information about (multi-point) spatial correlations that is intrinsic to the notion of Young measures where only one-point statistical quantities are constrained. The authors of [Fjordholm, Lanthaler, and Mishra \[ibid.\]](#) proceeded to add further information to measure-valued solutions by specifying multi-point correlations in a systematic manner as described below.

**5.1 Correlation measures and Statistical solutions.** We follow [Fjordholm, Lanthaler, and Mishra \[ibid.\]](#) and consider the framework of statistical solutions. As mentioned in the introduction, we are interested in the situation where instead of an initial data  $\bar{u} \in L^1(\mathbb{R}^d)$  for (1-1), we are given some  $\bar{\mu} \in \mathcal{P}(L^p(\mathbb{R}^d))$  for some  $p \in [1, \infty)$ , that is, a probability distribution over different initial data  $\bar{u} \in L^p(\mathbb{R}^d)$ . A statistical solution of this initial value problem is a map  $t \mapsto \mu_t \in \mathcal{P}(L^p(\mathbb{R}^d))$  which satisfies the PDE (1-1) in a certain sense. In [Fjordholm, Lanthaler, and Mishra \[ibid.\]](#), the authors showed that any probability measure  $\mu \in \mathcal{P}(L^p(\mathbb{R}^d))$  can be described equivalently as a *correlation measure*—a hierarchy  $\mathbf{v} = (v^1, v^2, \dots)$  in which each element  $v^k$  provides the joint probability distribution  $v^k_{x_1, \dots, x_k}$  of the solution values  $u(x_1), \dots, u(x_k)$  at any choice of spatial points  $x_1, \dots, x_k \in \mathbb{R}^d$ . To be more precise, a correlation measure is defined as

**Definition 5.1.** Let  $d, N \in \mathbb{N}$ , let  $q \in [1, \infty)$ , let  $D \subset \mathbb{R}^d$  be an open set (the “spatial domain”) and (for notational convenience) denote  $\mathcal{U} = \mathbb{R}^N$  (“phase space”). A *correlation measure from  $D$  to  $\mathcal{U}$*  is a collection  $\mathbf{v} = (v^1, v^2, \dots)$  of maps satisfying for every  $k \in \mathbb{N}$ :

- (i)  $v^k$  is a Young measure from  $D^k$  to  $\mathcal{U}^k$ .
- (ii) *Symmetry:* if  $\sigma$  is a permutation of  $\{1, \dots, k\}$  and  $f \in C_0(\mathcal{U}^k)$  then  $\langle v^k_{\sigma(x)}, f(\sigma(\xi)) \rangle = \langle v^k_x, f(\xi) \rangle$  for a.e.  $x \in D^k$ .
- (iii) *Consistency:* If  $f \in C_b(\mathcal{U}^k)$  is of the form  $f(\xi_1, \dots, \xi_k) = g(\xi_1, \dots, \xi_{k-1})$  for some  $g \in C_0(\mathcal{U}^{k-1})$ , then  $\langle v^k_{x_1, \dots, x_k}, f \rangle = \langle v^{k-1}_{x_1, \dots, x_{k-1}}, g \rangle$  for almost every  $(x_1, \dots, x_k) \in D^k$ .

(iv)  $L^q$  integrability:

$$(5-1) \quad \int_D \langle v_x^1, |\xi|^q \rangle dx < \infty.$$

(v) *Diagonal continuity (DC)*:  $\lim_{\varepsilon \rightarrow 0} d_\varepsilon^q(v^2) = 0$ , where

$$(5-2) \quad d_\varepsilon^q(v^2) := \left( \int_D \int_{B_\varepsilon(x)} \langle v_{x,y}^2, |\xi_1 - \xi_2|^q \rangle dy dx \right)^{1/q}.$$

(Here,  $f_B = \frac{1}{|B|} \int_B$ , the average over  $B$ .)

□

It was shown in [Fjordholm, Lanthaler, and Mishra \[ibid.\]](#) that every probability measure  $\mu \in \mathcal{P}(L^q(D; \mathcal{U}))$  is dual to a unique correlation measure  $\mathbf{v}$ , and *vice versa*. Using this duality, we can now state the definition of statistical solutions.

**Definition 5.2.** We say that a weak\*-measurable map  $t \mapsto \mu_t \in \mathcal{P}(L^p(D))$ , for some  $p \in [1, \infty)$ , with corresponding spatial correlation measures  $\mathbf{v}_t = (v_t^k)_{k \in \mathbb{N}}$ , is a *statistical solution* of (1-1) with initial data  $\bar{\mu} \in \mathcal{P}(L^p(D))$ , if

$$(5-3) \quad \int_{\mathbb{R}_+} \int_{D^k} \langle v_{t,x}^k, \xi^1 \otimes \cdots \otimes \xi^k \rangle : \partial_t \varphi + \\ + \sum_{i=1}^k \langle v_{t,x}^k, \xi^1 \otimes \cdots \otimes f(\xi^i) \otimes \cdots \otimes \xi^k \rangle : \nabla_{x_i} \varphi dx dt + \\ + \int_{D^k} \langle \bar{v}_x^k, \xi^1 \otimes \cdots \otimes \xi^k \rangle : \varphi|_{t=0} dx = 0,$$

for all  $\varphi \in C_c^\infty(D^k \times \mathbb{R}_+, (\mathbb{R}^N)^k)$  and all  $k \in \mathbb{N}$ .

□

**Remark 5.3.** It is straightforward to see that if the initial data is a Dirac measure i.e.,  $\bar{\mu} = \delta_{\bar{u}}$  for some  $\bar{u} \in L^p(D)$  and if the corresponding statistical solution is also a Dirac measure i.e.,  $\mu_t = \delta_{u(t)}$  for some  $u(t) \in L^p(D)$  for almost every  $t$ , then the notion of statistical solutions reduces to that of the standard weak solution (2-2). Moreover, setting  $k = 1$  in (5-3), we see that the one-point correlation marginal of a statistical solution is precisely a measure-valued solution in the sense of DiPerna (4-2). Thus, a statistical solution can be thought of as measure-valued solution, supplemented with additional constraints on all possible (multi-point) spatial correlations. A priori, a statistical solution contains significantly more information than a measure-valued solution.

□



In [Fjordholm, Lanthaler, and Mishra \[2017\]](#), the authors showed existence and uniqueness of statistical solutions, under an additional entropy condition, for multi-dimensional scalar conservation laws. The computation of these solutions in the scalar case is presented in a recent paper [Fjordholm, Lye, and Mishra \[2017a\]](#).

**5.2 Computation of statistical solutions.** In a forthcoming paper [Fjordholm, Lye, and Mishra \[2017b\]](#), we present the following variant of the Monte Carlo [Algorithm 4.4](#) for computing the statistical solutions of systems of conservation laws. For notational convenience, we restrict the description to the one-dimensional case (2-1) in the following.

**Algorithm 5.4.** Let  $\Delta = \Delta x$  denote the grid size parameter and let  $M \in \mathbb{N}$ . Let  $\bar{\mu} \in \mathcal{P}(L^p(D))$  be the initial data.

**Step 1:** For some probability space  $(\Omega, \mathcal{X}, P)$ , draw  $M$  independent and identically distributed random fields  $u_0^{\Delta,1}, \dots, u_0^{\Delta,M} : \Omega \rightarrow L^p(D)$ , all with the same probability law  $\bar{\mu}$ .

**Step 2:** For each  $1 \leq l \leq M$  and for any fixed  $\omega \in \Omega$ , use the finite volume scheme (2-5) to numerically approximate the conservation law (2-1) with initial data  $u_0^{\Delta,l}(\omega)$ . Denote  $u^{\Delta,l}(\omega; \cdot, t) = S_t^\Delta u_0^{\Delta,l}(\omega; \cdot)$ , with  $S_t^\Delta$  being the data to solution operator associated with the scheme (2-5).

**Step 3:** Define the approximate statistical solution in terms of the *empirical measure*,

$$(5-4) \quad \mu_t^{\Delta,M} := \frac{1}{M} \sum_{l=1}^M \delta_{u^{\Delta,l}(\omega;t)}.$$

□

**Remark 5.5.** The first step in [Algorithm 5.4](#) can be ensured by requiring that there exists a probability space  $(\Omega, \mathcal{X}, P)$  and a random field  $\bar{u} \in L^2(\Omega; L^p(D))$  such that the law of  $\bar{u}$  with respect to  $P$  is  $\bar{\mu}$ . In most real-world applications, the uncertainty in initial data is usually described in terms of such a random field  $\bar{u}$ , for instance, one given as a parametric function  $\bar{u} : [0, 1]^Q \times D \rightarrow \mathcal{U}$ , with possibly  $Q \gg 1$ . The initial Monte Carlo samples  $u_0^{\Delta,1}, \dots, u_0^{\Delta,M} : \Omega \rightarrow L^1(D)$  are realizations of  $\bar{u}$ . □

The convergence of this algorithm will be demonstrated in the forthcoming paper [Fjordholm, Lye, and Mishra \[ibid.\]](#). We summarize the convergence theorem below,

**Theorem 5.6.** [[Fjordholm, Lye, and Mishra \[ibid.\]](#)] Consider the system of conservation laws (2-1) with a strictly convex entropy function  $\eta$  and a entropy flux function  $q$ . Let

$\bar{\mu} \in \mathcal{P}(L^p(D))$  be the initial data. We follow the notation introduced in the description of [Algorithm 5.4](#) and denote  $U_j^l(t)$  as the approximate solution generated at time  $t$  by the semi-discrete finite volume scheme (2-5) for initial data  $u_0^l(\omega)$  for  $1 \leq l \leq M$ , corresponding to the  $k$ -th sample of the [Algorithm 5.4](#). Assume that the numerical approximations satisfy the following,

i. A discrete entropy inequality of the form,

$$(5-5) \quad \frac{d}{dt} \eta(U_j^l(t)) + \frac{1}{\Delta x} \left( Q_{j+1/2}^l(t) - Q_{j-1/2}^l(t) \right) \leq 0, \quad \forall 1 \leq l \leq M, \quad \omega \in \Omega.$$

for a numerical entropy flux  $Q_{j+1/2}^l(t) := Q(U_{j-r+1}^l, \dots, U_j^l, \dots, U_{j+r}^l)$  for some  $r \geq 1$ . The numerical flux function is assumed to be consistent with the entropy flux  $q$  in (2-3) i.e,  $Q(U, \dots, U) = q(U)$  and (locally) Lipschitz continuous in all its arguments

ii. Weak BV bound i.e,

$$(5-6) \quad \int_0^T \sum_j |U_{j+1}^l(t) - U_j^l(t)|^q dt \leq C, \quad \text{for some } q \in [p, \infty)$$

for all  $1 \leq l \leq M$  and  $\omega \in \Omega$ .

iii. Approximate scaling. Denote  $\mathbf{v}_t^{\Delta, M}$  as the correlation measure associated with the approximate statistical solution  $\mu_t^{\Delta, M}$  and define the corresponding diagonal deficiency as in (5-2) as

$$(5-7) \quad d_r^p(\mathbf{v}^{\Delta, M, 2}) := \left( \int_0^T \int_D \int_{B_r(x)} \langle \mathbf{v}_{t,x,y}^{\Delta, M, 2}, |\xi_1 - \xi_2|^p \rangle dy dx dt \right)^{1/p}.$$

Then, we assume that for all  $s \in N$ , the following holds,

$$(5-8) \quad d_{s\Delta}^p(\mathbf{v}^{\Delta, M, 2}) \leq C s^\lambda d_\Delta^p(\mathbf{v}^{\Delta, M, 2}),$$

for some  $0 < \lambda \leq 1$  and a constant  $C$  that are independent of  $\Delta$  but can depend on the initial probability measure  $\bar{\mu}$

Then, upto a subsequence, the probability measures  $\mu_t^{\Delta, M}$ , generated by the [Algorithm 5.4](#), converge in the following sense to a  $\mu_t \in \mathcal{P}(L^p(D))$ :

$$(5-9) \quad \lim_{(\Delta, M) \rightarrow (0, \infty)} \int_0^T \int_{D^k} |\langle \mathbf{v}_{t,x}^{\Delta, M, k}, g(x, t, \xi) \rangle - \mathbf{v}_{t,x}^k, g(x, t, \xi) \rangle| dx dt = 0,$$

for almost every  $t \in (0, T)$  and for any function  $g : D^k \times [0, T] \rightarrow C(\mathcal{U}^k)$  satisfying

$$|g(x, t, \xi)| \leq C_g \Pi_{i=1}^k (1 + |\xi_i|^p)$$

$$|g(x + z, t, \xi + \zeta) - g(x, t, \xi)| \leq k_1(z) k_2(t) \max(|\xi_i|, |\xi_i + \zeta_i|)^{p-1} \Pi_{j \neq i} (1 + |\xi_j|)^p,$$

for some  $k_1$  locally bounded at origin and  $k_2 \in L^1((0, T))$ . Moreover,  $\mu_t$  is a statistical solution of (1-1) i.e, it satisfies (5-3)

The proof of this convergence theorem will be provided in the forthcoming paper [Fjordholm, Lye, and Mishra \[2017b\]](#) and relies in a crucial manner on a novel topology on  $\mathcal{P}(L^p(D))$  that is induced by the associated correlation measures. Convergence in this topology ensures strong convergence for all multi-point statistical quantities of interest.

The discrete entropy inequality (5-5) is crucial in obtaining uniform  $L^p$  bounds on the approximate statistical solution for  $p = 2$ , on account of the strict convexity of the entropy function  $\eta$ . The weak-BV bound (5-6) provides a *uniform diagonal continuity* properties for the associated correlation measures at the grid scales. Both these requirements are satisfied by many finite volume schemes, such as the TeCNO schemes of [Fjordholm, Mishra, and Tadmor \[2012\]](#). On the other hand, the scaling requirement (5-8) is necessary to show uniform diagonal continuity at scales that are larger than the grid scale i.e, the so-called *intermediate scales*. Currently, we are not able to prove that standard numerical methods satisfy this scaling requirement. However, such a requirement is a weaker version of the scaling hypothesis of Kolmogorov that is standard in the literature on incompressible turbulence [Frisch \[1995\]](#).

The extension of [Algorithm 5.4](#) and the proof of its convergence, to several space dimensions is straightforward [Fjordholm, Lye, and Mishra \[2017b\]](#).

**5.3 Numerical results.** Our aim is to compute (multi-point) statistical quantities of interest with the ensemble Monte-Carlo [Algorithm 5.4](#). To this end, we consider the two-dimensional compressible Euler equations (3-10) in the domain  $[0, 1]^2$  with periodic boundary conditions. As initial data, we consider the probability measure on  $L^2(D)$  induced by the random field given in (3-11). For subsequent computations, we use  $M = 400$  Monte Carlo samples and compute up to  $t = 5$  using grid resolutions from  $128^2$  up to  $1024^2$  grid points.

In [Figure 3](#) we plot for each grid resolution  $\Delta x$  the mean of the density variable,

$$(5-10) \quad \bar{\rho}^{\Delta x}(x, t) := \frac{1}{M} \sum_{l=1}^M \rho^{\Delta x, l}(\omega; x, t)$$

(where  $\rho^{\Delta x, l}(\omega)$  is the mass density of each individual Monte Carlo sample). We observe that small scale features are averaged out in the mean and only large scale structures, such

as the strong reentrant shocks (recall the periodic boundary conditions) and mixing regions, are retained through the averaging process. The figure indicates that, unlike the individual samples shown in Figure 1, the mean converges as the mesh is refined. This convergence is quantified in Figure 4(a) where we plot the difference in the mean density for successive resolutions,

$$(5-11) \quad \|\bar{\rho}^{\Delta x}(\cdot, t) - \bar{\rho}^{\Delta x/2}(\cdot, t)\|_{L^1}.$$

The figure indicates that this quantity goes to zero, so the mean of the approximations form a Cauchy sequence and hence converge.

Figure 5 shows the *variance* of the mass density

$$\text{Var}(\rho^{\Delta x})(x, t) := \frac{1}{M} \sum_{k=1}^M \left( \rho^{\Delta x, k}(\omega; x, t) - \bar{\rho}^{\Delta x}(\omega; x, t) \right)^2.$$

As in Figure 3, the variance of the approximate statistical solution seem to converge as the mesh is refined. This is again quantified in Figure 4(b) where the  $L^1$  differences of the variances at successive mesh resolutions is plotted. Note from Figure 5 that the variance is concentrated at the shocks and even more so in the mixing layer around the original interface.

Both the mean and variance are one-point statistical quantities of interest for the statistical solution and can be expressible through a measure-valued solution. However, we have seen in this section that correlations play a key role in the whole concept of statistical solutions. As a representative quantity, we consider the so-called *two-point structure function*,

$$(5-12) \quad S_h^p(v^{\Delta, M, 2})(t) := \left( \int_D \int_{B_h(x)} \langle v_{t,x,y}^{\Delta, M, 2}, |\xi_1 - \xi_2|^p \rangle dy dx \right)^{1/p}.$$

Note that (5-12) is a time snapshot of the diagonal deficiency (5-7) that plays a critical role in the convergence Theorem 5.6. For the case of  $p = 2$ , the so-called structure functions are an important observable in the theory as well in experiments for turbulent fluid flows Frisch [1995]. We plot  $S_h^2$  (5-12) at time  $t = 5$  and for different mesh resolutions, as function of the length scale  $h$  in Figure 6 (a). The result shows that the structure function behaves as  $S_h^2 \sim Ch^\lambda$ . Here both the constant  $C$  and the exponent  $\lambda$  appear to be independent of the mesh size  $\Delta$ . The exponent quickly converges to a value of  $\lambda \approx 0.5$  in this case. This is consistent with the requirement in Theorem 5.6 of uniform (in mesh size) diagonal continuity for the approximate statistical solutions. Furthermore, we compute Cauchy rates for the structure function  $S_h^2$  as a function of resolution and plot the results in Figure 6 (b). As predicted by the theory, this result shows that the computation of the structure function converges on mesh refinement.

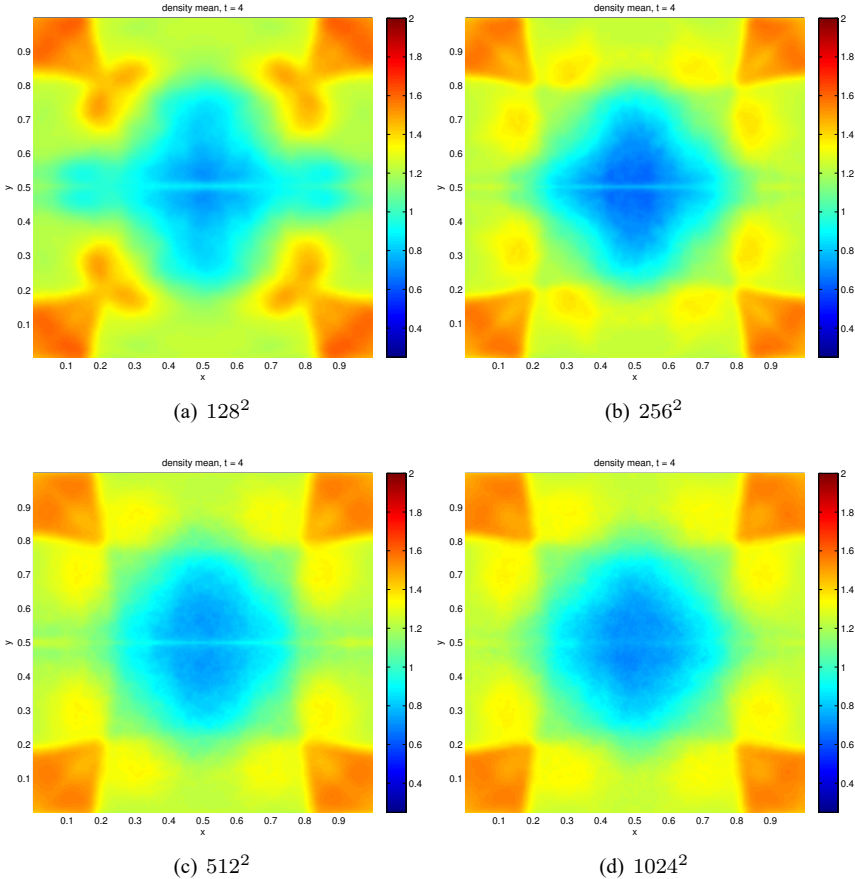


Figure 3: The mean density for the Richtmeyer-Meshkov problem with initial data (3-11) for different grid resolutions at time  $t = 4$ . All results are obtained with 400 Monte Carlo samples.

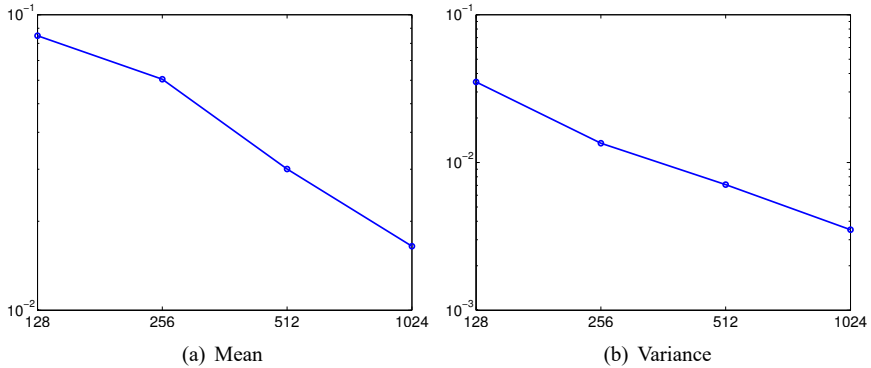


Figure 4: Cauchy rates (5-11) for the mean (left) and for variance (right) versus grid resolutions ( $x$ -axis) at time  $t = 4$  for the Richtmyer-Meshkov problem (3-11). All results are obtained with 400 Monte Carlo samples.

## 6 Conclusion

We considered the question of convergence, on mesh refinement, of state of the art numerical methods for hyperbolic systems of conservation laws (1-1). Although a large variety of numerical methods of the finite volume, finite difference and discontinuous Galerkin finite element type have been developed to approximate (1-1), rigorous proofs of convergence of these methods are relatively fewer. In the case of (multi-dimensional) scalar conservation laws, one can show convergence of the well known monotone schemes to the entropy solution by establishing bounds on the total variation. This program is harder to carry out for (arbitrary) high-order schemes due to the necessity of clipping at local extrema. On the other hand, compensated compactness techniques can be used to show convergence for high-order schemes in the scalar case.

In contrast to the scalar case, numerical examples show that numerical methods approximating the multi-dimensional system (1-1) may not necessarily converge to an entropy solution on mesh refinement. Structures are formed at finer and finer scales impeding convergence in spaces of integrable functions. One can weaken the notion of solutions by introducing entropy measure-valued solutions as a paradigm that characterizes the limit of numerical approximations of (1-1). Measure-valued solutions are Young measures and we can construct them using the Monte Carlo Algorithm 4.4 Fjordholm, Käppeli, Mishra, and Tadmor [2017]. One can prove weak-\* convergence of approximations, generated by the Algorithm 4.4, to measure-valued solutions and numerical experiments also illustrate the ability to compute one-point statistical quantities of interest.

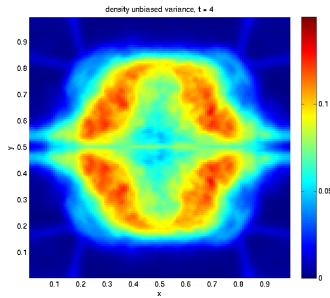
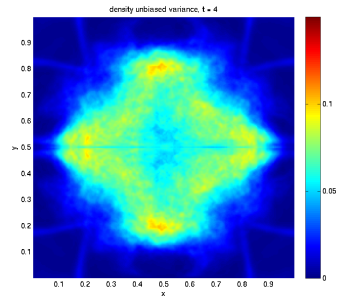
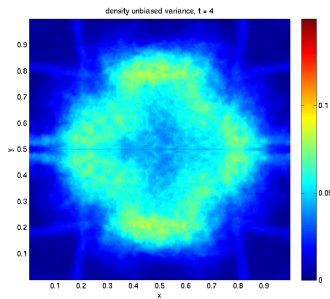
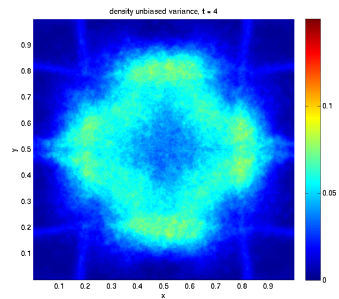
(a)  $128^2$ (b)  $256^2$ (c)  $512^2$ (d)  $1024^2$ 

Figure 5: Variance of the density with initial data (3-11) for different grid resolutions at time  $t = 4$ . All results are obtained with 400 Monte Carlo samples.

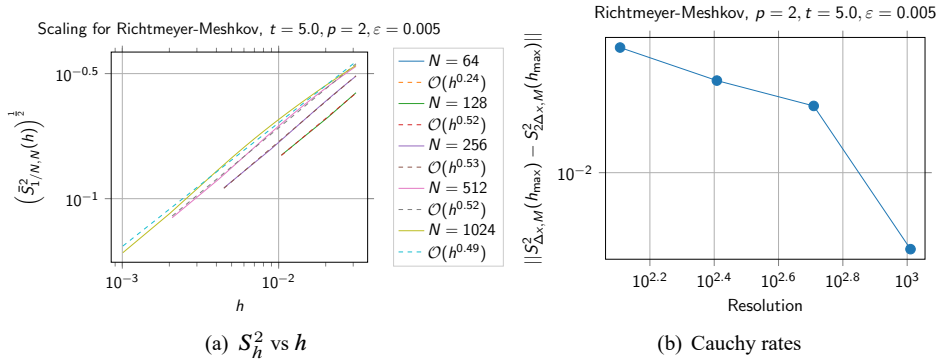


Figure 6: Computation of the two-point structure function  $S_h^2$  defined in (5-12) at  $t = 5$  for the Richtmyer-Meshkov initial data. Left  $S_h^2$  vs  $h$  at different mesh resolutions. Right. Cauchy rates for  $S_h^2$  for different mesh resolutions.

However, generic entropy measure-valued solutions are not necessarily unique, even for scalar conservation laws. Additional admissibility criteria need to be imposed. Following a recent paper [Fjordholm, Lanthaler, and Mishra \[2017\]](#), we consider statistical solutions as an appropriate solution paradigm for the multi-dimensional system of conservation laws (1-1). Statistical solutions are time parameterized probability measures on  $L^p$  spaces and are shown to be equivalent to a hierarchy of Young measures. One determines the time-evolution of statistical solutions in terms of an infinite family of moment equations (5-3). We present an ensemble Monte Carlo [Algorithm 5.4](#) and show under certain reasonable assumptions that the approximations generated by this algorithm converge to a statistical solution of systems of conservation laws in a suitable topology [Fjordholm, Lye, and Mishra \[2017b\]](#).

Summarizing, the state of the art answer to the question of convergence of numerical methods to systems of conservation laws appears to be that one cannot expect convergence in any pointwise or even integral sense for a single initial datum. On the other hand, statistical quantities, including very sophisticated multi-point correlations, converge for an ensemble of initial data. Thus, ensemble methods seem to be imperative when multi-dimensional systems of conservation laws are considered.

There are many outstanding issues in this direction. At the theoretical level, one needs further admissibility criteria (or entropy conditions) to uniquely specify the statistical solution of (1-1). Such conditions are hitherto undiscovered. Their discovery might require the design of novel numerical methods that are consistent with these criteria. Moreover,



Monte Carlo algorithms are very expensive computationally. There is a pressing need for the design of cheaper ensemble methods. Both these issues are topics of current research.

## References

- J. M. Ball (1989). “[A version of the fundamental theorem for Young measures](#)”. In: *PDEs and continuum models of phase transitions (Nice, 1988)*. Vol. 344. Lecture Notes in Phys. Springer, Berlin, pp. 207–215. MR: [1036070](#) (cit. on p. [3672](#)).
- Sylvie Benzon-Gavage and Denis Serre (2007). *Multidimensional hyperbolic partial differential equations*. Oxford Mathematical Monographs. First-order systems and applications. The Clarendon Press, Oxford University Press, Oxford, pp. xxvi+508. MR: [2284507](#) (cit. on p. [3660](#)).
- Stefano Bianchini and Alberto Bressan (2005). “[Vanishing viscosity solutions of nonlinear hyperbolic systems](#)”. *Ann. of Math. (2)* 161.1, pp. 223–342. MR: [2150387](#) (cit. on pp. [3660](#), [3668](#)).
- Alberto Bressan (2000). *Hyperbolic systems of conservation laws*. Vol. 20. Oxford Lecture Series in Mathematics and its Applications. The one-dimensional Cauchy problem. Oxford University Press, Oxford, pp. xii+250. MR: [1816648](#) (cit. on p. [3660](#)).
- Elisabetta Chiodaroli, Camillo De Lellis, and Ondřej Kreml (2015). “[Global ill-posedness of the isentropic system of gas dynamics](#)”. *Comm. Pure Appl. Math.* 68.7, pp. 1157–1190. MR: [3352460](#) (cit. on p. [3660](#)).
- Bernardo Cockburn and Chi-Wang Shu (1989). “[TVB Runge-Kutta local projection discontinuous Galerkin finite element method for conservation laws. II. General framework](#)”. *Math. Comp.* 52.186, pp. 411–435. MR: [983311](#) (cit. on pp. [3664](#), [3665](#)).
- G. M. Coclite, S. Mishra, and N. H. Risebro (2010). “[Convergence of an Engquist-Osher scheme for a multi-dimensional triangular system of conservation laws](#)”. *Math. Comp.* 79.269, pp. 71–94. MR: [2552218](#) (cit. on p. [3668](#)).
- Constantine M. Dafermos (2010). *Hyperbolic conservation laws in continuum physics*. Third. Vol. 325. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]. Springer-Verlag, Berlin, pp. xxxvi+708. MR: [2574377](#) (cit. on pp. [3659](#), [3660](#), [3662](#)).
- Camillo De Lellis and László Székelyhidi Jr. (2009). “[The Euler equations as a differential inclusion](#)”. *Ann. of Math. (2)* 170.3, pp. 1417–1436. MR: [2600877](#) (cit. on p. [3660](#)).
- Xia Xi Ding, Gui Qiang Chen, and Pei Zhu Luo (1989). “[Convergence of the fractional step Lax-Friedrichs scheme and Godunov scheme for the isentropic system of gas dynamics](#)”. *Comm. Math. Phys.* 121.1, pp. 63–84. MR: [985615](#) (cit. on p. [3668](#)).
- Ronald J. DiPerna (1985). “[Measure-valued solutions to conservation laws](#)”. *Arch. Rational Mech. Anal.* 88.3, pp. 223–270. MR: [775191](#) (cit. on pp. [3661](#), [3672](#)).

- U. S. Fjordholm (2013). “High-order accurate entropy stable numerical schemes for hyperbolic conservation laws”. Dissertation Nr. 21025. PhD thesis. ETH Zürich (cit. on pp. [3667](#), [3668](#)).
- U. S. Fjordholm, S. Lanthaler, and S. Mishra (2017). “Statistical solutions of hyperbolic conservation laws: foundations”. *Arch. Ration. Mech. Anal.* 226.2, pp. 809–849. MR: [3687882](#) (cit. on pp. [3661](#), [3676–3678](#), [3685](#)).
- Ulrik S. Fjordholm, Roger Käppeli, Siddhartha Mishra, and Eitan Tadmor (2017). “Construction of approximate entropy measure-valued solutions for hyperbolic systems of conservation laws”. *Found. Comput. Math.* 17.3, pp. 763–827. MR: [3648106](#) (cit. on pp. [3661](#), [3669](#), [3670](#), [3672–3675](#), [3683](#)).
- Ulrik S. Fjordholm, Siddhartha Mishra, and Eitan Tadmor (2012). “Arbitrarily high-order accurate entropy stable essentially nonoscillatory schemes for systems of conservation laws”. *SIAM J. Numer. Anal.* 50.2, pp. 544–573. MR: [2914275](#) (cit. on pp. [3664](#), [3667](#), [3668](#), [3670](#), [3675](#), [3680](#)).
- (2013). “ENO reconstruction and ENO interpolation are stable”. *Found. Comput. Math.* 13.2, pp. 139–159. MR: [3032678](#) (cit. on p. [3667](#)).
- (2016). “On the computation of measure-valued solutions”. *Acta Numer.* 25, pp. 567–679. MR: [3509212](#) (cit. on pp. [3661](#), [3669](#), [3670](#), [3672](#), [3675](#)).
- Ulrik Skre Fjordholm, Kjetil Lye, and Siddhartha Mishra (Oct. 2017a). “Numerical approximation of statistical solutions of scalar conservation laws”. arXiv: [1710.11173](#) (cit. on p. [3678](#)).
- (2017b). “Numerical approximation of statistical solutions of systems of conservation laws” (cit. on pp. [3661](#), [3678](#), [3680](#), [3685](#)).
- Uriel Frisch (1995). *Turbulence*. The legacy of A. N. Kolmogorov. Cambridge University Press, Cambridge, pp. xiv+296. MR: [1428905](#) (cit. on pp. [3680](#), [3681](#)).
- James Glimm (1965). “Solutions in the large for nonlinear hyperbolic systems of equations”. *Comm. Pure Appl. Math.* 18, pp. 697–715. MR: [0194770](#) (cit. on pp. [3660](#), [3668](#)).
- Edwige Godlewski and Pierre-Arnaud Raviart (1991). *Hyperbolic systems of conservation laws*. Vol. 3/4. Mathématiques & Applications (Paris) [Mathematics and Applications]. Ellipses, Paris, p. 252. MR: [1304494](#) (cit. on pp. [3660](#), [3663](#), [3665](#), [3667](#)).
- Sigal Gottlieb, Chi-Wang Shu, and Eitan Tadmor (2001). “Strong stability-preserving high-order time discretization methods”. *SIAM Rev.* 43.1, pp. 89–112. MR: [1854647](#) (cit. on p. [3664](#)).
- Ami Harten, Björn Engquist, Stanley Osher, and Sukumar R. Chakravarthy (1987). “Uniformly high-order accurate essentially nonoscillatory schemes. III”. *J. Comput. Phys.* 71.2, pp. 231–303. MR: [897244](#) (cit. on p. [3664](#)).

- Andreas Hiltebrand and Siddhartha Mishra (2014). “Entropy stable shock capturing space-time discontinuous Galerkin schemes for systems of conservation laws”. *Numer. Math.* 126.1, pp. 103–151. MR: [3149074](#) (cit. on pp. [3665](#), [3675](#)).
- Helge Holden and Nils Henrik Risebro (2011). *Front tracking for hyperbolic conservation laws*. Vol. 152. Applied Mathematical Sciences. First softcover corrected printing of the 2002 original. Springer, New York, pp. xii+361. MR: [2866066](#) (cit. on p. [3668](#)).
- Claes Johnson and Anders Szepessy (1987). “On the convergence of a finite element method for a nonlinear hyperbolic conservation law”. *Math. Comp.* 49.180, pp. 427–444. MR: [906180](#) (cit. on p. [3665](#)).
- R. Käppeli, S. C. Whitehouse, S. Scheidegger, U.-L. Pen, and M. Liebendörfer (2011). “FISH: A Three-dimensional Parallel Magnetohydrodynamics Code for Astrophysical Applications”. *The Astrophysical Journal Supplement* 195 (20) (cit. on pp. [3669](#), [3671](#)).
- Dietmar Kröner (1997). *Numerical schemes for conservation laws*. Wiley-Teubner Series Advances in Numerical Mathematics. John Wiley & Sons, Ltd., Chichester; B. G. Teubner, Stuttgart, pp. viii+508. MR: [1437144](#) (cit. on pp. [3660](#), [3665](#)).
- S. N. Kružkov (1970). “First order quasilinear equations with several independent variables”. *Mat. Sb. (N.S.)* 81 (123), pp. 228–255. MR: [0267257](#) (cit. on p. [3660](#)).
- Randall J. LeVeque (2002). *Finite volume methods for hyperbolic problems*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, pp. xx+558. MR: [1925043](#) (cit. on pp. [3660](#), [3663](#), [3664](#), [3667](#)).
- François Murat (1981). “Compacité par compensation: condition nécessaire et suffisante de continuité faible sous une hypothèse de rang constant”. *Ann. Scuola Norm. Sup. Pisa Cl. Sci. (4)* 8.1, pp. 69–102. MR: [616901](#) (cit. on p. [3668](#)).
- Steven Schochet (1989). “Examples of measure-valued solutions”. *Comm. Partial Differential Equations* 14.5, pp. 545–575. MR: [993820](#) (cit. on p. [3675](#)).
- Maria Elena Schonbek (1982). “Convergence of solutions to nonlinear dispersive equations”. *Comm. Partial Differential Equations* 7.8, pp. 959–1000. MR: [668586](#) (cit. on p. [3672](#)).
- L. Tartar (1979). “Compensated compactness and applications to partial differential equations”. In: *Nonlinear analysis and mechanics: Heriot-Watt Symposium, Vol. IV*. Vol. 39. Res. Notes in Math. Pitman, Boston, Mass.-London, pp. 136–212. MR: [584398](#) (cit. on pp. [3667](#), [3670](#), [3672](#)).
- Eleuterio F. Toro (1999). *Riemann solvers and numerical methods for fluid dynamics*. Second. A practical introduction. Springer-Verlag, Berlin, pp. xx+624. MR: [1717819](#) (cit. on pp. [3660](#), [3664](#)).

Received 2017-11-29.

SEMINAR FOR APPLIED MATHEMATICS

ETH ZÜRICH

RÄMISTRASSE 101

ZÜRICH

SWITZERLAND

[siddhartha.mishra@sam.math.ethz.ch](mailto:siddhartha.mishra@sam.math.ethz.ch)



# ON EFFECTIVE NUMERICAL METHODS FOR PHASE-FIELD MODELS

TAO TANG (汤涛)

## Abstract

In this article, we overview recent developments of modern computational methods for the approximate solution of phase-field problems. The main difficulty for developing a numerical method for phase field equations is a severe stability restriction on the time step due to nonlinearity and high order differential terms. It is known that the phase field models satisfy a nonlinear stability relationship called gradient stability, usually expressed as a time-decreasing free-energy functional. This property has been used recently to derive numerical schemes that inherit the gradient stability. The first part of the article will discuss implicit-explicit time discretizations which satisfy the energy stability. The second part is to discuss time-adaptive strategies for solving the phase-field problems, which is motivated by the observation that the energy functionals decay with time smoothly except at a few *critical* time levels. The classical operator-splitting method is a useful tool in time discretization. In the final part, we will provide some preliminary results using operator-splitting approach.

## 1 Introduction

Phase-field models have emerged as a powerful approach for modeling and predicting mesoscale morphological and microstructural evolution in materials. They were originally derived for the microstructure evolution and phase transition, but have been recently extended to many other physical phenomena, such as solid-solid transitions, growth of cancerous tumors, phase separation of block copolymers, dewetting and rupture of thin liquid films and infiltration of water into porous medium. In general, the phase-field models take two distinct values (for instance,  $+1$  and  $-1$ ) in each of the phases, with a smooth change

---

This work is partially supported by the Special Project on High-Performance Computing of the National Key R&D Program under No. 2016YFB0200604, Hong Kong Research Grants Council CERG grants, National Science Foundation of China, Hong Kong Baptist University FRG grants, and the SUSTech Start-Up Fund.

*MSC2010:* primary 65M15; secondary 65M70, 35Q35.

*Keywords:* Phase field equations, energy stability, adaptivity.

between both values in the zone around the interface, which is then diffused with a finite width. Many phenomenological macroscopic coarsening processes are energy driven in the sense that the dynamics is the gradient flow of a certain *energy functional* [Kohn \[2006\]](#).

Two of the phase-field models have attracted much attention: the molecular beam epitaxy (MBE) equation with slope selection

$$(1) \quad u_t = -\delta \Delta^2 u + \nabla \cdot f(\nabla u), \quad x \in \mathbb{R}^d, t \in (0, T],$$

and the Cahn-Hilliard (CH) equation

$$(2) \quad u_t = -\delta \Delta^2 u + \Delta f(u), \quad x \in \mathbb{R}^d, t \in (0, T].$$

In this paper, we consider

$$(3) \quad f(\phi) = \phi|\phi|^2 - \phi$$

for which the two phase-field models (1) and (2) become

$$(4) \quad u_t = -\delta \Delta^2 u + \nabla \cdot (|\nabla u|^2 \nabla u - \nabla u), \quad (x, y) \in \mathbb{R}^d, \quad t \in (0, T],$$

and

$$(5) \quad u_t = -\delta \Delta^2 u + \Delta(u^3 - u), \quad (x, y) \in \mathbb{R}^d, \quad t \in (0, T].$$

In (4),  $u$  is a scaled height function of epitaxial growth of thin films in a co-moving frame and the parameter  $\delta$  is a positive surface diffusion constant. In (5),  $u$  represents the concentration of one of the two metallic components of the alloy, and the positive parameter  $\delta$  represents the interfacial width, which is small compared to the characteristic length of the laboratory scale. An important feature of these two equations is that they can be viewed as the gradient flow of the following energy functionals:

$$(6) \quad E(u) = \int_{\Omega} \left[ \frac{\delta}{2} |\Delta u|^2 + \frac{1}{4} (|\nabla u|^2 - 1)^2 \right] dx$$

for the MBE equation and

$$(7) \quad E(u) = \int_{\Omega} \left[ \frac{\delta}{2} |\nabla u|^2 + \frac{1}{4} (|u|^2 - 1)^2 \right] dx$$

for the CH one. It is well known that both energy functionals decay in time

$$(8) \quad E(u(t)) \leq E(u(s)), \quad \forall t \geq s.$$

In this paper, we will review some recent works developing highly efficient numerical methods for phase field models. The main stability criteria is the energy decay principle (8). Among the time discretizations based on (8), [Eyre \[1993\]](#) convex splitting scheme should be specially mentioned. It is a first-order accurate unconditionally stable time-stepping scheme for gradient flows, which can be either linear or nonlinear depending on the ways of splitting. In particular, it has served as inspiration for many other time integration schemes in recent years, see, e.g., [Feng, Tang, and J. Yang \[2015\]](#), [Qiao and S. Sun \[2014\]](#), [Shen, C. Wang, X. Wang, and Wise \[2012\]](#), and [Shen, J. Xu, and J. Yang \[2017\]](#). Other significant works for higher order stable schemes for the phase field models can be found in [Gomez and Hughes \[2011\]](#), [Qiao, Z.-Z. Sun, and Z. Zhang \[2015\]](#), [Shen, C. Wang, X. Wang, and Wise \[2012\]](#), [Wise, C. Wang, and Lowengrub \[2009\]](#), [Xia, Y. Xu, and Shu \[2009\]](#), and [van der Zee, Oden, Prudhomme, and Hawkins-Daarud \[2011\]](#).

## 2 Time stablization by adding consistent terms

Since explicit schemes usually suffer from severe stability restrictions caused by the presence of high-order derivative terms and do not obey the energy decay property, semi-implicit schemes are widely used. It is known that explicit schemes usually suffer severe time step restrictions and generally do not obey energy conservation. To enforce the energy decay property and increase the time step, a good alternative is to use implicit-explicit (semi-implicit) schemes in which the linear part is treated implicitly (such as backward differentiation in time) and the nonlinear part is evaluated explicitly. For example, in [L.Q. Chen \[1998\]](#) Chen and Shen considered the semi-implicit Fourier-spectral scheme for (5) (set  $\delta = 1$ )

$$(1) \quad \frac{\widehat{u^{n+1}}(k) - \widehat{u^n}(k)}{\Delta t} = -|k|^4 \widehat{u^{n+1}}(k) - |k|^2 \widehat{f(u^n)}(k),$$

where  $\widehat{u^n}$  denotes the Fourier coefficient of  $u$  at time step  $t_n$ . On the other hand, the semi-implicit schemes can generate large truncation errors. As a result smaller time steps are usually required to guarantee accuracy and (energy) stability. To resolve this issue, a class of large time-stepping methods were proposed and analyzed in [Feng, Tang, and J. Yang \[2013\]](#), [He, Liu, and Tang \[2007\]](#), [Shen and X. Yang \[2010\]](#), [C. Xu and Tang \[2006\]](#), and [Zhu, Chen, Shen, and Tikare \[1999\]](#). The basic idea is to add an  $O(\Delta t)$  stabilizing term to the numerical scheme to alleviate the time step constraint whilst keeping energy stability.



The choice of the  $O(\Delta t)$  term is quite flexible. For example, in [Zhu, Chen, Shen, and Tikare \[1999\]](#) the authors considered the Fourier spectral approximation of the modified Cahn-Hilliard-Cook equation

$$(2) \quad \partial_t C = \nabla \cdot ((1 - aC^2)\nabla(C^3 - C - \kappa\nabla^2 C)).$$

The explicit Fourier spectral scheme is (see equation (16) therein)

$$(3) \quad \frac{\widehat{C^{n+1}}(k, t) - \widehat{C^n}(k, t)}{\Delta t} = ik \cdot \{(1 - aC^2)[ik'(\{-C + C^3\}_{k'}^n + \kappa|k'|^2\widehat{C^n}(k', t))]\}_k.$$

The time step for the above scheme has a severe constraint

$$(4) \quad \Delta t \cdot \kappa \cdot K^4 \leq 1,$$

where  $K$  is the number of Fourier modes in each coordinate direction. To increase the allowable time step, it is proposed in [Zhu, Chen, Shen, and Tikare \[ibid.\]](#) to add a term  $-Ak^4(\widehat{C^{n+1}} - \widehat{C^n})$  to the RHS of (3). Note that on the real side, this term corresponds to a fourth order dissipation, i.e.

$$-A\Delta^2(C^{n+1} - C^n)$$

which roughly is of order  $O(\Delta t)$ .

In [He, Liu, and Tang \[2007\]](#), a stabilized semi-implicit scheme was considered for the CH model, with the use of an order  $O(\Delta t)$  stabilization term

$$A\Delta(u^{n+1} - u^n).$$

Under a condition on  $A$  of the form:

$$(5) \quad A \geq \max_{x \in \Omega} \left\{ \frac{1}{2} |u^n(x)|^2 + \frac{1}{4} |u^{n+1}(x) + u^n(x)|^2 \right\} - \frac{1}{2}, \quad \forall n \geq 0,$$

one can obtain energy stability (8). Note that the condition (5) depends nonlinearly on the numerical solution. In other words, it implicitly uses the  $L^\infty$ -bound assumption on  $u^n$  in order to make  $A$  a controllable constant.

In 2010, Shen and Yang proved energy stability of semi-implicit schemes for the Allen-Cahn and the CH equations with truncated nonlinear term. More precisely it is assumed

that

$$(6) \quad \max_{u \in \mathbb{R}} |f'(u)| \leq L$$

which is what we referred to as the Lipschitz assumption on the nonlinearity in the abstract.

In 2011, Bertozzi et al. considered a nonlinear diffusion model of the form

$$\partial_t u = -\nabla \cdot (f(u) \nabla \Delta u) + \nabla \cdot (g(u) \nabla u),$$

where  $g(u) = f(u)\phi'(u)$ , and  $f, \phi$  are given smooth functions. In addition  $f$  is assumed to be non-negative. The numerical scheme considered in Bertozzi, Ju, and Lu [2011] takes the form

$$(7) \quad \frac{u^{n+1} - u^n}{\Delta t} = -A\Delta^2(u^{n+1} - u^n) - \nabla \cdot (f(u^n) \nabla \Delta u^n) + \nabla \cdot (g(u^n) \nabla u^n),$$

where  $A > 0$  is a parameter to be taken large. One should note the striking similarity between this scheme and the one introduced in Zhu, Chen, Shen, and Tikare [1999]. In particular in both papers the biharmonic stabilization of the form  $-A\Delta^2(u^{n+1} - u^n)$  was used. The analysis in Bertozzi, Ju, and Lu [2011] is carried out under the additional assumption that

$$(8) \quad \sup_n \|f(u^n)\|_\infty \leq A < \infty.$$

This is reminiscent of the  $L^\infty$  bound on  $u^n$ .

Roughly speaking, all prior analytical developments are conditional in the sense that either one makes a Lipschitz assumption on the nonlinearity, or one assumes certain a priori  $L^\infty$  bounds on the numerical solution. It is very desirable to *remove these technical restrictions* and establish a more reasonable stability theory.

In D. Li, Qiao, and Tang [2016], this problem is settled for the spectral Galerkin case. More precisely, the authors of D. Li, Qiao, and Tang [ibid.] considered a stabilized semi-implicit scheme introduced in He, Liu, and Tang [2007] following the earlier work C. Xu and Tang [2006]. It takes the form

$$(9) \quad \begin{cases} \frac{u^{n+1} - u^n}{\Delta t} = -\delta \Delta^2 u^{n+1} + A\Delta(u^{n+1} - u^n) + \Delta \Pi_N(f(u^n)), & n \geq 0, \\ u^0 = \Pi_N u_0. \end{cases}$$

where  $A > 0$  is the coefficient for the  $O(\Delta t)$  regularization term. For each integer  $N \geq 2$ , define

$$X_N = \text{span} \left\{ \cos(k \cdot x), \sin(k \cdot x) : k = (k_1, k_2) \in \mathbb{Z}^2, |k|_\infty = \max\{|k_1|, |k_2|\} \leq N \right\}.$$

Note that the space  $X_N$  includes the constant function (by taking  $k = 0$ ). The  $L^2$  projection operator  $\Pi_N : L^2(\Omega) \rightarrow X_N$  is defined by

$$(10) \quad (\Pi_N u - u, \phi) = 0, \quad \forall \phi \in X_N,$$

where  $(\cdot, \cdot)$  denotes the usual  $L^2$  inner product on  $\Omega$ . In yet other words, the operator  $\Pi_N$  is simply the truncation of Fourier modes of  $L^2$  functions to  $|k|_\infty \leq N$ . Since  $\Pi_N u_0 \in X_N$ , by induction it is easy to check that  $u^n \in X_N$  for all  $n \geq 0$ .

**Theorem 2.1** (Unconditional energy stability for 2D CH). *Consider (9) with  $\delta > 0$  and assume  $u_0 \in H^2(\Omega)$  with mean zero. Denote  $E_0 = E(u_0)$  the initial energy. There exists a constant  $\beta_c > 0$  depending only on  $E_0$  such that if*

$$(11) \quad A \geq \beta \cdot \left( \|u_0\|_{H^2}^2 + \delta^{-1} |\log \delta|^2 + 1 \right), \quad \beta \geq \beta_c,$$

then

$$E(u^{n+1}) \leq E(u^n), \quad \forall n \geq 0,$$

where  $E$  is defined by (7). Furthermore, let  $u_0 \in H^s$ ,  $s \geq 4$  with mean zero. Let  $u(t)$  be the solution to (5) with initial data  $u_0$ . Let  $u^n$  be defined according to (9) with initial data  $\Pi_N u_0$ . If  $A$  satisfies (11), then

$$\|u(t_m) - u^m\|_2 \leq A \cdot e^{C_1 t_m} \cdot C_2 \cdot (N^{-s} + \Delta t),$$

where  $t_m = m\Delta t$ ,  $C_1 > 0$  depends only on  $(u_0, \delta)$ ,  $C_2 > 0$  depends on  $(u_0, \delta, s)$ .

There is an analogue of Theorem 2.1 for the MBE Equation (4). Consider the following semi-implicit scheme for MBE (4):

$$(12) \quad \begin{cases} \frac{u^{n+1} - u^n}{\tau} = -\delta \Delta^2 u^{n+1} + A \Delta(u^{n+1} - u^n) + \Pi_N \nabla \cdot (g(\nabla u^n)), & n \geq 0, \\ u^0 = \Pi_N u_0. \end{cases}$$

This scheme was introduced and analyzed in C. Xu and Tang [2006] (see also Qiao, Z. Zhang, and Tang [2011]). The authors of C. Xu and Tang [2006] first introduced the

stabilized  $O(\Delta t)$  term of the form  $A\Delta(u^{n+1} - u^n)$  as given in (12), and provided an energy stability analysis based on the assumption that  $A$  depends implicitly on the  $L^\infty$  bound on the numerical solution  $u^n$ . Note that the result in D. Li, Qiao, and Tang [2016] provide a clean description on the size of the constant  $A$ , in the sense that  $A$  is independent of the  $L^\infty$  bound on the numerical solution. The energy-supercritical three-dimensional case is analysed in D. Li and Qiao [2017b] by exploiting discrete smoothing estimates.

Note that above results are restricted to the first-order time discretization. On the other hand, D. Li and Qiao [2017a] introduced recently several novel stabilization techniques for second-order schemes. Quite surprisingly, it is found that depending on the form of numerical discretization (such as  $f(2u^n - u^{n-1})$  v.s.  $2f(u^n) - f(u^{n-1})$ ) the corresponding scheme can have conditional stability or unconditional stability with the stabilization parameter depending only on initial data and the diffusion coefficient. Developing upon the second-order scheme in D. Li and Qiao [ibid.], Song and Shu [2017] constructed a new unconditionally stable second-order implicit–explicit local discontinuous Galerkin Method for the Cahn–Hilliard Equation.

### 3 Time stepping with $p$ -adaptivity

As the governing equations (4) and (5) involve the perturbed (i.e., the coefficient  $\delta \ll 1$ ) biharmonic operators and strong nonlinearities, it is very difficult to design efficient time discretization strategy which can resolve dynamics and steady state of the corresponding phase field models. Moreover, nonlinear energy stability which is intrinsic to the phase field models (see, e.g., Figure 1) is also a challenging issue for numerical approximations. Numerical evidences show that violating the energy stability may lead to non-physical oscillations. Consequently, a satisfactory numerical strategy needs to balance solution accuracy, efficiency and nonlinear stability.

Below we will briefly outline the motivation of this section. Our numerical evidences show that the lower order time discretizations may require very small time stepsizes in order to resolve the short time dynamics of the phase field problems. Figure 2 gives a typical example which gives energy evolutions for the Cahn–Hilliard Equation (5) with  $\Delta t = 1/1000, 1/100, 1/50$ . It is observed that a time step smaller than  $10^{-2}$  is needed in order to obtain accurate solutions.

For improvement, one quick idea is to use higher order time discretization. However, there has few higher order energy-stable schemes, particularly for order 3 or higher. Our idea is to use the so-called spectral deferred correction (SDC) method which was first introduced to solve initial value ordinary differential equations (ODEs) by Dutt, Greengard, and Rokhlin [2000]. The key idea of the SDC method is to first convert the original ODEs into the corresponding Picard equation and then apply a deferred correction procedure

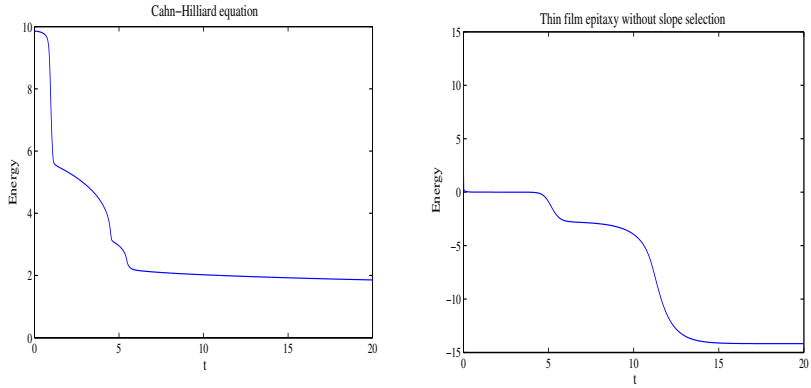


Figure 1: Illustrative energy curves for the three different models.

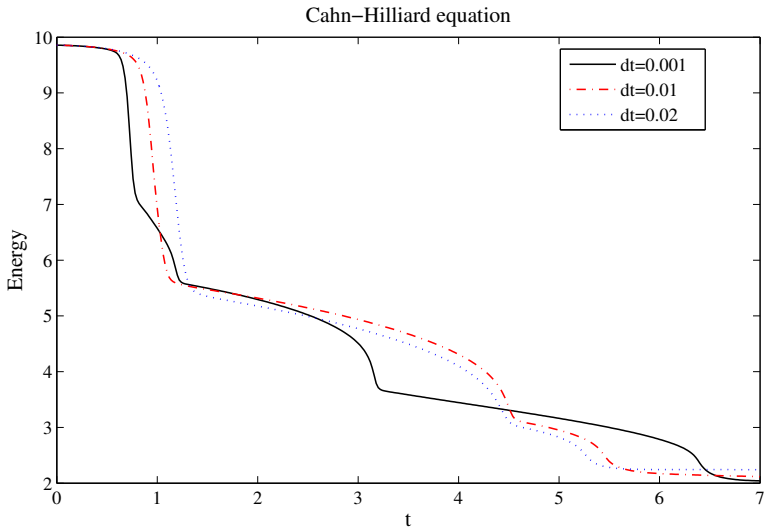


Figure 2: A typical example for the energy dependent on time steps for the Cahn-Hilliard equation.

in the integral formulation, aiming to achieve higher order accuracy in an *iterative* way.

The reasons for us to employ the SDC method are the following: iteration loops can improve the formal accuracy in a flexible and simple way; the SDC method was designed to handle stiff systems which are the case of our perturbed singularly nonlinear equations; and the flexibility of the order enhancement is useful for our local adaptive strategy to be described later. On the other hand, although the SDC method can solve the short-time dynamics very well (e.g., a 5th-order time discretization can fix the problem in Figure 2 with  $\Delta t = 1/20$ ), unfortunately, a higher order time discretization may yield numerical instability as the nonlinear stability can not be guaranteed for higher order time discretizations. A typical example is given in Figure 3, which solves the same example as in Figure 2 but with an 3rd order SDC method (i.e.  $Np = 2$  in the figure) and an 5th order SDC method (i.e.  $Np = 4$ ). It is observed that the discrete energies blow up before  $T = 30$ .

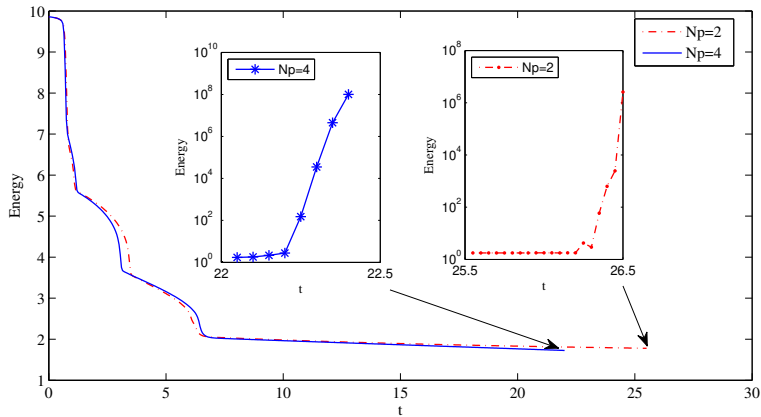


Figure 3: A typical energy blow-up with 3rd (right) or a 5th (left) order time discretization for the Cahn-Hilliard equation.

Note that the problem in Figs. 2 and 3 is partially due to the use of the central-difference approaches in space approximation (see Feng, Tang, and J. Yang [2015]). A more elegant approach using discontinuous Galerkin (DG) method together with SDC methods. Stable and accurate numerical results have been obtained in Guo and Y. Xu [2016] and Guo, Xia, and Y. Xu [2017]. On the other hand, for simple central-difference approaches in space, we can use a hybrid  $p$ -adaptive method which chooses appropriate order of accuracy at each time level. It is seen from the energy curves in Figure 1 that first-order methods should be good enough in most of time regimes, but in some critical stages with rapid

energy change appropriate adaptive strategies must be used. Some  $p$ -adaptive details will be reported and the relevant numerical results will be presented in this section.

**3.1 Convex splitting methods.** An important class of time discretization is the convex splitting method originally proposed by Eyre [1993], see also improved version of Shen, J. Xu, and J. Yang [2017] and Shen, C. Wang, X. Wang, and Wise [2012], which can produce *unconditional* energy stability (in the sense that the stability is irrelevant with the choice of the time steps). If we can express the free energy as the difference of two convex functional, namely  $E = E_c - E_e$ , where both  $E_c$  and  $E_e$  are convex about  $u$ , then we may use the concept of convex splitting due to Eyre [1993] to obtain highly stable numerical schemes.

Below we will demonstrate the convex splitting by considering the Cahn-Hilliard Equation (5). Using the splitting form

$$(1) \quad E_c(u) = \int_{\Omega} \left( \frac{\delta}{2} |\nabla u|^2 + \frac{\beta}{2} u^2 \right) dx, \quad E_e(u) = \int_{\Omega} \left( \frac{\beta}{2} u^2 - F(u) \right) dx,$$

where  $F = (|u|^2 - 1)^2/4$ , and the corresponding semi-discrete scheme to the Cahn-Hilliard Equation (5) is

$$(2) \quad \begin{aligned} \frac{u^{n+1} - u^n}{\Delta t} &= \Delta \left( \frac{\delta E_c(u^{n+1})}{\delta u} - \frac{\delta E_e(u^n)}{\delta u} \right) \\ &= -\epsilon^2 \Delta^2 u^{n+1} + \beta \Delta u^{n+1} - \beta \Delta u^n + \Delta f(u^n). \end{aligned}$$

It can be proven (see, e.g., Feng, Tang, and J. Yang [2015]) that if the constant  $\beta$  is sufficiently large then the semi-discrete scheme (2) is unconditionally energy stable, i.e.,  $E(u^{n+1}) \leq E(u^n)$ , where the energy  $E$  is defined by (7). Similarly, for the MBE model (4), using the convex splitting

$$(3) \quad E_c(u) = \int_{\Omega} \left( \frac{\epsilon^2}{2} |\Delta u|^2 + \frac{\beta}{2} |\nabla u|^2 \right) dx, \quad E_e(u) = \int_{\Omega} \left( \frac{\beta}{2} |\nabla u|^2 - F(\nabla u) \right) dx,$$

gives the corresponding semi-discrete scheme

$$(4) \quad \begin{aligned} \frac{u^{n+1} - u^n}{\Delta t} &= - \left( \frac{\delta E_c(u^{n+1})}{\delta u} - \frac{\delta E_e(u^n)}{\delta u} \right) \\ &= -\epsilon^2 \Delta^2 u^{n+1} + \beta \Delta u^{n+1} - \beta \Delta u^n + \nabla \cdot f(\nabla u^n). \end{aligned}$$

In practical computations, for both (4) and (5) with  $f(u)$  of the form (3),  $\beta = 1$  can guarantee the energy stability for (4), and  $\beta = 2$  can guarantee the energy stability for (2).

**3.2 Spectral deferred correction method.** Assume the time interval  $[0, T]$  into  $N$  non-overlapping intervals  $0 = t_0 < t_1 < \cdots < t_N = T$ . Let  $\Delta t_n = t_{n+1} - t_n$  and  $u_n$  denotes the numerical solution of  $u(t_n)$ , with  $u_0 = u(t_0)$ . Based on the convex splitting schemes presented in the above subsection, our convex splitting scheme can be written in the following form

$$(5) \quad u^{n+1} = u^n + \Delta t_n (F_E(u^n) + F_I(u^{n+1}))$$

for convenience, where  $F_N$  represents the explicit part and  $F_I$  represents the implicit part.

The SDC method is a one step, multi-stage method. Denoting the  $p + 1$  Legendre-Guass-Radau Ila nodes (cf. [Shen, Tang, and L.-L. Wang \[2011\]](#)) on  $[-1, 1]$  by  $-1 = r_0 < r_1 < \cdots < r_{p-1} < r_p = 1$  and letting

$$t_{n,i} = \frac{t_{n+1} - t_n}{2} r_i + \frac{t_{n+1} + t_n}{2}, \quad i = 0, 1, \dots, p,$$

we obtain the spectral nodes on interval  $[t_n, t_{n+1}]$  of the form  $t_n = t_{n,0} < t_{n,1} < \cdots < t_{n,p-1} < t_{n,p} = t_{n+1}$ . Then the interval  $[t_n, t_{n+1}]$  is divided into  $p$  subintervals. Let  $\Delta t_{n,m} = t_{n,m+1} - t_{n,m}$  and  $u_{n,m}^k$  denotes the  $k^{th}$  order approximation to  $u(t_{n,m})$ .

Note that we do the SDC procedure in every interval  $[t_n, t_{n+1}]$ . Given  $u_n$ , we wish to approximate  $u_{n+1}$ . Let  $u_{n,0}^1 = u_n$ . We first compute a first order accurate approximate solution  $u^1$  at the nodes  $\{t_{n,m}\}_{m=1}^p$ :

$$(6) \quad u_{n,m+1}^1 = u_{n,m}^1 + \Delta t_{n,m} (F_E(t_{n,m}, u_{n,m}^1) + F_I(t_{n,m+1}, u_{n,m+1}^1)).$$

We then do the successive corrections. For each  $1 \leq k \leq K$ , let  $u_{n,0}^{k+1} = u_n$ . For  $m = 0, \dots, p-1$ , we use

$$(7) \quad \begin{aligned} u_{n,m+1}^{k+1} = & u_{n,m}^{k+1} + \Delta t_{n,m} (F_E(t_{n,m}, u_{n,m}^{k+1}) - F_E(t_{n,m}, u_{n,m}^k) + F_I(t_{n,m+1}, u_{n,m+1}^{k+1}) \\ & - F_I(t_{n,m+1}, u_{n,m+1}^k)) + I_m^{m+1}(F_E(t, u^k) + F_I(t, u^k)), \end{aligned}$$

where the last part is the integral of the  $p$ -th degree interpolating polynomial on the  $p + 1$  points

$$(t_{n,m}, F_E(t_{n,m}, u_{n,m}^k) + F_I(t_{n,m}, u_{n,m}^k))_{m=0}^p$$

over the subinterval  $[t_{n,m}, t_{n,m+1}]$ , which is the numerical quadrature approximation of

$$\int_{t_{n,m}}^{t_{n,m+1}} (F_E(\tau, u(\tau)) + F_I(\tau, u(\tau))) d\tau.$$



The above procedure leads to  $u_{n+1} = u_{n,p}^{K+1}$ .

For more details of the SDC method, we refer the readers to [Dutt, Greengard, and Rokhlin \[2000\]](#) and [Minion \[2003\]](#) and the recent work [Guo and Y. Xu \[2016\]](#) and [Tang, Xie, and Yin \[2013\]](#).

**3.3 Efficiency enhancement with  $p$ -adaptivity.** It is known that the convex splitting method can preserve the energy stability but accuracy may not be satisfactory. Although using SDC may enhance accuracy, the SDC corrections may cause blow-up as demonstrated in Fig. 3. It remains to balance the accuracy and stability. To this end, an adaptive strategy adjusting the correction number was proposed in [Feng, Tang, and J. Yang \[2015\]](#) based on the discrete energies  $E_h(u^n)$  and  $E_h(u^{n-1})$ :

(8)

$$Np = \min\{Nmax, \max\{0, Nmax + \text{fix}[\log_\eta(|E_h(u^n) - E_h(u^{n-1})| + \eta^{-(Nmax+1)})]\}\},$$

where  $\eta$  is a positive constant,  $Nmax$  is the maximum number of corrections and  $\text{fix}[\cdot]$  represents the integer part of a number. Below we will explain the motivation of using (8) to predict  $Np$ . It is clear that more corrections are needed in the region where the energy decays fast. More specifically, the relationship between  $Np$  and the energy change is given as following:

$$(9) \quad Np = \begin{cases} 0, & \text{if } |E_h(u^n) - E_h(u^{n-1})| < \eta^{-Nmax} \\ k, & \text{if } \eta^{-Nmax+k} \leq |E_h(u^n) - E_h(u^{n-1})| < \eta^{-Nmax+k+1} \\ Nmax, & \text{if } |E_h(u^n) - E_h(u^{n-1})| \geq \eta^{-1} \end{cases},$$

where  $Nmax$  is upper bounded by  $2p - 1$  as the accuracy order of the interpolation on the  $p + 1$  Gauss-Radau nodes is  $2p$  and the parameter  $\eta$  can be fixed as 3 or 5.

Note that the energy decreasing property motivates us to use the energy difference at  $t_{n-1}$  and  $t_n$  for choosing the number of corrections. Firstly, as observed from the energy curves in [Figure 1](#), the energy variation in most time regimes is very small, so  $Np = 0$  should be chosen in most of the time intervals. This implies that only first order SDC method is used, which guarantees the energy stability in general. Secondly, in the transition regime, the energy variation is between  $\eta^{-Nmax}$  and  $\eta^{-1}$ , which indicates some variable value of  $Np$  is used based on the size of the energy variation. Thirdly, if the energy variation exceeds  $\eta^{-1}$ , then the maximum number of correction should be used. In the later two cases, the energy decreasing property may not be preserved locally. However, as the total number of the intervals relevant to the last two cases is very small, it is expected that the overall energy stability can be preserved well. In other words, the choice of (9) seems very useful to balance the accuracy and overall energy stability.

**Example 3.1.** We will use the adaptive SDC scheme for the Cahn-Hilliard Equation (5) with initial condition  $u_0(x, y) = 0.05 \sin x \sin y + 0.001$ ,  $0 \leq x, y \leq 2\pi$ , and the periodic boundary condition. The parameter  $\delta$  is chosen as 0.01.

The mesh grid in space is fixed as  $400 \times 400$ . We take the numerical solutions with small uniform time step  $dt = 0.001$  as the “reference” solution. We take  $p = 4$  in the SDC method, and  $\beta = 1$  in (2),  $\eta = 5$ ,  $N_{max} = 5$  in (9) and set  $Np = N_{max}$  at the first step.

In Figure 4, the numerical results using adaptive SDC scheme with  $dt = 0.04$  produce graphically indistinguishable energy curve as that for un-adaptive  $dt = 0.001$  results. On the other hand, the energy curve with un-adaptive  $dt = 0.04$  is quite far from the reference energy curve, especially before  $T = 10$ , which can be seen in the locally magnified energy curves from  $T = 2$  to 8.

The CPU time comparison is presented in Figure 5, where it is seen that our adaptive SDC scheme consumes more CPU time at beginning as more corrections are needed to capture the fast dynamical evolution. However, the adaptive SDC scheme can enhance the efficiency significantly in the long time computation. The numerical solutions at different time levels are presented in Figure 6, where it is observed that the solution dynamics can be captured correctly with larger time steps when adaptive strategy is employed.

## 4 Operator splitting method

Following the approach in Chertock, Kurganov, and Petrova [2009], we split Eq. (4) into the nonlinear part

$$(1) \quad u_t = \nabla \cdot (|\nabla u|^2 \nabla u),$$

and linear part

$$(2) \quad u_t = -\Delta u - \delta \Delta^2 u.$$

We denote by  $\mathcal{S}_{\mathbf{n}}$  the exact solution operator associated with (1) and by  $\mathcal{S}_{\mathcal{L}}$  the exact solution operator associated with (2). Notice that the corresponding energy functionals,

$$(3) \quad E_{\mathbf{n}}(u) = \frac{1}{4} \int_{\Omega} |\nabla u|^4 dx dy,$$

$$(4) \quad E_{\mathcal{L}}(u) = \int_{\Omega} \left( \frac{\delta}{2} |\Delta u|^2 - \frac{1}{2} |\nabla u|^2 + \frac{1}{4} \right) dx dy$$

decay. Then, introducing a (small) splitting step  $\Delta t$ , the solution of the original equation (4) (which is assumed to be available at time  $t$ ) is evolved using the Strang splitting method, one step of which can be written as

$$u(x, y, t + \Delta t) = \mathcal{S}_{\mathcal{L}}(\Delta t/2) \mathcal{S}_{\mathcal{N}}(\Delta t) \mathcal{S}_{\mathcal{L}}(\Delta t/2) u(x, y, t).$$

A similar splitting approach is applied to equation (5), for which the linear part is still (2) and the nonlinear one is

$$(5) \quad u_t = \Delta(u^3).$$

As in the case of the MBE equation, the corresponding energy functionals,

$$(6) \quad E_{\mathcal{N}}(u) = \frac{1}{4} \int_{\Omega} u^4 dx dy,$$

$$(7) \quad E_{\mathcal{L}}(u) = \int_{\Omega} \left( \frac{\delta}{2} |\nabla u|^2 - \frac{1}{2} u^2 + \frac{1}{4} \right) dx dy$$

decay. We stress that even though the linear parts of equations (4) and (5) are the same, the functionals (4) and (7) are different since they are associated with the corresponding parts of the energy functionals (6) and (7).

In order to implement the splitting method, the exact solution operators  $\mathcal{S}_{\mathcal{N}}$  and  $\mathcal{S}_{\mathcal{L}}$  have to be replaced by their numerical approximations. Note that one of the main advantages of the operator splitting technique is the fact that the nonlinear, (1) and (5), and linear, (2), subproblems, which are of different nature, can be solved numerically by different methods. First, using the method of lines, (1) and (5) can be reduced to systems of ODEs, which can be efficiently and accurately integrated by large stability domain explicit ODE solvers Abdulle [2002]. Second, since (2) is linear, one can solve it (practically) exactly using, for example, the pseudo-spectral method. This way, no stability restrictions on solving (2) are imposed.

**4.1 Finite-Difference Methods for (1) and (5).** In this section, we propose efficient explicit finite-difference methods for the degenerate parabolic equations (1) and (5). These methods are based on the semi-discretization of (1) and (5) followed by the use of an efficient and accurate ODE solver. The ODE solver will be utilized to evolve the solutions of (1) and (5) from time  $t$  to  $t + \Delta t$ . We note that in a general case the time-steps of the ODE solver denoted by  $\Delta t_{\text{ODE}}$  will be smaller than the splitting step  $\Delta t$  so that the approximation of  $\mathcal{S}_{\mathcal{N}}(\Delta t)$  will typically require several  $\Delta t_{\text{ODE}}$  steps.

We first design  $2m$ th-order centered-difference schemes for the 1-D version of (1):

$$(8) \quad u_t = (u_x^3)_x, \quad x \in [0, L], \quad t \in (0, T].$$

We consider a uniform grid with nodes  $x_j$ , such that  $x_{j+1} - x_j = \Delta x$ ,  $\forall j$ , and introduce the following  $2m$ th-order discrete approximation of the  $\frac{\partial}{\partial x}$  operator:

$$(9) \quad (\psi_x)_j := \sum_{p=-m}^m \alpha_p \psi_{j+p} = \psi_x(x_j) + \mathcal{O}((\Delta x)^{2m}).$$

For example, when  $m = 2$ , we obtain a fourth-order centered-difference approximation by taking

$$\alpha_1 = -\alpha_{-1} = \frac{2}{3\Delta x}, \quad \alpha_2 = -\alpha_{-2} = -\frac{1}{12\Delta x}.$$

Equipped with the above approximation of spacial derivatives, we discretize equation (8) using the method of lines as follows:

$$(10) \quad \frac{du_j}{dt}(t) = \sum_{p=-m}^m \alpha_p H_{j+p}(t) =: F_j(t),$$

where  $u_j(t)$  denotes the computed point value of the solution at  $(x_j, t)$ , and

$$(11) \quad H_j(t) := (u_x)_j^3(t) \quad \text{with} \quad (u_x)_j(t) := \sum_{p=-m}^m \alpha_p u_{j+p}(t).$$

Note that the above quantities depend on  $t$ , but for the sake of brevity we will suppress this dependence from now on.

It is proven in [Cheng, Kurganov, Qu, and Tang \[2015\]](#) that the semi-discrete schemes (10)-(11) satisfy the following energy decay property:

$$\frac{d}{dt} E_{\mathfrak{n}}^{\Delta} \leq 0,$$

where  $E_{\mathfrak{n}}^{\Delta}$  is a 1-D discrete version of the energy functional (3):  $E_{\mathfrak{n}}^{\Delta} := \frac{1}{4} \sum_j (u_x)_j^4 \Delta x$ .

We now consider the finite-difference schemes for  $u_t = \nabla \cdot (|\nabla u|^2 \nabla u)$ , i.e., (1). We consider a uniform grid with nodes  $(x_j, y_k)$ , such that  $x_{j+1} - x_j = \Delta x$ ,  $\forall j$ ,  $y_{k+1} - y_k = \Delta y$ ,  $\forall k$ , and introduce the following  $2m$ th-order discrete approximation of the  $\frac{\partial}{\partial x}$  and  $\frac{\partial}{\partial y}$

operators:

$$(12) \quad \begin{aligned} (\psi_x)_{j,k} &:= \sum_{p=-m}^m \alpha_p \psi_{j+p,k} = \psi_x(x_j, y_k) + \mathcal{O}((\Delta x)^{2m}), \\ (\psi_y)_{j,k} &:= \sum_{p=-m}^m \beta_p \psi_{j,k+p} = \psi_y(x_j, y_k) + \mathcal{O}((\Delta y)^{2m}). \end{aligned}$$

For example, when  $m = 2$ , we obtain a fourth-order centered-difference approximation by taking

$$\begin{aligned} \alpha_1 &= -\alpha_{-1} = \frac{2}{3\Delta x}, & \alpha_2 &= -\alpha_{-2} = -\frac{1}{12\Delta x}, \\ \beta_1 &= -\beta_{-1} = \frac{2}{3\Delta y}, & \beta_2 &= -\beta_{-2} = -\frac{1}{12\Delta y}. \end{aligned}$$

Equipped with the above approximation of spacial derivatives,  $2m$ th-order semi-discrete finite-difference schemes for (1) read:

$$(13) \quad \frac{du_{j,k}}{dt} = \sum_{p=-m}^m \alpha_p H_{j+p,k}^x + \sum_{p=-m}^m \beta_p H_{j,k+p}^y =: F_{j,k},$$

where

$$(14) \quad H_{j,k}^x := (u_x)_{j,k}^3 + (u_y)_{j,k}^2 (u_x)_{j,k} \quad \text{and} \quad H_{j,k}^y := (u_y)_{j,k}^3 + (u_x)_{j,k}^2 (u_y)_{j,k}$$

with

$$(15) \quad (u_x)_{j,k} := \sum_{p=-m}^m \alpha_p u_{j+p,k} \quad \text{and} \quad (u_y)_{j,k} := \sum_{p=-m}^m \beta_p u_{j,k+p}.$$

It is shown in [Cheng, Kurganov, Qu, and Tang \[2015\]](#) that the semi-discrete schemes (13)–(15) satisfy the following energy decay property:

$$\frac{d}{dt} E_{\mathfrak{n}}^{\Delta} \leq 0,$$

where  $E_{\mathfrak{n}}^{\Delta}$  is a 2-D discrete version of the energy functional (3):  $E_{\mathfrak{n}}^{\Delta} := \frac{1}{4} \sum_j |\nabla_h u_{j,k}|^4 \Delta x \Delta y$  with  $\nabla_h u_{j,k} := ((u_x)_{j,k}, (u_y)_{j,k})^T$ .

We now design semi-discrete finite-difference schemes for  $u_t = \Delta(u^3)$ , i.e., (5). We use the same grids and the same  $2m$ th-order discrete approximation of the  $\frac{\partial}{\partial x}$  and  $\frac{\partial}{\partial y}$

operators as above. Then,  $2m$ th-order semi-discrete finite-difference schemes for (5) read:

$$(16) \quad \frac{du_{j,k}}{dt} = \sum_{p=-m}^m \alpha_p H_{j+p,k}^x + \sum_{p=-m}^m \beta_p H_{j,k+p}^y =: F_{j,k},$$

where

$$(17) \quad H_{j,k}^x := \sum_{p=-m}^m \alpha_p u_{j+p,k}^3 \quad \text{and} \quad H_{j,k}^y := \sum_{p=-m}^m \beta_p u_{j,k+p}^3.$$

It can be shown that the semi-discrete schemes (16)-(17) satisfy the following energy decay property:

$$\frac{d}{dt} E_{\mathfrak{n}}^{\Delta} \leq 0,$$

where  $E_{\mathfrak{n}}^{\Delta}$  is a 2-D discrete version of the energy functional (6):  $E_{\mathfrak{n}}^{\Delta} := \frac{1}{4} \sum_j u_{j,k}^4 \Delta x \Delta y$ .

**4.2 Large Stability Domain Explicit ODE Solver.** The ODE systems (10), (13) and (16) have to be solved numerically. Recall that explicit ODE solvers typically require time-steps to be  $\Delta t_{\text{ODE}} \sim (\Delta x)^2$ , while implicit ODE solvers can be made unconditionally stable. However, the accuracy requirements would limit time-step size and since a large nonlinear algebraic system of equations has to be solved at each time-step, implicit methods may not be efficient. Here, we apply the explicit third-order large stability domain Runge-Kutta method, developed in Medovikov [1998] and Medovikov [n.d.], which allow one to use much larger time-steps compared with the standard explicit Runge-Kutta methods. In practice, when the problem is not too stiff as in the case of ODEs arising in finite-difference approximation of parabolic PDEs, these methods preserve all the advantages of explicit methods and are typically more efficient than implicit methods (see Abdulle [2002], Medovikov [1998], and Verwer, Sommeijer, and Hundsdorfer [2004] for details). We have implemented the code DUMKA3 Medovikov [n.d.], which incorporates the embedded formulas that permit an efficient stepsize control. The efficiency of DUMKA3 is further improved when the user provides an upper bound on the time-step stability restriction for the forward Euler method. Assume that the system of ODEs (10)-(11) is numerically integrated by the forward Euler method from time  $t$  to  $t + \Delta t_{\text{FE}}$  and that the following CFL condition holds:

$$(18) \quad \Delta t_{\text{FE}} \leq \frac{1}{am} \cdot \frac{1}{\max_j (u_x)_j^2}, \quad a := \sum_{p=-m}^m \alpha_p^2,$$

where  $\alpha_p$  are the coefficients in (9) and  $(u_x)_j$  are given by (11). It is shown in [Cheng, Kurganov, Qu, and Tang \[2015\]](#) that

$$(19) \quad \|u(t + \Delta t_{\text{FE}})\|_{L^2} \leq \|u(t)\|_{L^2},$$

where  $\|u(t)\|_{L^2} := \sqrt{\sum_j u_j^2(t) \Delta x}$ .

Similar theoretical results hold for (13)–(15) with the forward Euler method, and for (16)–(17) with the forward Euler method. Note that the code DUMKA3 automatically selects time-steps so that in average the selected time-steps  $\Delta t_{\text{ODE}}$  are much larger than  $\Delta t_{\text{FE}}$ .

#### 4.3 Pseudo-Spectral Methods for (2). We first consider the 1-D equation,

$$(20) \quad u_t = -u_{xx} - \delta u_{xxxx}, \quad x \in [0, L], \quad t \in (0, T],$$

subject to the  $L$ -periodic boundary conditions.

We first use the FFT algorithm to compute the discrete Fourier coefficients  $\{\widehat{u}_m(t)\}$  from the available point values  $\{u_j(t)\}$ . This gives us the following spectral approximation of  $u$  on  $[0, L]$ :

$$(21) \quad u(x, t) \approx \sum_m \widehat{u}_m(t) e^{i \frac{2\pi m x}{L}}.$$

We then substitute (21) into (20) and obtain very simple linear ODEs for the discrete Fourier coefficients of  $u$ ,

$$\frac{d}{dt} \widehat{u}_m(t) = (s - \delta s^2) \widehat{u}_m(t), \quad s = \left( \frac{2\pi m}{L} \right)^2,$$

which can be solved exactly:

$$\widehat{u}_m(t + \Delta t) = e^{(s - \delta s^2) \Delta t} \widehat{u}_m(t).$$

Finally, we use the inverse FFT algorithm to obtain the point values of the solution at the new time level,  $\{u_j(t + \Delta t)\}$ , out of the set of the discrete Fourier coefficients  $\{\widehat{u}_m(t + \Delta t)\}$ .

We now consider the 2-D equation (2),

$$u_t = -(u_{xx} + u_{yy}) - \delta(u_{xxxx} + 2u_{xxyy} + u_{yyyy}),$$

on a rectangular domain  $\Omega = [0, L_x] \times [0, L_y]$  with the  $L_x$ - and  $L_y$ -periodic boundary conditions in the  $x$ - and  $y$ -directions, respectively.

Similar to the 1-D case, we apply the FFT algorithm and obtain very simple linear ODEs for the discrete Fourier coefficients of  $u$ ,

$$(22) \quad \frac{d}{dt} \hat{u}_{m,\ell}(t) = (s - \delta s^2) \hat{u}_{m,\ell}(t), \quad s = \left( \frac{2\pi m}{L_x} \right)^2 + \left( \frac{2\pi \ell}{L_y} \right)^2.$$

The exact solution of (22) is

$$\hat{u}_{m,\ell}(t + \Delta t) = e^{(s - \delta s^2)\Delta t} \hat{u}_{m,\ell}(t).$$

Finally, we apply the inverse FFT algorithm to obtain the point values of the solution at the new time level,  $\{u_{j,k}(t + \Delta t)\}$ , out of the set of the discrete Fourier coefficients  $\{\hat{u}_{m,\ell}(t + \Delta t)\}$ .

As a numerical example, we again consider [Example 3.1](#) and compute its solution on a  $128 \times 128$  uniform grid with the constant splitting step  $\Delta t = 10^{-3}$ . The solution computed at times  $t = 1, 2, 5$  and  $20$  is shown in [Figure 7](#). These results are in good agreement with those reported in [Feng, Tang, and J. Yang \[2015\]](#) and with the SDC result reported in the last section.

We mention that the present operator-splitting approach can be combined with some time-adaptor strategy to speed up numerical simulations, see, e.g., [Cheng, Kurganov, Qu, and Tang \[2015\]](#), [Luo, Tang, and Xie \[2016\]](#), and [Qiao, Z. Zhang, and Tang \[2011\]](#).

We close this section by mentioning that some theoretical study for the operator splitting method outlined above was carried out in [X. Li, Qiao, and H. Zhang \[2017\]](#), where the finite difference scheme for the nonlinear part was improved so that larger time steps are allowed.

## 5 Concluding remarks

There have been considerable recent interests in developing highly stable and efficient numerical schemes for solving phase-field models. In this article, we present three classes of effective time discretization schemes. The first one is based on adding consistent terms so that the energy-decay property is satisfied. Some recent theory for this class of methods is reviewed. The second class is based on the time direction  $p$ -adaptivity, by combining lower-order convex-splitting methods and the SDC technique. It is demonstrated by



numerical experiments that this is a very efficient numerical approach. The third class method is based on the classical operator-splitting method. Some preliminary results show that this is a promising method for practical computations.

## References

- Assyr Abdulle (2002). “Fourth order Chebyshev methods with recurrence relation”. *SIAM J. Sci. Comput.* 23.6, pp. 2041–2054. MR: [1923724](#) (cit. on pp. [3700](#), [3703](#)).
- Andrea L. Bertozzi, Selim Esedoğlu, and Alan Gillette (2007). “Inpainting of binary images using the Cahn-Hilliard equation”. *IEEE Trans. Image Process.* 16.1, pp. 285–291. MR: [2460167](#).
- Andrea L. Bertozzi, Ning Ju, and Hsiang-Wei Lu (2011). “A biharmonic-modified forward time stepping method for fourth order nonlinear diffusion equations”. *Discrete Contin. Dyn. Syst.* 29.4, pp. 1367–1391. MR: [2773188](#) (cit. on p. [3691](#)).
- Yuanzhen Cheng, Alexander Kurganov, Zhuolin Qu, and Tao Tang (2015). “Fast and stable explicit operator splitting methods for phase-field models”. *J. Comput. Phys.* 303, pp. 45–65. MR: [3422699](#) (cit. on pp. [3701](#), [3702](#), [3704](#), [3705](#)).
- Alina Chertock, Alexander Kurganov, and Guergana Petrova (2009). “Fast explicit operator splitting method for convection-diffusion equations”. *Internat. J. Numer. Methods Fluids* 59.3, pp. 309–332. MR: [2484270](#) (cit. on p. [3699](#)).
- Alok Dutt, Leslie Greengard, and Vladimir Rokhlin (2000). “Spectral deferred correction methods for ordinary differential equations”. *BIT* 40.2, pp. 241–266. MR: [1765736](#) (cit. on pp. [3693](#), [3698](#)).
- David J. Eyre (1993). “Systems of Cahn-Hilliard equations”. *SIAM J. Appl. Math.* 53.6, pp. 1686–1712. MR: [1247174](#) (cit. on pp. [3689](#), [3696](#)).
- Xinlong Feng, Tao Tang, and Jiang Yang (2013). “Stabilized Crank-Nicolson/Adams-Bashforth schemes for phase field models”. *East Asian J. Appl. Math.* 3.1, pp. 59–80. MR: [3109557](#) (cit. on p. [3689](#)).
- (2015). “Long time numerical simulations for phase-field problems using  $p$ -adaptive spectral deferred correction methods”. *SIAM J. Sci. Comput.* 37.1, A271–A294. MR: [3304267](#) (cit. on pp. [3689](#), [3695](#), [3696](#), [3698](#), [3705](#)).
- Hector Gomez and Thomas J. R. Hughes (2011). “Provably unconditionally stable, second-order time-accurate, mixed variational methods for phase-field models”. *J. Comput. Phys.* 230.13, pp. 5310–5327. MR: [2799512](#) (cit. on p. [3689](#)).
- Ruihan Guo, Yinhua Xia, and Yan Xu (2017). “Semi-implicit spectral deferred correction methods for highly nonlinear partial differential equations”. *J. Comput. Phys.* 338, pp. 269–284. MR: [3628250](#) (cit. on p. [3695](#)).

- Ruihan Guo and Yan Xu (2016). “Local discontinuous Galerkin method and high order semi-implicit scheme for the phase field crystal equation”. *SIAM J. Sci. Comput.* 38.1, A105–A127. MR: [3439767](#) (cit. on pp. [3695](#), [3698](#)).
- Yinnian He, Yunxian Liu, and Tao Tang (2007). “On large time-stepping methods for the Cahn-Hilliard equation”. *Appl. Numer. Math.* 57.5-7, pp. 616–628. MR: [2322435](#) (cit. on pp. [3689](#)–[3691](#)).
- R. V. Kohn (2006). “Energy-driven pattern formation”. In: *Proceedings of the International Congress of Mathematicians, Madrid, 2006*. European Math. Soc., pp. 359–383 (cit. on p. [3688](#)).
- J. Shen L.Q. Chen (1998). “Applications of semi-implicit Fourier-spectral method to phase field equations”. *Comput. Phys. Comm.* 108, pp. 147–158 (cit. on p. [3689](#)).
- Dong Li and Zhonghua Qiao (2017a). “On second order semi-implicit Fourier spectral methods for 2D Cahn-Hilliard equations”. *J. Sci. Comput.* 70.1, pp. 301–341. MR: [3592143](#) (cit. on p. [3693](#)).
- (2017b). “On the stabilization size of semi-implicit Fourier-spectral methods for 3D Cahn-Hilliard equations”. *Commun. Math. Sci.* 15.6, pp. 1489–1506. MR: [3668944](#) (cit. on p. [3693](#)).
- Dong Li, Zhonghua Qiao, and Tao Tang (2016). “Characterizing the stabilization size for semi-implicit Fourier-spectral method to phase field equations”. *SIAM J. Numer. Anal.* 54.3, pp. 1653–1681. MR: [3507555](#) (cit. on pp. [3691](#), [3693](#)).
- Xiao Li, Zhonghua Qiao, and Hui Zhang (2017). “Convergence of a fast explicit operator splitting method for the epitaxial growth model with slope selection”. *SIAM J. Numer. Anal.* 55.1, pp. 265–285. MR: [3608748](#) (cit. on p. [3705](#)).
- Fuesheng Luo, Tao Tang, and Hehu Xie (2016). “Parameter-free time adaptivity based on energy evolution for the Cahn-Hilliard equation”. *Commun. Comput. Phys.* 19.5, pp. 1542–1563. MR: [3501222](#) (cit. on p. [3705](#)).
- A. A. Medovikov (n.d.). “DUMKA3 code” (cit. on p. [3703](#)).
- Alexei A. Medovikov (1998). “High order explicit methods for parabolic equations”. *BIT* 38.2, pp. 372–390. MR: [1638136](#) (cit. on p. [3703](#)).
- Michael L. Minion (2003). “Semi-implicit spectral deferred correction methods for ordinary differential equations”. *Commun. Math. Sci.* 1.3, pp. 471–500. MR: [2069941](#) (cit. on p. [3698](#)).
- Zhonghua Qiao and Shuyu Sun (2014). “Two-phase fluid simulation using a diffuse interface model with Peng-Robinson equation of state”. *SIAM J. Sci. Comput.* 36.4, B708–B728. MR: [3246906](#) (cit. on p. [3689](#)).
- Zhonghua Qiao, Zhi-Zhong Sun, and Zhengru Zhang (2015). “Stability and convergence of second-order schemes for the nonlinear epitaxial growth model without slope selection”. *Math. Comp.* 84.292, pp. 653–674. MR: [3290959](#) (cit. on p. [3689](#)).

- Zhonghua Qiao, Zhengru Zhang, and Tao Tang (2011). “An adaptive time-stepping strategy for the molecular beam epitaxy models”. *SIAM J. Sci. Comput.* 33.3, pp. 1395–1414. MR: [2813245](#) (cit. on pp. [3692](#), [3705](#)).
- J. Shen, J. Xu, and J. Yang (2017). “A new class of efficient and robust energy stable schemes for gradient flows” (cit. on pp. [3689](#), [3696](#)).
- Jie Shen, Tao Tang, and Li-Lian Wang (2011). *Spectral methods*. Vol. 41. Springer Series in Computational Mathematics. Algorithms, analysis and applications. Springer, Heidelberg, pp. xvi+470. MR: [2867779](#) (cit. on p. [3697](#)).
- Jie Shen, Cheng Wang, Xiaoming Wang, and Steven M. Wise (2012). “Second-order convex splitting schemes for gradient flows with Ehrlich-Schwoebel type energy: application to thin film epitaxy”. *SIAM J. Numer. Anal.* 50.1, pp. 105–125. MR: [2888306](#) (cit. on pp. [3689](#), [3696](#)).
- Jie Shen and Xiaofeng Yang (2010). “Numerical approximations of Allen-Cahn and Cahn-Hilliard equations”. *Discrete Contin. Dyn. Syst.* 28.4, pp. 1669–1691. MR: [2679727](#) (cit. on pp. [3689](#), [3690](#)).
- Huailing Song and Chi-Wang Shu (2017). “Unconditional energy stability analysis of a second order implicit-explicit local discontinuous Galerkin method for the Cahn-Hilliard equation”. *J. Sci. Comput.* 73.2-3, pp. 1178–1203. MR: [3719623](#) (cit. on p. [3693](#)).
- Tao Tang, Hehu Xie, and Xiaobo Yin (2013). “High-order convergence of spectral deferred correction methods on general quadrature nodes”. *J. Sci. Comput.* 56.1, pp. 1–13. MR: [3049939](#) (cit. on p. [3698](#)).
- J. G. Verwer, B. P. Sommeijer, and W. Hundsdorfer (2004). “RKC time-stepping for advection-diffusion-reaction problems”. *J. Comput. Phys.* 201.1, pp. 61–79. MR: [2098853](#) (cit. on p. [3703](#)).
- S. M. Wise, C. Wang, and J. S. Lowengrub (2009). “An energy-stable and convergent finite-difference scheme for the phase field crystal equation”. *SIAM J. Numer. Anal.* 47.3, pp. 2269–2288. MR: [2519603](#) (cit. on p. [3689](#)).
- Yinhua Xia, Yan Xu, and Chi-Wang Shu (2009). “Application of the local discontinuous Galerkin method for the Allen-Cahn/Cahn-Hilliard system”. *Commun. Comput. Phys.* 5.2-4, pp. 821–835. MR: [2513717](#) (cit. on p. [3689](#)).
- Chuanju Xu and Tao Tang (2006). “Stability analysis of large time-stepping methods for epitaxial growth models”. *SIAM J. Numer. Anal.* 44.4, pp. 1759–1779. MR: [2257126](#) (cit. on pp. [3689](#), [3691](#), [3692](#)).
- Kristoffer G. van der Zee, J. Tinsley Oden, Serge Prudhomme, and Andrea Hawkins-Daarud (2011). “Goal-oriented error estimation for Cahn-Hilliard models of binary phase transition”. *Numer. Methods Partial Differential Equations* 27.1, pp. 160–196. MR: [2743604](#) (cit. on p. [3689](#)).

- J. Zhu, L.-Q. Chen, J. Shen, and V. Tikare (1999). “Coarsening kinetics from a variable-mobility Cahn-Hilliard equation: Application of a semi-implicit Fourier spectral method”. *Phys. Rev. E* 60, pp. 3564–3572 (cit. on pp. [3689–3691](#)).

Received 2018-02-26.

TAO TANG (汤涛)

DEPARTMENT OF MATHEMATICS, SOUTHERN UNIVERSITY OF SCIENCE AND TECHNOLOGY, NANSHAN DISTRICT, SHENZHEN, GUANGDONG 518055, CHINA

[tangt@sustc.edu.cn](mailto:tangt@sustc.edu.cn)

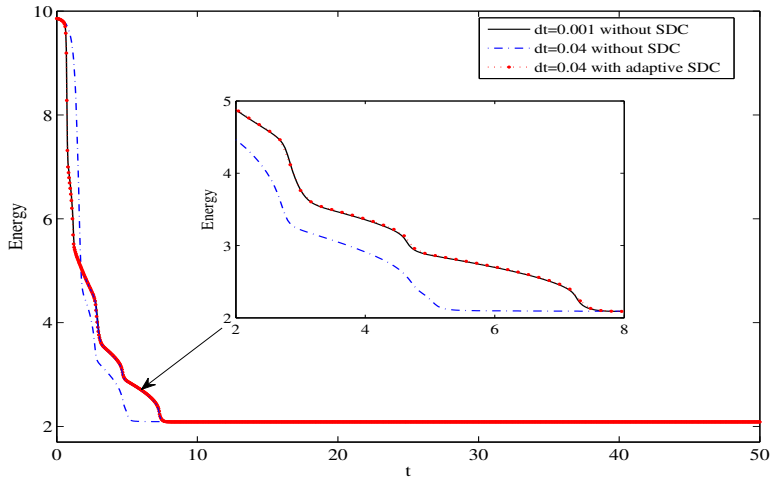


Figure 4: [Example 3.1](#): Energy curves of the Cahn-Hilliard equation by different schemes with different time steps.

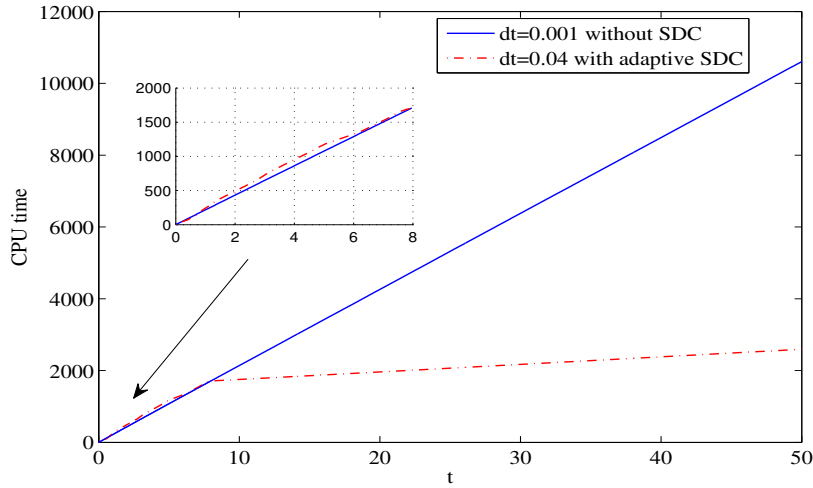


Figure 5: [Example 3.1](#): CPU time comparison between different schemes for Cahn-Hilliard equation.

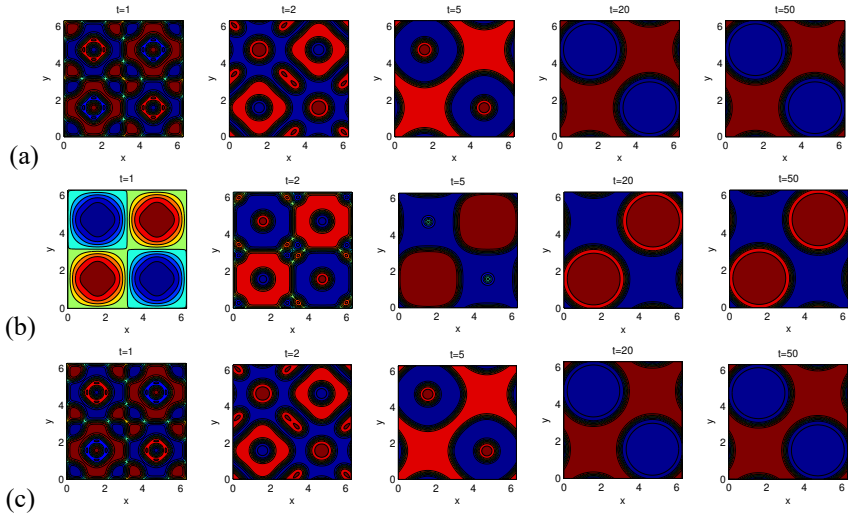


Figure 6: [Example 3.1](#): Solution variation at different time, using (a) direct energy convex splitting scheme without SDC and  $dt = 0.001$ ; (b) direct energy convex splitting scheme without SDC and  $dt = 0.04$ ; and (c) adaptive SDC scheme with  $dt = 0.04$ .

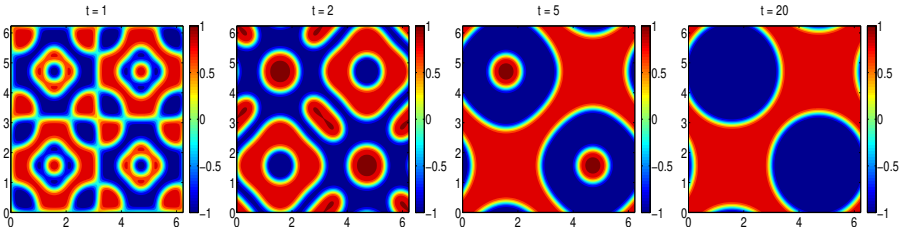


Figure 7: [Example 3.1](#):  $u$  computed with splitting time-stepping with  $\Delta t = 10^{-3}$ .



# FFT BASED SPECTRAL EWALD METHODS AS AN ALTERNATIVE TO FAST MULTIPOLE METHODS

ANNA-KARIN TORNBORG

## Abstract

In this paper, we review a set of fast and spectrally accurate methods for rapid evaluation of three dimensional electrostatic and Stokes potentials. The algorithms use the so-called Ewald decomposition and are FFT-based, which makes them naturally most efficient for the triply periodic case. Two key ideas have allowed efficient extension of these Spectral Ewald (SE) methods to problems with periodicity in only one or two dimensions: an adaptive 3D FFT that apply different upsampling rates locally combined with a new method for FFT based solutions of free space harmonic and biharmonic problems. The latter approach is also used to extend to the free space case, with no periodicity. For the non-radial kernels of Stokes flow, the structure of their Fourier transform is exploited to extend the applicability from the radial harmonic and biharmonic kernels.

A window function is convolved with the point charges to assign values on the FFT grid. Spectral accuracy is attained with a variable number of points in the support of the window function, tuning a shape parameter according to this choice. A new window function, recently introduced in the context of a non-uniform FFT algorithm, allows for further reduction in the computational time as compared to the truncated Gaussians previously used in the SE method.

## 1 Introduction

The direct evaluation of so called  $N$ -body problems yields a computational cost proportional to  $N^2$ . One example of such a problem is the evaluation of an electrostatic potential

---

This work has been supported by the Swedish Research Council under grants no 2011-3178 and 2015-04998 and by the Göran Gustafsson Foundation for Research in Natural Sciences and Medicine. The author gratefully acknowledge this support.

MSC2010: primary 65M80; secondary 65T50, 65R30, 76O07, 65M38.



owing to  $N$  particles at locations  $\mathbf{x}_n, n = 1, \dots, N$  with charges  $q_n$ , at each of the particle locations,

$$(1) \quad \phi^{0P}(\mathbf{x}_m) = \sum_{n=1}^{N'} q_n \frac{1}{|\mathbf{x}_m - \mathbf{x}_n|},$$

where  $N'$  indicates that the term  $m = n$  is excluded from the sum.

Such sums also arise when solving boundary integral equations numerically. Discretizing the following integral over the boundary  $\partial\Omega$  of  $\Omega \subset \mathbb{R}^3$

$$\int_{\partial\Omega} \frac{f(\mathbf{y})}{|\mathbf{x} - \mathbf{y}|} dS_{\mathbf{y}},$$

$q_n$  is for each  $n$  the product of the function  $f$  evaluated at the quadrature point  $\mathbf{x}_n$  and the quadrature weight at that point. Since the integrand is singular at  $\mathbf{x} = \mathbf{y}$ , the point  $\mathbf{x}_m$  can naturally not be included in the sum.

As we use integral equations to solve the Stokes equations, we need instead to evaluate integrals that contain fundamental solutions of Stokes flow, like the Stokeslet

$$(2) \quad S(\mathbf{r}) = \frac{1}{r} \mathbf{I} + \frac{1}{r^3} \mathbf{r} \mathbf{r}, \quad \text{or} \quad S_{j\ell} = \frac{\delta_{j\ell}}{r} + \frac{r_j r_\ell}{r^3} \quad j, \ell = 1, 2, 3,$$

with  $r = |\mathbf{r}|$  and where  $\delta_{j\ell}$  is the Kronecker delta. Another fundamental solution of Stokes flow, the Stresslet, will be introduced later. The discrete sum containing the Stokeslet corresponding to Equation (1) becomes

$$(3) \quad \mathbf{u}^{0P}(\mathbf{x}_m) = \sum_{n=1}^{N'} S(\mathbf{x}_m - \mathbf{x}_n) \mathbf{f}_n, \quad m = 1, \dots, N.$$

In electrostatic calculations, periodic boundary conditions are typically applied to accurately capture properties of a larger aggregate. Periodicity in only one or two directions can be applied for systems with different structures, such as membranes or nanopores. Similarly, it is common to apply periodic boundary conditions in some directions for fluid flows.

Assume that we have  $N$  particles with charge  $q_n$  located at  $\mathbf{x}_n, n = 1, \dots, N$ , in a domain  $\Omega = [0, L_1] \times [0, L_2] \times [0, L_3]$ , where the system is charge neutral, i.e.  $\sum_{n=1}^N q_n \equiv 0$ . The electrostatic potential due to these charges, evaluated at these same locations, is given by the sum

$$(4) \quad \phi^{DP}(\mathbf{x}_m) = \sum_{\mathbf{p} \in P_D} \sum_{n=1}^{N'} \frac{q_n}{|\mathbf{x}_m - \mathbf{x}_n + \mathbf{p}|}, \quad m = 1, \dots, N.$$

The sum over  $\mathbf{p}$  is a periodic replication of the charges, and  $D = 0, 1, 2, 3$  indicates the number of periodic directions. The  $N'_$  indicates that the term ( $n = m, \mathbf{p} = \mathbf{0}$ ) is excluded from the sum. We define

$$(5) \quad \begin{aligned} P_3 &= \{(jL_1, lL_2, pL_3) : (j, l, p) \in \mathbb{Z}^3\}, & P_2 &= \{(jL_1, lL_2, 0) : (j, l) \in \mathbb{Z}^2\}, \\ P_1 &= \{(0, 0, pL_3) : p \in \mathbb{Z}\}, & P_0 &= \{(0, 0, 0)\}. \end{aligned}$$

Here, we have chosen  $x$  and  $y$  as the periodic directions and  $z$  as the free direction in the doubly periodic case (2P), and  $x$  and  $y$  as the free and  $z$  as the periodic direction in the singly periodic case (1P).

In the triply periodic case, the sum given above is only conditionally convergent also for charge neutral systems, and the result will depend on the summation order, as is shown e.g. in the much cited paper by [de Leeuw, Perram, and E. R. Smith \[1980\]](#). This is further discussed more recently by [E. Smith \[2008\]](#). The Ewald summation formula for the triply periodic case was derived by [Ewald \[1921\]](#) in 1921. The resulting formula imposes two choices: a spherical summation order and an assumption that the dielectric constant of the surrounding medium is infinite, i.e. that it is a conductor. This is often referred to as “tin foil” boundary conditions. As shown in [E. Smith \[2008\]](#), charge neutrality is necessary also in the singly and doubly periodic cases for the sums to be convergent, but the results are independent on the summation order.

In the Ewald summation formula [Ewald \[1921\]](#), the potential is computed by splitting the contribution from each charge into a rapidly decaying part and a smooth part which is summed in Fourier space. The Ewald sum for evaluating the potential at a source location  $\mathbf{x}_m, m = 1, \dots, N$  under triply periodic boundary conditions is

$$(6) \quad \begin{aligned} \phi^{3P}(\mathbf{x}_m) &= \sum_{\mathbf{p} \in P_3} \sum_{n=1}^{N'_} q_n \frac{\text{erfc}(\xi |\mathbf{x}_m - \mathbf{x}_n + \mathbf{p}|)}{|\mathbf{x}_m - \mathbf{x}_n + \mathbf{p}|} + \\ &+ \frac{4\pi}{V} \sum_{\mathbf{k} \neq \mathbf{0}} \sum_{n=1}^N q_n \frac{e^{-k^2/4\xi^2}}{k^2} e^{-i\mathbf{k} \cdot (\mathbf{x}_m - \mathbf{x}_n)} - \frac{2\xi}{\sqrt{\pi}} q_m. \end{aligned}$$

Here, the  $N'_$  indicates that the term ( $n = m, \mathbf{p} = \mathbf{0}$ ) is excluded from the real space sum and  $P_3$  is given in [Equation \(5\)](#). The  $\mathbf{k}$ -vectors form the discrete set  $\{2\pi(\frac{n_1}{L_1}, \frac{n_2}{L_2}, \frac{n_3}{L_3}) : (n_1, n_2, n_3) \in \mathbb{Z}^3\}$ ,  $k^2 = |\mathbf{k}|^2$  and  $V = L_1 L_2 L_3$ . Here,  $\xi > 0$  is the decomposition parameter. The result is independent of this parameter, but it controls the relative decay of the real and reciprocal space sums. The last term is the so called self correction term. When evaluating the potential at a charge location, no contribution from this charge itself should be included, and this term is added for this purpose. The Ewald sums for the energy and electrostatic force are easily obtained from the expression for the potential, see e.g. [Deserno and Holm \[1998\]](#).

Ewald sums have also been derived for the doubly and singly periodic cases, see [Grzybowski, Gwózdź, and Bródka \[2000\]](#) and [Porto \[2000\]](#) and references therein. The derivation of the singly periodic sum in [Porto \[ibid.\]](#), however left an integral expression for which no closed form was given, that could later be obtained following [Fripiat, Delhalle, Flamant, and Harris \[2010\]](#). In [Tornberg \[2016\]](#), derivations of the Ewald  $3P$ ,  $2P$  and  $1P$  sums are presented in a unified framework that gives a natural starting point for the design of a fast method.

The Stokeslet sum ([Equation \(3\)](#)) can be extended similarly to [Equation \(4\)](#). Hasimoto derived an Ewald type decomposition for the triply periodic case in 1959 [Hasimoto \[1959\]](#). Instead of charge neutrality, here we have an assumption that there is a mean pressure gradient that balances the net force. [Pozrikidis \[1996\]](#) derived an alternative sum for the Stokes  $3P$  case, using a different decomposition due to [Beenakker \[1986\]](#). He further discussed also the  $2P$  and  $1P$  Stokeslet sums, however not stating all formulas explicitly. Explicit formulas for the  $2P$  Stokeslet sum with the Hasimoto decomposition can be found in [Lindbo and Tornberg \[2011a\]](#).

**1.1 Fast methods and the development of the Spectral Ewald method.** As was noted above, the direct evaluation of the sum in [Equation \(1\)](#) has a computational complexity of  $O(N^2)$  and so does the Ewald sum for the triply periodic case as given in [Equation \(6\)](#), but in addition with a much larger cost for the same value of  $N$ .

For free space problems such as [Equation \(1\)](#), the Fast Multipole Method (FMM) can reduce the  $O(N^2)$  cost to  $O(N)$  work, where the constant multiplying  $N$  will depend on the required accuracy. FMM was first introduced by Greengard and Rokhlin for the harmonic kernel in 2D and later in 3D [Greengard and Rokhlin \[1987\]](#) and [Cheng, Greengard, and Rokhlin \[1999\]](#) and has since been extended to other kernels, including the fundamental solutions of Stokes flow [Tornberg and Greengard \[2008\]](#) and [Wang, Lei, Li, Huang, and Yao \[2007\]](#). The FMM has not been as popular for periodic problems, even if it can be extended to this case at an additional cost, see e.g. [Gumerov and Duraiswami \[2014\]](#) and the references therein.

For triply periodic problems in electrostatics, FFT-based methods have been the most popular and successfully used since the early 1990s. Here, the Ewald decomposition is used, with  $\xi$  in [Equation \(6\)](#) chosen such that the real space terms decay rapidly, and more work is put into the Fourier sum, which is accelerated with an FFT based method. With a proper scaling of  $\xi$  as  $N$  grows, the full algorithm yields a cost of  $O(N \log N)$ . One early method for evaluation of the electrostatic potential and force was the Particle Mesh Ewald (PME) method by [Darden, York, and Pedersen \[1993\]](#), later refined to the Smooth Particle Mesh Ewald (SPME) method by [Essmann, Perera, Berkowitz, Darden, Lee, and Pedersen \[1995\]](#). See also the survey by [Deserno and Holm \[1998\]](#). The SPME method

was extended to the fast evaluation of the triply periodic stokeslet sum by [Saintillan, Darve, and Shaqfeh \[2005\]](#).

**The Spectral Ewald method.** The Spectral Ewald (SE) method was first introduced for the triply periodic stokeslet sum in [Lindbo and Tornberg \[2010\]](#), and soon thereafter for the electrostatic problem [Lindbo and Tornberg \[2011b\]](#). The PME methods mentioned above have a polynomial order of accuracy, and require a refinement of the FFT grid to reduce approximation errors. Specifically, in the SPME method, the point sources are convolved with B-splines of a fixed regularity and support to assign values on the FFT grid. In contrast, the SE method as it was introduced in [Lindbo and Tornberg \[2010, 2011b\]](#), is spectrally accurate. By using suitably scaled and truncated Gaussians, the approximation error is reduced spectrally fast as the number of points in the support of the truncated Gaussians is increased, and is not tied to the grid size. The idea of using Gaussians as window functions was of course not new, not in PME like methods, nor in the closely related non-uniform FFT methods, as discussed in [Lindbo and Tornberg \[2010, 2011b\]](#). The key in the performance of the SE method was to tie a shape parameter of the Gaussian to the number of points in its support in order to minimize approximation errors.

Recently, we compared the use of Gaussians to a window function that was recently introduced by Barnett and Magland in connection to a non-uniform FFT algorithm [Barnett and Magland \[2017\]](#). This window function is an approximation to the Kaiser-Bessel function that retains the desirable properties while reducing the cost of evaluation. Similarly to the Gaussians, we adjust a shape parameter for this window function with the number of points in the support. In [Saffar Shamshirgar and Tornberg \[2017a\]](#), we showed that this new Barnett-Magland window function is superior to the Gaussian and that the computational cost is further reduced for the same target accuracy.

FFT based methods are most efficient for the triply periodic case. In this case, FFTs can be used in all directions without any oversampling. As soon as there is a non-periodic direction, the grid has to be extended in that direction. In the doubly periodic case, [De Joannis, Arnold, and Holm \[2002\]](#) devised a method where the problem is extended to full periodicity, with a larger length in the non-periodic direction, and where a correction term is applied to improve on the result. Here, the increased length in the non-periodic direction simply means a zero-padding of the FFT, increasing in the number of grid points in that direction. The SE2P method by [Lindbo and Tornberg \[2012\]](#) takes a different approach, which needs a “mixed” transform; a discrete Fourier transform in the periodic variables and an approximation to the continuous Fourier integral transform in the free dimension. Also in this case the grid in the free dimension must be oversampled for an accurate approximation.

In the doubly periodic Ewald sum, there is a term that includes the contribution from the zero wave number in the periodic directions, i.e., that depends only on the variable in the free direction. An expansion based on Chebyshev polynomials offered an efficient evaluation if this 1D sum [Lindbo and Tornberg \[2012\]](#).

**Recent developments.** There were two main challenges to overcome when extending the Spectral Ewald method to the singly periodic (1P) case. With two free dimensions, an oversampling factor of four to six in each would increase the cost of FFTs by a factor of 16 to 64, which is clearly not desirable. Furthermore, the zero wave number in the periodic direction here yields a 2D sum as opposed to a 1D sum in the doubly periodic case, and it is not feasible to extend the approach in [Lindbo and Tornberg \[ibid.\]](#).

In [Saffar Shamshirgar and Tornberg \[2017b\]](#) we showed that it is sufficient to upsample only for small discrete wave numbers, and introduced an adaptive FFT and IFFT (denoted by AFT and AIFT) that only upsample for a select number of discrete modes in the periodic direction. As for the second challenge, the 2D sum is the free space solution to a 2D Poisson problem, and a recent idea for how to solve free space problems by the means of FFTs [Vico, Greengard, and Ferrando \[2016\]](#) can therefore be used. The treatment of the zero periodic wave number can now be treated in the same framework as the other modes, and will be included in the AFT mentioned above. This is done at a negligible extra cost. A typical increase in cost of the FFTs performed in the 1P case as compared to 3P is a factor of 2 – 3. The gridding cost when applying the window function is essentially the same in both cases. The ratio of the total runtime cost for the SE1P method and the SE3P method is therefore even smaller.

In [af Klinteberg, Saffar Shamshirgar, and Tornberg \[2017\]](#), the approach to solve free space problems with FFTs was used to extend the SE method to problems without periodicity. The original idea in [Vico, Greengard, and Ferrando \[2016\]](#) is applicable for the harmonic and biharmonic kernels, here an extension was introduced such that sum of free space potentials could be evaluated for stokeslets, stresslets and rotlets.

We have very recently unified the treatment from free space up to triply periodic for the electrostatic problem [Saffar Shamshirgar and Tornberg \[2017a\]](#). The 2P algorithm from [Lindbo and Tornberg \[2012\]](#) was here modified to make use of the advances made when developing the 1P method [Saffar Shamshirgar and Tornberg \[2017b\]](#). The software is available [Lindbo, af Klinteberg, and Saffar Shamshirgar \[2016\]](#), including also the implementation of the new window function [Barnett and Magland \[2017\]](#).

Recently, [Nestler, Pippig, and Potts \[2015\]](#) developed an FFT based fast algorithm based on Ewald decomposition for triply, double and singly periodic problems. To the best of our knowledge, this is the only Ewald method with  $O(N \log(N))$  complexity for singly periodic problems except our own. Their approach is however quite different as

compared to ours, as instead of discretizing the continuous Fourier transforms, they work with the analytical formulas containing special functions that are obtained from them.

Any method based on Ewald summation and acceleration by FFTs will be most efficient in the triply periodic case. As soon as there is one or more non-periodic directions, there will be a need for some oversampling of FFTs, which will increase the computational cost. For the fast multipole method (FMM), the opposite is true. The free space problem is the fastest to compute, and any periodicity will invoke an additional cost, which will become substantial or even overwhelming if the base periodic box has a large aspect ratio. Hence, implementing the FFT-based Spectral Ewald method for a free-space problem and comparing it to an FMM method will be the worst possible case for the SE method. Still, we did so for the free space summation of Stokes potentials in [af Klinteberg, Safar Shamshirgar, and Tornberg \[2017\]](#), using an open source implementation of the FMM [Greengard \[2012\]](#). It turned out that our SE method was competitive and often performed clearly better than the FMM (one can, however, expect this adaptive FMM to perform better for highly non-uniform point distributions).

**Outline.** The structure of this review is as follows: in [Section 2](#) we discuss the derivation of the Ewald sums, and highlight the differences that occur due to different periodicities. We also introduce modifications based on the ideas in [Vico, Greengard, and Ferrando \[2016\]](#) to get formulas on a form amenable to numerical treatment for all Fourier modes. In [Section 2](#), we introduce the triply periodic SE method, and discuss all the steps of the algorithm. This is the simplest case, and the extension of the Spectral Ewald method to different periodicities is discussed in the following section. These sections are all concerned with the evaluation of the electrostatic potential, and in [Section 5](#) we discuss the extension to potentials of Stokes flow, before we summarize and conclude.

## 2 Ewald formulas for electrostatics

There is more than one way to derive the Ewald summation formula. One can e.g. utilize the fact that the electrostatic potential can be found as the solution to the Poisson equation

$$(7) \quad -\Delta\phi = 4\pi f^{DP}(\mathbf{x}), \quad f^{DP}(\mathbf{x}) = \sum_{\mathbf{p} \in P_D} \sum_n q_n \delta(\mathbf{x} - \mathbf{x}_n + \mathbf{p}), \quad \mathbf{x} \in \mathbb{R}^3.$$

The sum over  $\mathbf{p}$  is a replication of the charges in the periodic directions, and  $D = 0, 1, 2, 3$  indicates the number of periodic directions with  $P_D$  defined in [Equation \(5\)](#). We introduce a charge screening function,  $\gamma(\xi, \mathbf{x})$  to decompose  $f^{DP}$  into two parts:

$$f^{DP}(\mathbf{x}) = \underbrace{f^{DP}(\mathbf{x}) - (f^{DP} * \gamma)(\mathbf{x})}_{:= f^{DP,R}(\xi, \mathbf{x})} + \underbrace{(f^{DP} * \gamma)(\mathbf{x})}_{:= f^{DP,F}(\xi, \mathbf{x})}.$$

The Poisson equation can be solved for each of the two parts of the right hand side to find  $\phi^{PD,R}$  and  $\phi^{PD,F}$ , that can then be added. The screening function for which the classical Ewald decomposition is obtained is a Gaussian  $\gamma(\xi, \mathbf{x})$ , with the Fourier transform  $\widehat{\gamma}(\xi, \mathbf{k})$ ,  $\xi > 0$ ,

$$(8) \quad \gamma(\xi, \mathbf{x}) = \xi^3 \pi^{-3/2} e^{-\xi^2 |\mathbf{x}|^2}, \quad \widehat{\gamma}(\xi, \mathbf{k}) = e^{-|\mathbf{k}|^2 / 4\xi^2}.$$

The function  $f^{DP,F}(\xi, \mathbf{x})$  is smooth, and a Fourier representation of the solution  $\phi^{PD,F}$  will hence converge rapidly.

The Ewald sum for evaluating the potential at a source location  $\mathbf{x}_m$ ,  $m = 1, \dots, N$  under different periodicity conditions becomes

$$(9) \quad \phi^{DP}(\mathbf{x}_m) = \phi^{DP,R}(\mathbf{x}_m, \xi) + \phi^{DP,F}(\mathbf{x}_m, \xi) - \frac{2\xi}{\sqrt{\pi}} q_m,$$

where

$$(10) \quad \phi^{DP,R}(\mathbf{x}_m, \xi) = \sum_{\mathbf{p} \in P_D} \sum_{n=1}^{N'} q_n \frac{\text{erfc}(\xi |\mathbf{x}_m - \mathbf{x}_n + \mathbf{p}|)}{|\mathbf{x}_m - \mathbf{x}_n + \mathbf{p}|}, \quad D = 0, 1, 2, 3.$$

This term can be derived by evaluating a convolution integral of  $\gamma(\xi, \mathbf{x} - \mathbf{x}_n)$  with the harmonic Green's function (see e.g. Appendix A of [Tornberg \[2016\]](#)), then summing over all sources including periodic copies. Here, the  $N'$  indicates that the term ( $n = m$ ,  $\mathbf{p} = \mathbf{0}$ ) is excluded from the real space sum and  $P_D$  is given in [Equation \(5\)](#). The last term in [Equation \(9\)](#) is the so called self correction term. When evaluating the potential at a charge location, no contribution from this charge itself should be included, and this term is added for this purpose.

For the Fourier space contribution, it remains to solve

$$(11) \quad -\Delta \phi^{DP,F} = 4\pi f^{DP,F}(\mathbf{x}, \xi), \quad f^{DP,F}(\mathbf{x}, \xi) = \sum_{\mathbf{p} \in P_D} \sum_n q_n \gamma(\xi, \mathbf{x} - \mathbf{x}_n + \mathbf{p}), \quad \mathbf{x} \in \mathbb{R}^3.$$

with  $\gamma(\xi, \mathbf{x})$  as defined in [Equation \(8\)](#), under appropriate boundary conditions.

Expanding  $\phi^{3P,F}(\mathbf{x}, \xi)$  in a triply periodic Fourier sum, and using the expression for  $\widehat{\gamma}(\xi, \mathbf{k})$  from [Equation \(8\)](#) to do the same for  $f^{3P,F}(\mathbf{x}, \xi)$ , we can solve [Equation \(11\)](#) and obtain

$$(12) \quad \phi^{3P,F}(\mathbf{x}_m, \xi) = \frac{4\pi}{V} \sum_{\mathbf{k} \neq \mathbf{0}} \sum_{n=1}^N q_n \frac{e^{-k^2 / 4\xi^2}}{k^2} e^{-i\mathbf{k} \cdot (\mathbf{x}_m - \mathbf{x}_n)}.$$

which is the summation over  $\mathbf{k}$  in the second line of [Equation \(6\)](#).

For the doubly periodic case, we can expand  $\phi^{2P,F}(\mathbf{x}, \xi)$  and  $f^{2P,F}(\mathbf{x}, \xi)$  in Fourier series in the periodic  $x$  and  $y$  directions. The Fourier coefficients  $\hat{\phi}_{\bar{\mathbf{k}}}(z)$  and  $\hat{f}_{\bar{\mathbf{k}}}(z)$  will be indexed by  $\bar{\mathbf{k}} = (k_1, k_2)$ . These coefficients can be represented in terms of a Fourier transform in the non-periodic  $z$ -direction. Alternatively, we can insert these doubly periodic Fourier series into [Equation \(11\)](#), use orthogonality and for each wave vector  $\bar{\mathbf{k}}$  obtain

$$(-\partial_z^2 + |\bar{\mathbf{k}}|^2)\hat{\phi}_{\bar{\mathbf{k}}}(z) = 4\pi \hat{f}_{\bar{\mathbf{k}}}(z).$$

In light of how we will later proceed with constructing a fast method to evaluate  $\phi^{2P,F}(\mathbf{x}, \xi)$ , we take the first view point for  $\bar{\mathbf{k}} \neq 0$ , and the second for  $\bar{\mathbf{k}} = 0$ . We write

$$(13) \quad \phi^{2P,F}(\mathbf{x}, \xi) = \bar{\phi}^{2P,F}(\mathbf{x}, \xi) + \phi_0^{2P,F}(z, \xi),$$

with

$$(14) \quad \bar{\phi}^{2P,F}(\mathbf{x}, \xi) = \frac{2}{L_1 L_2} \sum_{\bar{\mathbf{k}} \neq 0} \sum_{n=1}^N q_n \int_{\mathbb{R}} \frac{1}{k^2} e^{-k^2/4\xi^2} e^{-i\bar{\mathbf{k}} \cdot (\mathbf{x} - \mathbf{x}_n)} d\kappa_3.$$

Here, we use  $\mathbf{k} = (k_1, k_2, \kappa_3)$  to emphasize that  $\kappa_3$  is a continuous variable. The term for  $\bar{\mathbf{k}} = 0$ , i.e.  $k_1 = k_2 = 0$  is given by the free space solution of the 1D Poisson equation

$$(15) \quad -\frac{d^2}{dz^2} \phi_0^{2P,F}(z, \xi) = 4\pi \frac{\xi \pi^{-1/2}}{L_1 L_2} \sum_{n=1}^N q_n e^{-\xi^2 |z - z_n|^2}.$$

The integral in [Equation \(14\)](#) can be evaluated analytically, and the [Equation \(15\)](#) has an explicit solution. The result is stated e.g. in section 9 of [Tornberg \[ibid.\]](#). Those formulas are however only used for validation of the fast method. To develop the fast method, we will continue along a different path.

In the singly periodic case, we can similarly to the double periodic case expand in a Fourier series in the periodic direction  $z$ , and index coefficients by  $k_3$ . Also, in this case, we use the continuous Fourier transform to express the coefficients as long as  $k_3 \neq 0$ , and formulate the PDE for the  $k_3 = 0$  coefficient. The explicit formulas can be found in [Saffar Shamshirgar and Tornberg \[2017b\]](#).

In the doubly periodic case, we need to solve a one-dimensional free space problem ([Equation \(15\)](#)), and in the singly periodic case a two-dimensional free space problem. For the free space case, the full problem is a three dimensional free space problem, i.e. [Equation \(11\)](#) for  $D = 0$  under the boundary conditions  $\phi^{0P,F}(\mathbf{x}, \xi) \rightarrow 0$  as  $|\mathbf{x}| \rightarrow 0$ .

This solution can be expressed as a 3D Fourier integral in  $k$ -space, with a  $1/k^2$  factor. It is integrable, and a change to spherical coordinates will for example remove the singularity.



The integral can however not be accurately approximated with values on a regular grid, which is needed for a fast treatment with FFTs.

**2.1 Free space formulas with truncated Green's functions.** Assume that we want to solve

$$(16) \quad -\Delta\phi = 4\pi f(\mathbf{x}), \quad \mathbf{x} \in \mathbb{R}^{\dim}.$$

with free space boundary conditions ( $\phi \rightarrow 0$  as  $|\mathbf{x}| \rightarrow 0$ ), and  $\dim = 1, 2$  or  $3$ .

Assume now that  $f(\mathbf{x})$  is compactly supported within a domain  $\tilde{\mathfrak{D}}$ , a box with sides  $\tilde{\mathbf{L}}$ ,  $\tilde{\mathfrak{D}} = \{\mathbf{x} \mid x_i \in [0, \tilde{L}_i]\}$ , and that we seek the solution for  $\mathbf{x} \in \tilde{\mathfrak{D}}$ . The largest point-to-point distance in the domain is  $|\tilde{\mathbf{L}}|$ . Let  $\mathfrak{R} \geq |\tilde{\mathbf{L}}|$ . Without changing the solution, we can then replace the Green's function with a truncated version

$$G^{\mathfrak{R}}(r) = G(r) \text{rect}\left(\frac{r}{2\mathfrak{R}}\right), \quad \text{rect}(\mathbf{x}) = \begin{cases} 1 & \text{for } |\mathbf{x}| \leq 1/2, \\ 0 & \text{for } |\mathbf{x}| > 1/2. \end{cases}$$

The Fourier transform of the truncated Green's function in 3D, where  $G(r) = 1/r$  is [Vico, Greengard, and Ferrando \[2016\]](#)

$$\hat{G}^{\mathfrak{R}}(k) = 8\pi \left( \frac{\sin(\mathfrak{R}k/2)}{k} \right)^2,$$

with the well defined limit

$$\hat{G}^{\mathfrak{R}}(0) = \lim_{k \rightarrow 0} \hat{G}^{\mathfrak{R}}(k) = 2\pi\mathfrak{R}^2.$$

We then have

$$(17) \quad \begin{aligned} \phi(\mathbf{x}) &= \int_{\mathbb{R}^3} G(|\mathbf{x} - \mathbf{y}|) f(\mathbf{y}) d\mathbf{y} = \frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} \hat{G}(k) \hat{f}(\mathbf{k}) e^{i\mathbf{k} \cdot \mathbf{x}} d\mathbf{k} \\ &= \frac{1}{(2\pi)^3} \int_{\mathbb{R}^3} \hat{G}^{\mathfrak{R}}(k) \hat{f}(\mathbf{k}) e^{i\mathbf{k} \cdot \mathbf{x}} d\mathbf{k}, \end{aligned}$$

as long as the assumption introduced above is fulfilled s.t.  $\text{supp}(f) \in \tilde{\mathfrak{D}}$ ,  $\mathbf{x} \in \tilde{\mathfrak{D}}$  and  $\mathfrak{R}$  is chosen appropriately. Similar formulas can be derived in one and two dimensions.

**2.2 Formulas for the k-space contributions.** The decomposition of  $\phi^{DP}$  into a real space part  $\phi^{DP,R}$  and a Fourier space part  $\phi^{DP,F}$  was introduced in [Equation \(10\)](#). The terms in the real space part have the same form independent of the periodicity of the problem, only the summation over periodic images changes, as given in [Equation \(10\)](#). For the

$k$ -space contribution, we need to work with Fourier series in any periodic direction, and Fourier integrals elsewhere. The simplest case is hence the triply periodic case, where a discrete sum is obtained directly, and the result was given in Equation (12).

For the cases with mixed periodicity, there will be a sum over the discrete Fourier modes in the periodic direction(s). For the Fourier mode where the discrete wave number/wave vector is zero, the integral has a singularity. For this zero mode, we instead define the contribution as the solution to a free space problem. In the previous section, we discussed how to work with truncated Green's functions to obtain an expression in Fourier space for a mollified Green's function that has no singularity. Similarly, for the 0P case, we solve a 3D free space problem, with the right hand side given in Equation (11).

This treatment introduces no approximation under the assumption of a compactly supported right hand side. Our right hand sides are however sums of Gaussians and are not compactly supported. The Gaussians do however decay exponentially fast, and as we construct a numerical algorithm based on these formulas, this error source can be controlled and made vanishingly small by the choice of  $\mathcal{R}$ .

With  $\mathbf{k} = (k_1, k_2, \kappa_3)$ ,  $k = |\mathbf{k}|$ , we write for the 2P case,

$$(18) \quad \phi^{2P,F}(\mathbf{x}, \xi) \approx \frac{2}{L_1 L_2} \sum_{k_1, k_2} \sum_{n=1}^N q_n \int_{\mathbb{R}} \hat{G}_{\mathcal{R}}^{2P}(\mathbf{k}) e^{-k^2/4\xi^2} e^{-i\mathbf{k} \cdot (\mathbf{x} - \mathbf{x}_n)} d\kappa_3.$$

where

$$(19) \quad \hat{G}_{\mathcal{R}}^{2P}(\mathbf{k}) = \begin{cases} 1/k^2 & k_1^2 + k_2^2 \neq 0 \\ (\mathcal{R}k \sin(\mathcal{R}k) + \cos(\mathcal{R}k) - 1)/k, & k_1 = 0, k_2 = 0, \kappa_3 \neq 0 \\ \mathcal{R}^2/2 & \mathbf{k} = 0. \end{cases}$$

With  $\mathbf{k} = (\kappa_1, \kappa_2, k_3)$ ,  $k = |\mathbf{k}|$ , for the 1P case we write

$$(20) \quad \phi^{1P,F}(\mathbf{x}, \xi) \approx \frac{1}{\pi L_3} \sum_{k_3} \sum_{n=1}^N q_n \int_{\mathbb{R}^2} \hat{G}_{\mathcal{R}}^{1P}(\mathbf{k}) e^{-k^2/4\xi^2} e^{-i\mathbf{k} \cdot (\mathbf{x} - \mathbf{x}_n)} d\kappa_1 d\kappa_2,$$

where

$$(21) \quad \hat{G}_{\mathcal{R}}^{1P}(\mathbf{k}) = \begin{cases} 1/k^2 & k_3 \neq 0 \\ (1 - J_0(\mathcal{R}k))/k^2 - \mathcal{R} \log(\mathcal{R}) J_1(\mathcal{R}k)/k & k_3 = 0, \kappa_1^2 + \kappa_2^2 \neq 0 \\ \mathcal{R}^2(1 - 2 \log(\mathcal{R}))/4. & \mathbf{k} = 0. \end{cases}$$

In the above,  $k_i \in \{2\pi n/L_i, n \in \mathbb{Z}\}$ . For the free space case, we have no discrete modes, and use  $\mathbf{k} = (\kappa_1, \kappa_2, \kappa_3)$ ,  $k = |\mathbf{k}|$ , as we write

$$(22) \quad \phi^{0P,F}(\mathbf{x}, \xi) \approx \frac{1}{2\pi^2} \sum_{n=1}^N q_n \int_{\mathbb{R}^3} \hat{G}_{\mathcal{R}}^{0P}(k) e^{-k^2/4\xi^2} e^{-i\mathbf{k} \cdot (\mathbf{x} - \mathbf{x}_n)} d\kappa_1 d\kappa_2 d\kappa_3,$$

where

$$(23) \quad \hat{G}_{\mathfrak{R}}^{0P}(k) = \begin{cases} 2 \sin^2(\mathfrak{R}k/2)/k^2, & k \neq 0 \\ \mathfrak{R}^2/2 & k = 0. \end{cases}$$

Here,  $\hat{G}_{\mathfrak{R}}^{0P}$  is scaled with a factor of  $1/(4\pi)$  as compared to  $\hat{G}^{\mathfrak{R}}$  introduced in the previous section. Again, we want to emphasize that the  $\approx$  sign in these equations arise due to the fact that we formally do not have compactly supported right hand sides for the free space problems. In practice, if the domain of support is set such that the Gaussians are sufficiently decayed, and the parameter  $\mathfrak{R}$  is chosen according to this, there will be no noticeable errors from this source in the FFT based algorithm that we will develop based on these formulas.

**2.3 Truncation errors.** Both the real space and  $k$ -space sums need to be truncated. They decay exponentially fast, and can for the triply periodic case be truncated such as to only include terms for which  $|\mathbf{x}_m - \mathbf{x}_n + \mathbf{p}| < r_c$  in Equation (10) and  $k = |\mathbf{k}| < 2\pi k_\infty/L$  (assuming  $L_i = L$ ,  $i = 1, 2, 3$  for a simpler expression. Excellent error estimates were derived by [Kolafa and Perram \[1992\]](#), and given  $\xi$  and an error tolerance,  $r_c$  and  $k_\infty$  can be appropriately chosen. See also [Lindbo and Tornberg \[2011b\]](#). Even though the error estimates were derived for the triply periodic case, they work remarkably well also for the singly and doubly periodic cases [Lindbo and Tornberg \[2012\]](#) and [Saffar Shamshirgar and Tornberg \[2017b\]](#) and even for the free space case [af Klinteberg, Saffar Shamshirgar, and Tornberg \[2017\]](#) (see the discussion on the rotlet). For the Fourier space contribution this means that the discretized integrals are truncated at the corresponding  $k_\infty$ .

### 3 The spectral Ewald method with full periodicity

As was just discussed, contributions to the real space sum will be ignored if the distance between the source location and the target evaluation point is larger than a cut-off radius  $r_c$ . Typically, a linked cell list or a Verlet list algorithm can be used to efficiently obtain a list of nearest neighbors [Lindbo and Tornberg \[2011b\]](#). The real space sum in Equation (10) includes a summation over the periodic dimensions, which means that contribution from periodic images of the sources are also included if they are within this distance. We will now proceed to discuss the evaluation of the Fourier space sum (Equation (12)) in the triply periodic case.

**3.1 Formulas and algorithmic steps.** The triply periodic case is the most straight forward, and also computationally most efficient since periodicity is naturally handled by FFTs. The fast method that we propose follows the structure of methods within the PME

family: point charges are distributed on a uniform grid using an interpolation (window) function, an FFT is applied, a multiplication is made in  $k$ -space with an appropriately modified Green's function (depending on the choice of window function), an inverse FFT is applied, and the window function is used once more to evaluate the result at irregular evaluation (target) points.

Let us denote the window function by  $\mathcal{W}(\mathbf{x})$ , and assume  $\mathcal{W}(-\mathbf{x}) = \mathcal{W}(\mathbf{x})$ . Note the trivial identity  $\hat{\mathcal{W}}_{\mathbf{k}} \hat{\mathcal{W}}_{\mathbf{k}} \hat{\mathcal{W}}_{\mathbf{k}}^{-2} \equiv 1$  and introduce

$$(24) \quad \hat{\hat{H}}_{\mathbf{k}} = \frac{e^{-k^2/4\xi^2}}{k^2} \hat{\mathcal{W}}_{\mathbf{k}}^{-2} \hat{H}_{\mathbf{k}}$$

where

$$(25) \quad \hat{H}_{\mathbf{k}} = \sum_{n=1}^N q_n \hat{\mathcal{W}}_{\mathbf{k}} e^{-i\mathbf{k}\mathbf{x}_n}.$$

With this, the expression for  $\phi^{3P,F}(\mathbf{x}_m, \xi)$  in Equation (12) becomes

$$(26) \quad \phi^{3P,F}(\mathbf{x}_m, \xi) = \frac{4\pi}{V} \sum_{\mathbf{k} \neq \mathbf{0}} \hat{\mathcal{W}}_{\mathbf{k}} \hat{\hat{H}}_{-\mathbf{k}} e^{-i\mathbf{k}\mathbf{x}_m}.$$

The fact that a product in Fourier space is equal to a convolution in real space implies that  $H(\mathbf{x})$  is given by

$$(27) \quad H(\mathbf{x}) = \sum_{n=1}^N q_n \int_{\Omega} \delta(\mathbf{x} - \mathbf{x}_n) \mathcal{W}(\mathbf{y} - \mathbf{x})_* d\mathbf{y} = \sum_{n=1}^N q_n \mathcal{W}(\mathbf{x} - \mathbf{x}_n)_*,$$

where  $\mathcal{W}(\mathbf{x})_* = \sum_{\mathbf{p} \in P_D} \mathcal{W}(\mathbf{x} + \mathbf{p})$ .

Furthermore, Parseval's formula yields

$$(28) \quad \begin{aligned} \phi^{3P,F}(\mathbf{x}_m, \xi) &= 4\pi \int_{\Omega} \widetilde{H}(\mathbf{x}) \left[ \int_{\Omega} \delta(\mathbf{y} - \mathbf{x}_m) \mathcal{W}(\mathbf{y} - \mathbf{x})_* d\mathbf{y} \right] d\mathbf{x} \\ &= 4\pi \int_{\Omega} \widetilde{H}(\mathbf{x}) \mathcal{W}(\mathbf{x} - \mathbf{x}_m)_* d\mathbf{x}, \end{aligned}$$

where we have suppressed the dependence on  $\xi$  in the notation for  $\widetilde{H}(\mathbf{x})$ .

For simplicity, we will in the following assume that  $L_1 = L_2 = L_3 = L$  such that the periodic domain as defined above (Equation (4)) is  $\Omega = [0, L]^3$ .

1. Introduce a uniform grid over  $\Omega$  of size  $M^3$  and evaluate  $H(\mathbf{x})$  on this grid using Equation (27).

2. Apply an FFT to evaluate  $\hat{H}$ .
3. Evaluate  $\hat{H}_{\mathbf{k}}$  according to Equation (24).
4. Apply an IFFT to evaluate  $\tilde{H}(\mathbf{x})$  on the uniform grid.
5. Evaluate the integral (Equation (28)) with the trapezoidal rule to arrive at the final result,  $\phi^{3P,F}(\mathbf{x}_m, \xi)$ .

There are two sources of errors. Truncation errors arise as only a finite number of Fourier modes are included. Given an error tolerance, the grid size  $M$  is chosen from the truncation error estimate as  $M = 2k_\infty$ . Approximation errors enter due to the approximation of the integral in Equation (28). We will discuss window functions that do not have compact support, and hence, truncation of the window function will also contribute to the approximation error.

**3.2 Window functions and approximation errors.** In the Spectral Ewald method as presented in e.g. Lindbo and Tornberg [2011b], truncated Gaussians have been used as window functions. Here, we use  $\mathcal{W}(\mathbf{x}) = g(\mathbf{x}, \xi, \eta)$ , where

$$(29) \quad g(\mathbf{x}, \xi, \eta) = \left( \frac{2\xi^2}{\pi\eta} \right) e^{-2\xi^2|\mathbf{x}|^2/\eta}$$

The function has been normalized to 1. The Fourier transform is known,  $\hat{g}(\mathbf{k}, \xi, \eta) = e^{-\eta|\mathbf{k}|^2/8\xi^2}$ . With this choice of window function, the scaling step in Equation (24) becomes

$$\hat{H}_{\mathbf{k}} = \frac{e^{-(1-\eta)k^2/4\xi^2}}{k^2} \hat{H}_{\mathbf{k}}.$$

This relation to the Gaussian factor in the Ewald formula is the reason for defining  $g$  as in Equation (29) with both  $\xi$  and the new shape parameter  $\eta$ .

Another class of window functions that has been commonly used is cardinal B-splines Essmann, Perera, Berkowitz, Darden, Lee, and Pedersen [1995] and Deserno and Holm [1998]. The degree of the B-spline is chosen, which gives a fixed (small) support size, and a certain regularity. If the FFT grid has a grid size  $h = L/M$ , an approximation error of  $O(h^p)$  will be introduced, where  $p$  depends on the regularity. Hence, to decrease the approximation error, the grid size  $M$  must be increased.

For the Gaussian window function, we truncate at  $|\mathbf{x}| = |\mathbf{y}| = |\mathbf{z}| = w$ , where  $2w = Ph$  such that we have  $P^3$  points in the support. With  $\eta = (2w\xi/m)^2$ , we can show Lindbo and Tornberg [2011b] that the error committed in approximating Equation (28) by the trapezoidal rule can be bounded by

$$(30) \quad C \left( e^{-\pi^2 P^2/(2m)^2} + \operatorname{erfc}(m/\sqrt{2}) \right).$$

The first term is the quadrature error, and the second term is due to the truncation of the Gaussians. With  $m = c\sqrt{\pi P}$  (where  $c = 1$  found close to optimal for electrostatics), we obtain an exponential decay of the error with  $P$ .

Hence, for any given  $P$ , we scale the window function to achieve the optimal balance between resolution and truncation. We do not need to increase the grid size to reduce the approximation errors - we instead increase  $P$  and scale the window function properly. This allows for the grid size to be selected solely according to the Kolafa-Perram estimate for the truncation of the Fourier Ewald sum.

Recently, Barnett and Magland introduced a new window function in their work in the non-uniform FFT method [Barnett and Magland \[2017\]](#). This new window function is an approximation of the so called Kaiser-Bessel function, which can be shown to yield low error levels but is expensive to compute. To use this window function, we set  $\mathcal{W}(\mathbf{x}) = B(x, \beta)B(y, \beta)B(z, \beta)$ , where

$$B(x, \beta) = \begin{cases} e^{\beta\sqrt{1-(x/w)^2}}/e^{\beta} & -w \leq x \leq w \\ 0 & \text{otherwise.} \end{cases}$$

This definition effectively yields a truncation, and again with  $2w = Ph$ , there are  $P^3$  points in the support. The Fourier transform of this window function is not analytically known. By the structure of the function, it is sufficient to compute a 1D FFT (or at most three 1D FFTs if all dimensions are different), to obtain the transform numerically. This can then be used in the scaling step ([Equation \(24\)](#)).

Although not proven yet, from numerical evidence [Saffar Shamshirgar and Tornberg \[2017a\]](#) we can predict that the approximation error comparable to [Equation \(30\)](#) is

$$(31) \quad C \left( \beta^2 e^{-2\pi P^2/\beta} + \operatorname{erfc}(\sqrt{\beta}) \right).$$

Hence, also here we can choose a parameter ( $\beta$ ) to balance the resolution and truncation. In [Saffar Shamshirgar and Tornberg \[ibid.\]](#) we find  $\beta = 2.5P$  close to optimal, and with this an approximation error that decays like  $Ce^{-2.5P}$ .

Hence, this window function shares many properties with the Gaussian, and the approximation errors decay faster with  $P$  ( $Ce^{-2.5P}$  as compared to  $Ce^{-\pi P/2}$ ). In [Saffar Shamshirgar and Tornberg \[ibid.\]](#) it is shown that the evaluation costs for the two window functions are comparable for the same  $P$ , and hence that the new BM window function is computationally more efficient.

## 4 The spectral Ewald method for different periodicities

In the previous section, we introduced the Spectral Ewald method for the triply periodic electrostatic problem. In the case of one or more non-periodic direction(s), we have to

make some modifications. Our formulas now involve integrals defining inverse Fourier transforms in [Equations \(18\), \(20\) and \(22\)](#), respectively. In addition, the evaluation of  $H(\mathbf{x})$  in [Equation \(27\)](#) will be slightly different, and we start at this end.

In the triply periodic case, we introduced a uniform grid of grid size  $M^3$  on  $[0, L]^3$  with  $h = L/M$ . Now, we need to extend the grid in the non-periodic direction to accomodate the support of the window functions. We set  $\tilde{L} = L + Ph$ , and  $\tilde{M} = M + P$  s.t.  $h = L/M = \tilde{L}/\tilde{M}$ . Approximating the Fourier integrals with the trapezoidal rule, we obtain discrete sums that can be evaluated with FFTs. The errors that we introduce are similar to the errors introduced by discretization of the integral in [Equation \(28\)](#), as was discussed in [Section 3.2](#).

In analogy with the triply periodic case, we define a  $\hat{\tilde{H}}(\mathbf{k})$  as  $\hat{\tilde{H}}_{\mathbf{k}}$  in [Equation \(24\)](#) with the factor  $1/k^2$  replaced by  $\hat{G}^{DP}(\mathbf{k})$  (i.e.  $\hat{G}^{3P}(\mathbf{k}) = 1/|\mathbf{k}|^2 = 1/k^2$ ). Consider e.g. the singly periodic case, then we have

$$\tilde{H}(\mathbf{x}) = \frac{1}{(2\pi)^2} \sum_{k_3} \int_{\mathbb{R}^2} \hat{\tilde{H}}(\mathbf{k}) e^{i\mathbf{k}\cdot\mathbf{x}} d\kappa_1 d\kappa_2$$

where  $\mathbf{k} = (\kappa_1, \kappa_2, k_3)$ . For  $k_3 \neq 0$  we have  $\hat{G}^{1P}(\mathbf{k}) = 1/|\mathbf{k}|^2 = 1/(\kappa_1^2 + \kappa_2^2 + k_3^2)$  ([Equation \(21\)](#)). To evaluate this integral accurately for all discrete  $k_3$  we need to discretize it on a finer grid in  $k$ -space than a regular FFT of  $H(x)$  above would yield. Hence, for the  $1P/2P$  cases, there are  $2/1$  non-periodic directions that need this refinement.

The simplest way to achieve this is to define a global upsampling factor  $s_g$  and extend the domain to  $s_g \tilde{L}$  in any non-periodic direction. This is a so-called zero padding in real space, which leads to a denser sampling of modes in Fourier space. Applying an FFT on a grid of size  $s_g \tilde{M}$  in a non-periodic direction, yields a sampling in  $\kappa$  for  $\kappa = 2\pi n/(s_g \tilde{L})$ ,  $n = -s_g \tilde{M}/2, \dots, s_g \tilde{M}/2 - 1$ . An upsampling such that  $s_g \tilde{M}$  is 4 up to 6 times larger than  $M$  can be needed depending on the accuracy requirement. This yields a large extra cost especially in the 1P case, with FFTs that are 16 or even 36 times larger than in the triply periodic case.

In [Saffar Shamshirgar and Tornberg \[2017b\]](#), we show that it is sufficient to apply up-sampling to a band of discrete  $k_3$  modes with small magnitude, and introduce an adaptive FFT and IFFT. With the AFFT, a typical increase in computational cost relative to the FFT without oversampling is a factor of  $2 - 3$ , with global upsampling it is  $16 - 36$ . For the doubly periodic case, upsampling is only needed in one dimension. Hence, the cost of global upsampling is not so overwhelming and was used in the first 2P implementation in [Lindbo and Tornberg \[2012\]](#).

In this implementation [Lindbo and Tornberg \[ibid.\]](#), the treatment of the zero mode  $k_1 = k_2 = 0$ , was done separately. In the final Ewald formula, there is an explicit 1D

sum for this term (see e.g. section 9 of [Tornberg \[2016\]](#)), and an expansion with Chebyshev polynomials was used for efficient evaluation. Again, moving to the singly periodic (1P) case, the corresponding term is a 2D sum, and cannot be evaluated as efficiently. This is where the idea of solving free space problems with FFTs as introduced by [Vico, Green-gard, and Ferrando \[2016\]](#) enters. Now we consider the 2D free space Poisson problem corresponding to  $k_3 = 0$  instead of the solution written as a sum in the Ewald formula. By introducing the truncated Green's function for the free space problem and then it's Fourier transform, we arrived at the definition in [Equation \(21\)](#) of  $\hat{G}^{1P}(\mathbf{k})$  for  $k_3 = 0$ . It has a finite limit as  $|\mathbf{k}| \rightarrow 0$ , also given in the definition. Hence, this can be treated as all the other discrete  $k_3$  modes. An oversampling factor of  $s_0 = 1 + \sqrt{2}$  is sufficient for this mode. With this approach, the  $k_3 = 0$  mode can be included almost for free.

This approach to solve free space problems is used for the free space case with no periodicity. With  $\hat{G}^{0P}(\mathbf{k})$  defined in [Equation \(23\)](#), an upsampling factor of  $1 + \sqrt{3}$  in each dimension is sufficient for full accuracy. A precomputation can however be made, to reduce the needed upsampling to a factor of 2, which is the minimum upsampling for computing an aperiodic convolution. This was first introduced for potentials of Stokes flow in [af Klinteberg, Saffar Shamshirgar, and Tornberg \[2017\]](#).

An extension of the domain length from  $L$  to  $\tilde{L}$  to fit the support of the window function does not guarantee that the Gaussians in the right hand side of [Equation \(11\)](#) will be sufficiently decayed in this domain. As was discussed in [af Klinteberg, Saffar Shamshirgar, and Tornberg \[ibid.\]](#), an additional extension is however needed only if  $\tilde{M}$  is picked larger than necessary for a given error tolerance (for a fixed  $P$  this reduces the support width  $Ph$  of the window function and hence  $\tilde{L}$ ).

Very recently, we have treated all cases of periodicity in a unified framework [Saffar Shamshirgar and Tornberg \[2017a\]](#), also introducing the new window function as suggested by [Barnett and Magland \[2017\]](#). Hence, this includes an implementation of the 2P method with both adaptive FFT and free space FFT treatment of the  $k_3 = 0$  term, features that differ from the original 2P method [Lindbo and Tornberg \[2012\]](#).

## 5 Extension to fundamental solutions for Stokes flow

One fundamental solution for Stokes flow, the stokeslet, was introduced in [Equation \(2\)](#). The triply periodic SE method for the stokeslet is very similar in structure to that for electrostatics. We however work with vector point sources (forces) and the Fourier representation of the stokeslet is a matrix for each  $\mathbf{k}$ . Another important fundamental solution, needed e.g. when formulating second kind integral equations for Stokes flow, is the stresslet, as given by

$$(32) \quad T_{j\ell m}(\mathbf{r}) = -6 \frac{r_j r_\ell r_m}{r^5} \quad j, \ell, m = 1, 2, 3.$$



The stresslet has three indicies and the triply periodic SE method will involve 6 FFTs for the source components, and three inverse FFTs for the components of the solution. Unlike the stokeslet, the stresslet does not by construction generate a divergence free velocity field. The correction term needed to impose a zero mean flow through a periodic cell was derived in [af Klinteberg and Tornberg \[2014\]](#). Available truncation error estimates for the Ewald sums from [Lindbo and Tornberg \[2010\]](#) and [af Klinteberg and Tornberg \[2014\]](#) are summarized in [af Klinteberg, Saffar Shamshirgar, and Tornberg \[2017\]](#), as additional estimates needed for the free space case are derived.

For the free space case in [af Klinteberg, Saffar Shamshirgar, and Tornberg \[ibid.\]](#), we write the stokeslet and the stresslet as a differential operator acting on  $r$ , which is the Green's function for the biharmonic equation. For the stokeslet, this becomes

$$S_{j\ell} = (\delta_{j\ell} \nabla^2 - \nabla_j \nabla_\ell) r \quad j, l = 1, 2, 3.$$

Using the approach from [Vico, Greengard, and Ferrando \[2016\]](#) for the biharmonic kernel, which is a radial kernel, we can use this structure to extend the treatment to the stokeslet and the stresslet. Similarly, the rotlet is based on the harmonic kernel.

The doubly periodic case for the stokeslet was treated in [Lindbo and Tornberg \[2011a\]](#), in line with the 2P treatment of electrostatics in [Lindbo and Tornberg \[2012\]](#). The extension to the 2P and 1P cases for both the stokeslet and stresslet in line with the new unified treatment of electrostatics as discussed in the previous section has not yet been done.

## 6 Summary and future work

We have in this paper reviewed the development of the Spectral Ewald methods. We have mainly considered their application to electrostatic problems, but also discussed the extension to Stokes flow. With the recent developments in [Saffar Shamshirgar and Tornberg \[2017b\]](#) and [Saffar Shamshirgar and Tornberg \[2017a\]](#), we now have a method for electrostatics that offers a unified treatment for problems with different periodicities, from triply periodic down to free space.

Compared to the triply periodic case, which is the most efficient, each non-periodic dimension increases the computational cost, but with the adaptive FFTs only to a limited amount. The cost of the algorithm associated with the window functions is essentially independent of the FFT grid size. Since that cost is reduced with the new Barnett-Magland window function, the increase in the FFT cost will have a larger impact when measuring computational cost relative to the triply periodic case. In this setting, and for typical parameter choices, we noted in [Saffar Shamshirgar and Tornberg \[2017b\]](#) that the doubly periodic case is only marginally more expensive than the triply periodic, and the singly periodic and free space cases are up to two and four times as expensive, respectively.

These FFT based methods are alternatives to fast multipole methods for evaluating electrostatic and Stokes potentials, and we have shown that the SE method is competitive with FMM [Greengard \[2012\]](#) for the free space summation of Stokes potentials, where it is at its largest disadvantage [of Klinteberg, Saffar Shamshirgar, and Tornberg \[2017\]](#). We expect the SE method to do better the more periodic directions we have, and the FMM to do better relative to the SE method the more non-uniform the distribution of points get. Hence, this is not to conclude that one method is always better than the other, but only to remark that an FFT based SE method can be a competitive alternative to the FMM method. There is an additional value in having a method that can be used for different periodicities, thereby keeping the structure intact and easing the integration with the rest of the simulation code, concerning, e.g., modifications of quadrature methods in a boundary integral method to handle near interactions.

Future work involves extending the unified treatment for the harmonic kernel of electrostatics also to stokeslets and stresslets. In order for this to be possible it remains first to derive the appropriate Ewald summation formulas for the singly periodic stokeslet and for the singly and doubly periodic stresslet sums, as we are not aware of any suitable decompositions.

## References

- A. Barnett and J. F. Magland (2017). “[FINUFFT: a fast and lightweight non-uniform fast Fourier transform library](#)” (cit. on pp. [3713](#), [3714](#), [3723](#), [3725](#)).
- C. W. J. Beenakker (1986). “[Ewald sum of the Rotne–Prager tensor](#)”. *J. Chem. Phys.* 85.3, p. 1581 (cit. on p. [3712](#)).
- H. Cheng, L. Greengard, and V. Rokhlin (Nov. 1999). “[A Fast Adaptive Multipole Algorithm in Three Dimensions](#)”. *J. Comput. Phys.* 155.2, pp. 468–498 (cit. on p. [3712](#)).
- T. Darden, D. York, and L. Pedersen (1993). “Particle mesh Ewald: An  $O(N) \log(N)$  method for Ewald sums in large systems”. *J. Chem. Phys.* 98, p. 10089 (cit. on p. [3712](#)).
- J. De Joannis, A. Arnold, and C. Holm (2002). “[Electrostatics in periodic slab geometries. II](#)”. *J. Chem. Phys.* 117.6, p. 2503. arXiv: [cond-mat/0202400](#) (cit. on p. [3713](#)).
- M. Deserno and C. Holm (1998). “How to mesh up Ewald sums. I. A theoretical and numerical comparison of various particle mesh routines”. *J. Chem. Phys.* 109, 7678–7693 (cit. on pp. [3711](#), [3712](#), [3722](#)).
- U. Essmann, L. Perera, M. Berkowitz, T. Darden, H. Lee, and L. Pedersen (1995). “A smooth particle mesh Ewald method”. *J. Chem. Phys.* 103, p. 8577 (cit. on pp. [3712](#), [3722](#)).
- P. Ewald (1921). “Die Berechnung optischer und elektrostatischer Gitterpotentiale”. *Ann. Phys.* 64, 253–287 (cit. on p. [3711](#)).

- J.G. Fripiat, J. Delhalle, I. Flamant, and F. E. Harris (2010). “Ewald-type formulas for Gaussian-basis Bloch states in one-dimensionally periodic systems”. *J. Chem. Phys.* 132, p. 044108 (cit. on p. 3712).
- L. Greengard (2012). “Fast multipole methods for the Laplace, Helmholtz and Stokes equations in three dimensions” (cit. on pp. 3715, 3727).
- L Greengard and V Rokhlin (1987). “A Fast Algorithm for Particle Simulations”. *J. Comput. Phys.* 73, pp. 325–348 (cit. on p. 3712).
- A. Grzybowski, E. Gwóźdź, and A. Bródka (2000). “Ewald summation of electrostatic interactions in molecular dynamics of a three-dimensional system with periodicity in two directions”. *Phys. Rev. B* 61, 6706–6712 (cit. on p. 3712).
- N.A. Gumerov and R. Duraiswami (2014). “A method to compute periodic sums”. *J. Comput. Phys.* 272, pp. 307–326 (cit. on p. 3712).
- H. Hasimoto (Feb. 1959). “On the periodic fundamental solutions of the Stokes equations and their application to viscous flow past a cubic array of spheres”. English. *J. Fluid Mech.* 5.02, p. 317 (cit. on p. 3712).
- L. af Klinteberg, D. Saffar Shamshirgar, and A.-K. Tornberg (2017). “Fast Ewald summation for free-space Stokes potentials”. *Res. Math. Sci.* 4, p. 1 (cit. on pp. 3714, 3715, 3720, 3725–3727).
- L. af Klinteberg and A.-K. Tornberg (2014). “Fast Ewald summation for Stokesian particle suspensions”. *Int. J. Numer. Methods Fluids* 76.10, pp. 669–698 (cit. on p. 3726).
- J. Kolafa and J. W. Perram (1992). “Cutoff Errors in the Ewald Summation Formulae for Point Charge Systems”. *Mol. Simul.* 9.5, pp. 351–368 (cit. on p. 3720).
- S. W. de Leeuw, J. W. Perram, and E. R. Smith (1980). “Simulation of electrostatic systems in periodic boundary conditions. I. Lattice sums and dielectric constants”. English. *Proc. Royal Soc. London A* 373, 27–56 (cit. on p. 3711).
- D. Lindbo, L. af Klinteberg, and D. Saffar Shamshirgar (2016). “The Spectral Ewald Unified package” (cit. on p. 3714).
- D. Lindbo and A.-K. Tornberg (2010). “Spectrally accurate fast summation for periodic Stokes potentials”. *J. Comput. Phys.* 229.23, pp. 8994–9010 (cit. on pp. 3713, 3726).
- (2011a). “Fast and spectrally accurate summation of 2-periodic Stokes potentials”. arXiv: 1111.1815 (cit. on pp. 3712, 3726).
- (2011b). “Spectral accuracy in fast Ewald-based methods for particle simulations”. *J. Comput. Phys.* 230.24, pp. 8744–8761 (cit. on pp. 3713, 3720, 3722).
- (2012). “Fast and spectrally accurate Ewald summation for 2-periodic electrostatic systems”. *J. Chem. Phys.* 136, p. 164111 (cit. on pp. 3713, 3714, 3720, 3724–3726).
- F. Nestler, M. Pippig, and D. Potts (2015). “Fast Ewald summation based on NFFT with mixed periodicity”. *J. Comput. Phys.* 285, pp. 280–315 (cit. on p. 3714).

- M. Porto (2000). “Ewald summation of electrostatic interactions of systems with finite extent in two of three dimensions”. *J. Phys. A: Math. Gen.* 33, pp. 6211–6218 (cit. on p. [3712](#)).
- C. Pozrikidis (1996). “Computation of periodic Green’s functions of Stokes flow”. *J. Eng. Math.* 30, pp. 79–96 (cit. on p. [3712](#)).
- D. Saffar Shamshirgar and A.-K. Tornberg (2017a). “Fast Ewald summation for electrostatic potentials with arbitrary periodicity” (cit. on pp. [3713](#), [3714](#), [3723](#), [3725](#), [3726](#)).
- (2017b). “The Spectral Ewald method for singly periodic domains”. *J. Comput. Phys.* 347, pp. 341–366 (cit. on pp. [3714](#), [3717](#), [3720](#), [3724](#), [3726](#)).
- D. Saintillan, E. Darve, and E. Shaqfeh (2005). “A smooth particle-mesh Ewald algorithm for Stokes suspension simulations: The sedimentation of fibers”. *Phys. Fluids* 17.3 (cit. on p. [3713](#)).
- E.R. Smith (2008). “Electrostatic potentials in systems periodic in one, two, and three dimensions”. *J. Chem. Phys.* 128, p. 174104 (cit. on p. [3711](#)).
- A.-K. Tornberg (2016). “The Ewald sums for singly, doubly and triply periodic electrostatic systems”. *Adv. Comput. Math.* 42.1, pp. 227–248. arXiv: [1404 . 3534](#) (cit. on pp. [3712](#), [3716](#), [3717](#), [3725](#)).
- A.-K. Tornberg and L. Greengard (2008). “A fast multipole method for the three-dimensional Stokes equations”. *J. Comput. Phys.* 227.3, pp. 1613–1619 (cit. on p. [3712](#)).
- F. Vico, L. Greengard, and M. Ferrando (2016). “Fast convolution with free-space Green’s functions”. *J. Comput. Phys.* 323, pp. 191–203 (cit. on pp. [3714](#), [3715](#), [3718](#), [3725](#), [3726](#)).
- H. Wang, T. Lei, J. Li, J. Huang, and Z. Yao (2007). “A parallel fast multipole accelerated integral equation scheme for 3D Stokes equations”. *Int. J. Numer. Methods Eng.* 70.7, pp. 812–839 (cit. on p. [3712](#)).

Received 2018-02-08.

ANNA-KARIN TORNBORG  
DEPARTMENT OF MATHEMATICS  
ROYAL INSTITUTE OF TECHNOLOGY (KTH)  
SWEDEN  
[akto@kth.se](mailto:akto@kth.se)



# WORST-CASE EVALUATION COMPLEXITY AND OPTIMALITY OF SECOND-ORDER METHODS FOR NONCONVEX SMOOTH OPTIMIZATION

CORALIA CARTIS, NICHOLAS I. M. GOULD AND PHILIPPE L. TOINT

## Abstract

We establish or refute the optimality of inexact second-order methods for unconstrained nonconvex optimization from the point of view of worst-case evaluation complexity, improving and generalizing our previous results. To this aim, we consider a new general class of inexact second-order algorithms for unconstrained optimization that includes regularization and trust-region variations of Newton’s method as well as of their linesearch variants. For each method in this class and arbitrary accuracy threshold  $\epsilon \in (0, 1)$ , we exhibit a smooth objective function with bounded range, whose gradient is globally Lipschitz continuous and whose Hessian is  $\alpha$ –Hölder continuous (for given  $\alpha \in [0, 1]$ ), for which the method in question takes at least  $\lfloor \epsilon^{-(2+\alpha)/(1+\alpha)} \rfloor$  function evaluations to generate a first iterate whose gradient is smaller than  $\epsilon$  in norm. Moreover, we also construct another function on which Newton’s takes  $\lfloor \epsilon^{-2} \rfloor$  evaluations, but whose Hessian is Lipschitz continuous on the path of iterates. These examples provide lower bounds on the worst-case evaluation complexity of methods in our class when applied to smooth problems satisfying the relevant assumptions. Furthermore, for  $\alpha = 1$ , this lower bound is of the same order in  $\epsilon$  as the upper bound on the worst-case evaluation complexity of the cubic regularization method and other algorithms in a class of methods recently proposed by Curtis, Robinson and Samadi or by Royer and Wright, thus implying that these methods have optimal worst-case evaluation complexity within a wider class of second-order methods, and that Newton’s method is suboptimal.

## 1 Introduction

Newton’s method has long represented a benchmark for rapid asymptotic convergence when minimizing smooth, unconstrained objective functions [Dennis and Schnabel \[1983\]](#).

It has also been efficiently safeguarded to ensure its global convergence to first- and even second-order critical points, in the presence of local nonconvexity of the objective using linesearch [Nocedal and Wright \[1999\]](#), trust-region [Conn, Gould, and Toint \[2000\]](#) or other regularization techniques [Griewank \[1981\]](#), [Nesterov and Polyak \[2006\]](#), and [Cartis, Gould, and Toint \[2011a\]](#). Many variants of these globalization techniques have been proposed. These generally retain fast local convergence under non-degeneracy assumptions, are often suitable when solving large-scale problems and sometimes allow approximate rather than true Hessians to be employed. We attempt to capture the common features of these methods in the description of a general class of second-order methods, which we denote by  $\mathfrak{M}.\alpha$  in what follows.

In this paper, we are concerned with establishing *lower bounds* on the worst-case evaluation complexity of the  $\mathfrak{M}.\alpha$  methods<sup>(1)</sup> when applied to “sufficiently smooth” nonconvex minimization problems, in the sense that we exhibit objective functions on which these methods take a large number of function evaluations to obtain an approximate first-order point.

There is a growing literature on the global worst-case evaluation complexity of first- and second-order methods for nonconvex smooth optimization problems (for which we provide a partial bibliography with this paper). In particular, it is known [Vavasis \[1993\]](#), [Nesterov \[2004, p. 29\]](#) that steepest-descent method with either exact or inexact line-searches takes at most<sup>(2)</sup>  $\mathcal{O}(\epsilon^{-2})$  iterations/function-evaluations to generate a gradient whose norm is at most  $\epsilon$  when started from an arbitrary initial point and applied to nonconvex smooth objectives with gradients that are globally Lipschitz continuous within some open convex set containing the iterates generated. Furthermore, this bound is essentially sharp (for inexact [Cartis, Gould, and Toint \[2010\]](#) and exact [Cartis, Gould, and Toint \[2012c\]](#) line-searches). Similarly, trust-region methods that ensure at least a Cauchy (steepest-descent-like) decrease on each iteration satisfy a worst-case evaluation complexity bound of the same order under identical conditions [Gratton, Sartenauer, and Toint \[2008\]](#). It follows that Newton’s method globalized by trust-region regularization has the same  $\mathcal{O}(\epsilon^{-2})$  worst-case evaluation upper bound; such a bound has also been shown to be essentially sharp [Cartis, Gould, and Toint \[2010\]](#).

From a worst-case complexity point of view, one can do better when a cubic regularization/perturbation of the Newton direction is used [Griewank \[1981\]](#), [Nesterov and Polyak \[2006\]](#), [Cartis, Gould, and Toint \[2011a\]](#), and [Curtis, Robinson, and Samadi \[2017b\]](#)—such a method iteratively calculates step corrections by (exactly or approximately) minimizing a cubic model formed of a quadratic approximation of the objective and the cube

<sup>(1)</sup>And, as an aside, on that of the steepest-descent method.

<sup>(2)</sup>When  $\{a_k\}$  and  $\{b_k\}$  are two sequences of real numbers, we say that  $a_k = \mathcal{O}(b_k)$  if the ratio  $a_k/b_k$  is bounded.

of a weighted norm of the step. For such a method, the worst-case global complexity improves to be  $\mathcal{O}(\epsilon^{-3/2})$  [Nesterov and Polyak \[2006\]](#) and [Cartis, Gould, and Toint \[2011a\]](#), for problems whose gradients and Hessians are Lipschitz continuous as above; this bound is also essentially sharp [Cartis, Gould, and Toint \[2010\]](#). If instead powers between two and three are used in the regularization, then an “intermediate” worst-case complexity of  $\mathcal{O}(\epsilon^{-(2+\alpha)/(1+\alpha)})$  is obtained for such variants when applied to functions with globally  $\alpha$ -Hölder continuous Hessian on the path of iterates, where  $\alpha \in (0, 1]$  [Cartis, Gould, and Toint \[2011d\]](#). It is finally possible, as proposed in [Royer and Wright \[2017\]](#), to obtain the desired  $\mathcal{O}(\epsilon^{-3/2})$  order of worst-case evaluation complexity using a purely quadratic regularization, at the price of mixing iterations using the regularized and unregularized Hessian with iterations requiring the computation of its left-most eigenpair.

These (essentially tight) upper bounds on the worst-case evaluation complexity of such second-order methods naturally raise the question as to whether other second-order methods might have better worst-case complexity than cubic (or similar) regularization over certain classes of sufficiently smooth functions. To attempt to answer this question, we define a general, parametrized class of methods that includes Newton’s method, and that attempts to capture the essential features of globalized Newton variants we have mentioned. Our class includes for example, the algorithms discussed above as well as multiplier-adjusting types such as the Goldfeld-Quandt-Trotter approach [Goldfeld, Quandt, and Trotter \[1966\]](#). The methods of interest take a potentially-perturbed Newton step at each iteration so long as the perturbation is “not too large” and the subproblem is solved “sufficiently accurately”. The size of the perturbation allowed is simultaneously related to the parameter  $\alpha$  defining the class of methods and the rate of the asymptotic convergence of the method. For each method in each  $\alpha$ -parametrized class and each  $\epsilon \in (0, 1)$ , we construct a function with globally  $\alpha$ -Hölder-continuous Hessian and Lipschitz continuous gradient for which the method takes precisely  $\lceil \epsilon^{-(2+\alpha)/(1+\alpha)} \rceil$  function evaluations to drive the gradient norm below  $\epsilon$ . As such counts are the same order as the worst-case upper complexity bound of regularization methods, it follows that the latter methods are optimal within their respective  $\alpha$ -class of methods. As  $\alpha$  approaches zero, the worst-case complexity of these methods approaches that of steepest descent, while for  $\alpha = 1$ , we recover that of cubic regularization. We also improve the examples proposed in [Cartis, Gould, and Toint \[2010, 2011d\]](#) in two ways. The first is that we now employ objective functions with bounded range, which allows refining the associated definition of sharp worst-case evaluation complexity bounds, the second being that the new examples now have finite isolated global minimizers.

The structure of the paper is as follows. Section 2 describes the parameter-dependent class of methods and objectives of interest; Section 2.1 gives properties of the methods such as their connection to fast asymptotic rates of convergence while Section 2.2 reviews some well-known examples of methods covered by our general definition of the class.



Section 3 then introduces two examples of inefficiency of these methods and Section 4 discusses the consequences of these examples regarding the sharpness and possible optimality of the associated worst-case evaluation complexity bounds. Further consequences of our results on the new class proposed by [Curtis, Robinson, and Samadi \[2017b\]](#) and [Royer and Wright \[2017\]](#) are developed in Section 5 and 6, respectively. Section 7 draws our conclusions.

**Notation.** Throughout the paper,  $\|\cdot\|$  denotes the Euclidean norm on  $\mathbb{R}^n$ ,  $I$  the  $n \times n$  identity matrix, and  $\lambda_{\min}(H)$  and  $\lambda_{\max}(H)$  the left- and right-most eigenvalue of any given symmetric matrix  $H$ , respectively. The condition number of a symmetric positive definite matrix  $M$  is denoted by  $\kappa(M) \stackrel{\text{def}}{=} \lambda_{\max}(M)/\lambda_{\min}(M)$ . If  $M$  is only positive-semidefinite which we denote by  $M \succeq 0$ , and  $\lambda_{\min}(M) = 0$ , then  $\kappa(0) \stackrel{\text{def}}{=} +\infty$  unless  $M = 0$ , in which case we set  $\kappa(0) \stackrel{\text{def}}{=} 1$ . Positive definiteness of  $M$  is written as  $M \succ 0$ .

## 2 A general parametrized class of methods and objectives

Our aim is to minimize a given  $C^2$  objective function  $f(x)$ ,  $x \in \mathbb{R}^n$ . We consider methods that generate sequences of iterates  $\{x_k\}$  for which  $\{f(x_k)\}$  is monotonically decreasing, we let

$$f_k \stackrel{\text{def}}{=} f(x_k), \quad g_k \stackrel{\text{def}}{=} g(x_k) \quad \text{and} \quad H_k \stackrel{\text{def}}{=} H(x_k).$$

where  $g(x) = \nabla_x f(x)$  and  $H(x) = \nabla_{xx} f(x)$ .

Let  $\alpha \in [0, 1]$  be a fixed parameter and consider iterative methods whose iterations are defined as follows. Given some  $x_0 \in \mathbb{R}^n$ , let

$$(2.1) \quad x_{k+1} = x_k + s_k, \quad k \geq 0,$$

where  $s_k$  satisfies

$$(2.2) \quad (H_k + M_k)s_k = -g_k + r_k \quad \text{with} \quad \|r_k\| \leq \min[\kappa_{rg}\|g_k\|, \kappa_{rs}\|M_k s_k\|]$$

for some residual  $r_k$  and constants  $\kappa_{rg} \in [0, 1)$  and  $\kappa_{rs} > 0$ , and for some symmetric matrix  $M_k$  such that

$$(2.3) \quad M_k \succeq 0, \quad H_k + M_k \succeq 0$$

and

$$(2.4) \quad \lambda_{\min}(H_k) + \lambda_{\min}(M_k) \leq \kappa_\lambda \max\left\{|\lambda_{\min}(H_k)|, \|g_k\|^{\frac{\alpha}{1+\alpha}}\right\}$$

for some  $\kappa_\lambda > 1$  independent of  $k$ . Without loss of generality, we assume that  $s_k \neq 0$ . Furthermore, we require that no infinite steps are taken, namely

$$(2.5) \quad \|s_k\| \leq \kappa_s$$

for some  $\kappa_s > 0$  independent of  $k$ . The  $\mathfrak{M}.\alpha$  class of second-order methods consists of all methods whose iterations satisfy (2.1)–(2.5). The particular choices  $M_k = \lambda_k I$  and  $M_k = \lambda_k N_k$  (with  $N_k$  symmetric, positive definite and with bounded condition number) will be of particular interest in what follows<sup>(3)</sup>. Note that the definition of  $\mathfrak{M}.\alpha$  just introduced generalizes that of  $\mathcal{M}.\alpha$  in [Cartis, Gould, and Toint \[2011d\]](#).

Typically, the expression (2.2) for  $s_k$  is derived by minimizing (possibly approximately) the second-order model

$$(2.6) \quad m_k(s) = f_k + g_k^T s + \frac{1}{2} s^T (H_k + \beta_k M_k) s, \quad \text{with } \beta_k \stackrel{\text{def}}{=} \beta_k(s) \geq 0 \quad \text{and} \quad \beta_k \leq 1$$

of  $f(x_k + s)$ —possibly with an explicit regularizing constraint—with the aim of obtaining a sufficient decrease of  $f$  at the new iterate  $x_{k+1} = x_k + s_k$  compared to  $f(x_k)$ . In the definition of an  $\mathfrak{M}.\alpha$  method however, the issue of (sufficient) objective-function decrease is not explicitly addressed/required. There is no loss of generality in doing so here since although local refinement of the model may be required to ensure function decrease, the number of function evaluations to do so (at least for known methods) does not increase the overall worst-case evaluation complexity by more than a constant multiple and thus does not affect quantitatively the worst-case bounds derived; see for example, [Cartis, Gould, and Toint \[2010, 2011b\]](#) and [Gratton, Sartenaer, and Toint \[2008\]](#) and also Section 2.2. Furthermore, the examples of inefficiency proposed in Section 3 are constructed in such a way that each iteration of the method automatically provides sufficient decrease of  $f$ .

Having defined the classes of methods we shall be concerned with, we now specify the problem classes that we shall apply the methods in each class to, in order to demonstrate slow convergence. Given a method in  $\mathfrak{M}.\alpha$ , we are interested in minimizing functions  $f$  that satisfy

A. $\alpha$   $f : \mathbb{R}^n \rightarrow \mathbb{R}$  is twice continuously differentiable and bounded below, with gradient  $g$  being globally Lipschitz continuous on  $\mathbb{R}^n$  with constant  $L_g$ , namely,

$$(2.7) \quad \|g(x) - g(y)\| \leq L_g \|x - y\|, \quad \text{for all } x, y \in \mathbb{R}^n;$$

and the Hessian  $H$  being globally  $\alpha$ –Hölder continuous on  $\mathbb{R}^n$  with constant  $L_{H,\alpha}$ , i.e.,

$$(2.8) \quad \|H(x) - H(y)\| \leq L_{H,\alpha} \|x - y\|^\alpha, \quad \text{for all } x, y \in \mathbb{R}^n.$$

□

---

<sup>(3)</sup>Note that (2.4) is slightly more general than a maybe more natural condition involving  $\lambda_{\min}(H_k + M_k)$  instead of  $\lambda_{\min}(H_k) + \lambda_{\min}(M_k)$ .

The case when  $\alpha = 1$  in  $\mathcal{A}.\alpha$  corresponds to the Hessian of  $f$  being globally Lipschitz continuous. Moreover, (2.7) implies (2.8) when  $\alpha = 0$ , so that the  $\mathcal{A}.0$  class is that of twice continuously differentiable functions with globally Lipschitz continuous gradient. Note also that (2.7) and the existence of  $H(x)$  imply that

$$(2.9) \quad \|H(x)\| \leq L_g$$

for all  $x \in \mathbb{R}^n$  Nesterov [2004, Lemma 1.2.2], and that every function  $f$  satisfying  $\mathcal{A}.\alpha$  with  $\alpha > 1$  must be quadratic. As we will see below, it turns out that we could weaken the conditions defining  $\mathcal{A}.\alpha$  by only requiring (2.7) and (2.8) to hold in an open set containing all the segments  $[x_k, x_k + s_k]$  (the “path of iterates”), but these segments of course depend themselves on  $f$  and the method applied.

The next subsection provides some background and justification for the technical condition (2.4) by relating it to fast rates of asymptotic convergence, which is a defining feature of second-order algorithms. In Section 2.2, we then review some methods belonging to  $\mathcal{M}.\alpha$ .

**2.1 Properties of the methods in  $\mathcal{M}.\alpha$ .** We first state inclusions properties for  $\mathcal{M}.\alpha$  and  $\mathcal{A}.\alpha$ .

**Lemma 2.1.** *1. Consider a method of  $\mathcal{M}.\alpha_1$  for  $\alpha_1 \in [0, 1]$  and assume that it generates bounded gradients. Then it belongs to  $\mathcal{M}.\alpha_2$  for  $\alpha_2 \in [0, \alpha_1]$ .*

*2.  $\mathcal{A}.\alpha_1$  implies  $\mathcal{A}.\alpha_2$  for  $\alpha_2 \in [0, \alpha_1]$ , with  $L_{H,\alpha_2} = \max[L_{H,\alpha_1}, 2L_g]$ .*

*Proof.* By assumption,  $\|g_k\| \leq \kappa_g$  for some  $\kappa_g \geq 1$ . Hence, if  $\|g_k\| \geq 1$ ,

$$(2.10) \quad \|g_k\|^{\frac{\alpha_1}{1+\alpha_1}} \leq \kappa_g^{\frac{\alpha_1}{1+\alpha_1}} \leq \kappa_g \leq \kappa_g \|g_k\|^{\frac{\alpha_2}{1+\alpha_2}}$$

for any  $\alpha_2 \in [0, \alpha_1]$ . Moreover, (2.10) also holds if  $\|g_k\| \leq 1$ , proving the first statement of the lemma. Now we obtain from (2.9), that, if  $\|x - y\| > 1$ , then

$$\|H(x) - H(y)\| \leq \|H(x)\| + \|H(y)\| \leq 2L_g \leq 2L_g \|x - y\|^\alpha$$

for any  $\alpha \in [0, 1]$ . When  $\|x - y\| \leq 1$ , we may deduce from (2.8) that, if  $\alpha_1 \geq \alpha_2$ , then (2.8) with  $\alpha = \alpha_1$  implies (2.8) with  $\alpha = \alpha_2$ . This proves the second statement.  $\square$

Observe if a method is known to be globally convergent in the sense that  $\|g_k\| \rightarrow 0$  when  $k \rightarrow \infty$ , then it obviously generates bounded gradients and thus the globally convergent methods of  $\mathcal{M}.\alpha_1$  are included in  $\mathcal{M}.\alpha_2$  ( $\alpha_2 \in [0, \alpha_1]$ ).

We next give a sufficient, more concise, condition on the algorithm-generated matrices  $M_k$  that implies the bound (2.4).

**Lemma 2.2.** Let (2.2) and (2.3) hold. Assume also that the algorithm-generated matrices  $M_k$  satisfies

$$(2.11) \quad \lambda_{\min}(M_k) \leq \bar{\kappa}_\lambda \|s_k\|^\alpha, \text{ for some } \bar{\kappa}_\lambda > 1 \text{ and } \alpha \in [0, 1] \text{ independent of } k.$$

Then (2.4) holds with  $\kappa_\lambda \stackrel{\text{def}}{=} 2\bar{\kappa}_\lambda^{\frac{1}{1+\alpha}}(1 + \kappa_{rg})$ .

*Proof.* Clearly, (2.4) holds when  $\lambda_{\min}(H_k + M_k) = 0$ . When  $\lambda_{\min}(H_k + M_k) > 0$  and hence  $H_k + M_k \succ 0$ , (2.2) implies that

$$(2.12) \quad \|s_k\| \leq \frac{\|g_k\| + \|r_k\|}{\lambda_{\min}(H_k + M_k)} \leq \frac{(1 + \kappa_{rg})\|g_k\|}{\lambda_{\min}(H_k) + \lambda_{\min}(M_k)}.$$

This and (2.11) give the inequality

$$(2.13) \quad \psi(\lambda_{\min}(M_k)) \leq 0 \quad \text{with} \quad \psi(\lambda) \stackrel{\text{def}}{=} \lambda^{\frac{1}{\alpha}}(\lambda + \lambda_{\min}(H_k)) - \bar{\kappa}_\lambda^{\frac{1}{1+\alpha}}(1 + \kappa_{rg})\|g_k\|.$$

Now note that  $\psi(0) = \psi(-\lambda_{\min}(H_k)) = -\bar{\kappa}_\lambda^{\frac{1}{1+\alpha}}(1 + \kappa_{rg})\|g_k\|$  and thus

$$(2.14) \quad \psi(\lambda_{1,k}) < 0 \quad \text{with} \quad \lambda_{1,k} = \max\{0, -\lambda_{\min}(H_k)\}.$$

Moreover, the form of  $\psi(\lambda)$  implies that  $\psi(\lambda)$  is strictly increasing for  $\lambda \geq \lambda_{1,k}$ . Define now

$$(2.15) \quad \lambda_{2,k} \stackrel{\text{def}}{=} -\lambda_{\min}(H_k) + 2 \max \left\{ |\lambda_{\min}(H_k)|, \bar{\kappa}_\lambda^{\frac{1}{1+\alpha}}(1 + \kappa_{rg})^{\frac{\alpha}{1+\alpha}} \|g_k\|^{\frac{\alpha}{1+\alpha}} \right\} > \lambda_{1,k}.$$

Suppose first that  $\lambda_{\min}(H_k) < 0$  and  $|\lambda_{\min}(H_k)| \geq \bar{\kappa}_\lambda^{\frac{1}{1+\alpha}}(1 + \kappa_{rg})^{\frac{\alpha}{1+\alpha}} \|g_k\|^{\frac{\alpha}{1+\alpha}}$ . Then one verifies that  $\lambda_{2,k} = 3|\lambda_{\min}(H_k)|$  and

$$\begin{aligned} \psi(\lambda_{2,k}) &= (3|\lambda_{\min}(H_k)|)^{\frac{1+\alpha}{\alpha}} - (3|\lambda_{\min}(H_k)|)^{\frac{1}{\alpha}} |\lambda_{\min}(H_k)| - \bar{\kappa}_\lambda^{\frac{1}{1+\alpha}}(1 + \kappa_{rg})^{\frac{\alpha}{1+\alpha}} \|g_k\| \\ &= 2 \cdot 3^{\frac{1}{\alpha}} |\lambda_{\min}(H_k)|^{\frac{1+\alpha}{\alpha}} - \bar{\kappa}_\lambda^{\frac{1}{1+\alpha}}(1 + \kappa_{rg})^{\frac{\alpha}{1+\alpha}} \|g_k\| > 0 \end{aligned}$$

Suppose now that  $\lambda_{\min}(H_k) \geq 0$  and  $|\lambda_{\min}(H_k)| \geq \bar{\kappa}_\lambda^{\frac{1}{1+\alpha}}(1 + \kappa_{rg})^{\frac{\alpha}{1+\alpha}} \|g_k\|^{\frac{\alpha}{1+\alpha}}$ . Then  $\lambda_{2,k} = \lambda_{\min}(H_k)$  and

$$\psi(\lambda_{2,k}) = (\lambda_{\min}(H_k))^{\frac{1+\alpha}{\alpha}} + (\lambda_{\min}(H_k))^{\frac{1}{\alpha}} |\lambda_{\min}(H_k)| - \bar{\kappa}_\lambda^{\frac{1}{1+\alpha}}(1 + \kappa_{rg})^{\frac{\alpha}{1+\alpha}} \|g_k\| > 0.$$

Thus we deduce that  $\psi(\lambda_{2,k}) > 0$  whenever  $|\lambda_{\min}(H_k)| \geq \bar{\kappa}_\lambda^{\frac{1}{1+\alpha}}(1 + \kappa_{rg})^{\frac{\alpha}{1+\alpha}} \|g_k\|^{\frac{\alpha}{1+\alpha}}$ . Moreover the same inequality obviously holds if

$$|\lambda_{\min}(H_k)| < \bar{\kappa}_\lambda^{\frac{1}{1+\alpha}}(1 + \kappa_{rg})^{\frac{\alpha}{1+\alpha}} \|g_k\|^{\frac{\alpha}{1+\alpha}}$$

because  $\psi(\lambda)$  is increasing with  $\lambda$ . As a consequence,  $\psi(\lambda_{2,k}) > 0$  in all cases. We now combine this inequality, (2.14) and the monotonicity of  $\psi(\lambda)$  for  $\lambda \geq \lambda_{1,k}$  to obtain that either  $\lambda_{\min}(M_k) \leq \lambda_{1,k} < \lambda_{2,k}$  or  $\lambda_{\min}(M_k) \in [\lambda_{1,k}, \lambda_{2,k})$  because of (2.13). Thus  $\lambda_{\min}(M_k) \leq \lambda_{2,k}$ , which, due to (2.15) and  $\bar{\kappa}_\lambda > 1$ , implies (2.4).  $\square$

Thus a method satisfying (2.1)–(2.5) and (2.11) belongs to  $\mathfrak{M}.\alpha$ , but not every method in  $\mathfrak{M}.\alpha$  needs to satisfy (2.11). This latter requirement implies the following property regarding the length of the step generated by methods in  $\mathfrak{M}.\alpha$  satisfying (2.11) when applied to functions satisfying A. $\alpha$ .

**Lemma 2.3.** *Assume that an objective function  $f$  satisfying A. $\alpha$  is minimized by a method satisfying (2.1), (2.2), (2.11) and such that the conditioning of  $M_k$  is bounded in that  $\kappa(M_k) \leq \kappa_\kappa$  for some  $\kappa_\kappa \geq 1$ . Then there exists  $\bar{\kappa}_{s,\alpha} > 0$  independent of  $k$  such that, for  $k \geq 0$ ,*

$$(2.16) \quad \|s_k\| \geq \bar{\kappa}_{s,\alpha} \|g_{k+1}\|^{\frac{1}{1+\alpha}}.$$

*Proof.* The triangle inequality provides

$$(2.17) \quad \|g_{k+1}\| \leq \|g_{k+1} - (g_k + H_k s_k)\| + \|g_k + H_k s_k\|.$$

From (2.1),  $g_{k+1} = g(x_k + s_k)$  and Taylor expansion provides

$$g_{k+1} = g_k + \int_0^1 H(x_k + \tau s_k) s_k d\tau$$

This and (2.8) now imply

$$\begin{aligned} \|g_{k+1} - (g_k + H_k s_k)\| &\leq \left\| \int_0^1 [H(x_k + \tau s_k) - H(x_k)] d\tau \right\| \cdot \|s_k\| \\ &\leq L_{H,\alpha} (1 + \alpha)^{-1} \|s_k\|^{1+\alpha} \end{aligned}$$

so that (2.17) and (2.2) together give that

$$\|g_{k+1}\| \leq L_{H,\alpha} (1 + \alpha)^{-1} \|s_k\|^{1+\alpha} + (1 + \kappa_{rs}) \|M_k\| \|s_k\|.$$

If  $M_k \neq 0$ , this inequality and the fact that  $\kappa(M_k)$  is bounded then imply that

$$\|g_{k+1}\| \leq L_{H,\alpha} (1 + \alpha)^{-1} \|s_k\|^{1+\alpha} + \kappa(M_k) (1 + \kappa_{rs}) \lambda_{\min}(M_k) \|s_k\|,$$

while we may ignore the last term on the right-hand side if  $M_k = 0$ . Hence, in all cases,

$$\|g_{k+1}\| \leq L_{H,\alpha} (1 + \alpha)^{-1} \|s_k\|^{1+\alpha} + \kappa_\kappa (1 + \kappa_{rs}) \lambda_{\min}(M_k) \|s_k\|,$$

where we used that  $\kappa(M_k) \leq \kappa_\kappa$  by assumption. This bound and (2.11) then imply (2.16)

with  $\bar{\kappa}_{s,\alpha} \stackrel{\text{def}}{=} [L_{H,\alpha} (1 + \alpha)^{-1} + \kappa_\kappa (1 + \kappa_{rs}) \bar{\kappa}_\lambda]^{-\frac{1}{1+\alpha}}$ .  $\square$

Property (2.16) will be central for proving (in Appendix A2) desirable properties of a class of methods belonging to  $\mathfrak{M}.\alpha$ . In addition, we now show that (2.16) is a necessary condition for fast local convergence of methods of type (2.2), under reasonable assumptions; fast local rate of convergence in a neighbourhood of well-behaved minimizers is a “trademark” of what is commonly regarded as second-order methods.

**Lemma 2.4.** *Let  $f$  satisfy assumptions A. $\alpha$ . Apply an algorithm to minimizing  $f$  that satisfies (2.1) and (2.2) and for which*

$$(2.18) \quad \|M_k\| \leq \bar{\kappa}_\lambda, \quad k \geq 0, \quad \text{for some } \bar{\kappa}_\lambda > 0 \text{ independent of } k.$$

*Assume also that convergence at linear or faster than linear rate occurs, namely,*

$$(2.19) \quad \|g_{k+1}\| \leq \kappa_c \|g_k\|^{1+\alpha}, \quad k \geq 0,$$

*for some  $\kappa_c > 0$  independent of  $k$ , with  $\kappa_c \in (0, 1)$  when  $\alpha = 0$ . Then (2.16) holds.*

*Proof.* Let

$$(2.20) \quad 0 \leq \alpha_k \stackrel{\text{def}}{=} \frac{\|s_k\|}{\|g_{k+1}\|^{\frac{1}{1+\alpha}}}, \quad k \geq 0.$$

From (2.19) and the definition of  $\alpha_k$  in (2.20), we have that, for  $k \geq 0$ ,

$$\begin{aligned} (1 - \kappa_{rg}) \frac{\|s_k\|}{\alpha_k} &\leq \kappa_{c,\alpha} (1 - \kappa_{rg}) \|g_k\| \leq \kappa_{c,\alpha} \|g_k + r_k\| \\ &= \kappa_{c,\alpha} \|(H_k + M_k)s_k\| \leq \kappa_{c,\alpha} \|H_k + M_k\| \cdot \|s_k\|, \end{aligned}$$

where  $\kappa_{c,\alpha} \stackrel{\text{def}}{=} \kappa_c^{\frac{1}{1+\alpha}}$  and where we used (2.2) to obtain the first equality. It follows that

$$(2.21) \quad \|H_k + M_k\| \geq \frac{(1 - \kappa_{rg})}{\alpha_k \kappa_{c,\alpha}}, \quad k \geq 0.$$

The bounds (2.9) and (2.18) imply that  $\{H_k + M_k\}$  is uniformly bounded above for all  $k$ , namely,

$$(2.22) \quad \|H_k + M_k\| \leq \kappa_{hl}, \quad k \geq 0,$$

where  $\kappa_{hl} \stackrel{\text{def}}{=} L_g + \bar{\kappa}_\lambda$ . Now (2.21) and (2.22) give that  $\alpha_k \geq 1/(\kappa_{hl}\kappa_{c,\alpha}) > 0$ , for all  $k \geq 0$ , and so it follows from (2.20), that (2.16) holds with  $\bar{\kappa}_{s,\alpha} \stackrel{\text{def}}{=} (1 - \kappa_{rg})/(\kappa_{c1}\kappa_{c,\alpha})$ .  $\square$

It is clear from the proof of Lemma 2.4 that (2.19) is only needed asymptotically, that is for all  $k$  sufficiently large; for simplicity, we have assumed it holds globally.

Note that letting  $\alpha = 1$  in [Lemma 2.4](#) provides a necessary condition for quadratically convergent methods satisfying (2.1), (2.2) and (2.18). Also, similarly to the above proof, one can show that if superlinear convergence of  $\{g_k\}$  to zero occurs, then (2.16) holds with  $\alpha = 0$  for all  $\bar{\kappa}_{s,\alpha} > 0$ , or equivalently,  $\|g_{k+1}\|/\|s_k\| \rightarrow 0$ , as  $k \rightarrow \infty$ .

Summarizing, we have shown that (2.16) holds for a method in  $\mathfrak{M}.\alpha$  if (2.11) holds and  $\kappa(M_k)$  is bounded, or if linear or faster asymptotic convergence takes place for unit steps.

**2.2 Some examples of methods that belong to the class  $\mathfrak{M}.\alpha$ .** Let us now illustrate some of the methods that either by construction or under certain conditions belong to  $\mathfrak{M}.\alpha$ . This list of methods does not attempt to be exhaustive and other practical methods may be found to belong to  $\mathfrak{M}.\alpha$ .

**Newton's method** [Dennis and Schnabel \[1983\]](#). Newton's method for convex optimization is characterised by finding a correction  $s_k$  that satisfies  $H_k s_k = -g_k$  for nonzero  $g_k \in \text{Range}(H_k)$ . Letting

$$(2.23) \quad M_k = 0, \quad r_k = 0 \quad \text{and} \quad \beta_k = 0$$

in (2.2) and (2.6), respectively, yields Newton's method. Provided additionally that both  $g_k \in \text{Range}(H_k)$  and  $H_k$  is positive semi-definite,  $s_k$  is a descent direction and (2.3) holds. Since (2.4) is trivially satisfied in this case, it follows that Newton's method belongs to the class  $\mathfrak{M}.\alpha$ , for any  $\alpha \in [0, 1]$ , provided it does not generate infinite steps to violate (2.5). As Newton's method is commonly embedded within trust-region or regularization frameworks when applied to nonconvex functions, (2.5) will in fact, hold as it is generally enforced for the latter methods. Note that allowing  $\|r_k\| > 0$  subject to the second part of (2.2) then covers inexact variants of Newton's method.

**Regularization algorithms** [Griewank \[1981\]](#), [Nesterov \[2004\]](#), and [Cartis, Gould, and Toint \[2011b\]](#). In these methods, the step  $s_k$  from the current iterate  $x_k$  is computed by (possibly approximately) globally minimizing the model

$$(2.24) \quad m_k(s) = f_k + g_k^T s + \frac{1}{2} s^T H_k s + \frac{\sigma_k}{2 + \alpha} \|s\|^{2+\alpha},$$

where the regularization weight  $\sigma_k$  is adjusted to ensure sufficient decrease of  $f$  at  $x_k + s_k$ . We assume here that the minimization of (2.24) is carried accurately enough to ensure that  $\nabla_{ss}^2 m_k - (s) = H_k + \sigma_k \|s\| I$  is positive semidefinite, which is always possible because of [Cartis, Gould, and Toint \[2011a, Theorem 3.1\]](#). The scalar  $\alpha$  is the same fixed parameter as in the definition of  $\mathcal{A}.\alpha$  and  $\mathfrak{M}.\alpha$ , so that for each  $\alpha \in [0, 1]$ , we have a different regularization term and hence what we shall call an  $(2 + \alpha)$ -regularization method. For

$\alpha = 1$ , we recover the cubic regularization (ARC) approach [Griewank \[1981\]](#), [Weiser, Deuffhard, and Erdmann \[2007\]](#), [Nesterov and Polyak \[2006\]](#), and [Cartis, Gould, and Toint \[2011a,b\]](#). For  $\alpha = 0$ , we obtain a quadratic regularization scheme, reminiscent of the Levenberg–Morrison–Marquardt method [Nocedal and Wright \[1999\]](#). For these  $(2+\alpha)$ -regularization methods, we have

$$(2.25) \quad \alpha \in [0, 1], \quad M_k = \sigma_k \|s_k\|^\alpha I, \quad \text{and} \quad \beta_k = \frac{2}{2+\alpha}$$

in (2.2) and (2.6). If scaling the regularization term is considered, then the second of these relation is replaced by  $M_k = \sigma_k \|s_k\|^\alpha N_k$  for some fixed scaling symmetric positive definite matrix having a bounded condition number. Note that, by construction,  $\kappa(M_k) = 1$ . Since  $\alpha \geq 0$ , we have  $0 \leq \beta_k \leq 1$  which is required in (2.6). A mechanism of successful and unsuccessful iterations and  $\sigma_k$  adjustments can be devised similarly to ARC [Cartis, Gould, and Toint \[2011a, Alg. 2.1\]](#) in order to deal with steps  $s_k$  that do not give sufficient decrease in the objective. An upper bound on the number of unsuccessful iterations which is constant multiple of successful ones can be given under mild assumptions on  $f$  [Cartis, Gould, and Toint \[2011b, Theorem 2.1\]](#). Note that each (successful or unsuccessful) iteration requires one function- and at most one gradient evaluation.

We now show that for each  $\alpha \in [0, 1]$ , the  $(2+\alpha)$ -regularization method based on the model (2.24) satisfies (2.5) and (2.4) when applied to  $f$  in  $A.\alpha$ , and so it belongs to  $\mathfrak{M}.\alpha$ .

**Lemma 2.5.** *Let  $f$  satisfy  $A.\alpha$  with  $\alpha \in (0, 1]$ . Consider minimizing  $f$  by applying an  $(2+\alpha)$ -regularization method based on the model (2.24), where the step  $s_k$  is chosen as the global minimizer of the local  $\alpha$ -model, namely of  $m_k(s)$  in (2.6) with the choice (2.25), and where the regularization parameter  $\sigma_k$  is chosen to ensure that*

$$(2.26) \quad \sigma_k \geq \sigma_{\min}, \quad k \geq 0,$$

for some  $\sigma_{\min} > 0$  independent of  $k$ . Then (2.5) and (2.11) hold, and so the  $(2+\alpha)$ -regularization method belongs to  $\mathfrak{M}.\alpha$ .

(see Appendix A2 for details). The same argument that is used in [Cartis, Gould, and Toint \[2011a, Lem.2.2\]](#) for the  $\alpha = 1$  case (see also Appendix A2) provides

$$(2.27) \quad \|s_k\| \leq \max \left\{ \left( \frac{3(2+\alpha)L_g}{4\sigma_k} \right)^{\frac{1}{\alpha}}, \left( \frac{3(2+\alpha)\|g_k\|}{\sigma_k} \right)^{\frac{1}{1+\alpha}} \right\}, \quad k \geq 0,$$

so long as  $A.\alpha$  holds, which together with (2.26), implies

$$(2.28) \quad \|s_k\| \leq \max \left\{ \left( \frac{3(2+\alpha)L_g}{4\sigma_{\min}} \right)^{\frac{1}{\alpha}}, \left( \frac{3(2+\alpha)\|g_k\|}{\sigma_{\min}} \right)^{\frac{1}{1+\alpha}} \right\}, \quad k \geq 0.$$



The assumptions  $A.\alpha$ , that the model is minimized globally imply that the  $\alpha \leq 1$  analog of [Cartis, Gould, and Toint \[2011a, Corollary 2.6\]](#) holds, which gives  $\|g_k\| \rightarrow 0$  as  $k \rightarrow \infty$ , and so  $\{\|g_k\|\}, k \geq 0$ , is bounded above. The bound (2.5) now follows from (2.28).

Using the same techniques as in [Cartis, Gould, and Toint \[ibid., Lemma 5.2\]](#) that applies when  $f$  satisfies A.1, it is easy to show for the more general  $A.\alpha$  case that  $\sigma_k \leq c_\sigma \max(\sigma_0, L_{H,\alpha})$  for all  $k$ , where  $c_\sigma$  is a constant depending solely on  $\alpha$  and algorithm parameters. It then follows from (2.25) that (2.11) holds and therefore that the  $(2 + \alpha)$ -regularization method belongs to  $\mathfrak{M}.\alpha$  for  $\alpha \in (0, 1]$ .  $\square$

We cannot extend this result to the  $\alpha = 0$  case unless we also assume that  $H_k$  is positive semi-definite. If this is the case, further examination of the proof of [Cartis, Gould, and Toint \[ibid., Lem.2.2\]](#) allows us to remove the first term in the max in (2.28), and the remainder of the proof is valid.

We note that bounding the regularization parameter  $\sigma_k$  away from zero in (2.26) appears crucial when establishing the bounds (2.5) and (2.4). Requiring (2.26) implies that the Newton step is always perturbed, but does not prevent local quadratic convergence of ARC [Cartis, Gould, and Toint \[2011b\]](#).

**Goldfeld-Quandt-Trotter-type (GQT) methods** [Goldfeld, Quandt, and Trotter \[1966\]](#). Let  $\alpha \in (0, 1]$ . These algorithms set  $M_k = \lambda_k I$ , where

$$(2.29) \quad \lambda_k = \begin{cases} 0, & \text{when } \lambda_{\min}(H_k) \geq \omega_k \|g_k\|^{\frac{\alpha}{1+\alpha}}; \\ -\lambda_{\min}(H_k) + \omega_k \|g_k\|^{\frac{\alpha}{1+\alpha}}, & \text{otherwise,} \end{cases}$$

in (2.2), where  $\omega_k > 0$  is a parameter that is adjusted so as to ensure sufficient objective decrease. (Observe that replacing  $\frac{\alpha}{1+\alpha}$  by 1 in the exponent of  $\|g_k\|$  in (2.29) recovers the original method of Goldfeld et al. [Goldfeld, Quandt, and Trotter \[ibid.\]](#).) It is straightforward to check that (2.3) holds for the choice (2.29). Thus the GQT approach takes the pure Newton step whenever the Hessian is locally sufficiently positive definite, and a suitable regularization of this step otherwise. The parameter  $\omega_k$  is increased by a factor, say  $\gamma_1 > 1$ , and  $x_{k+1}$  left as  $x_k$  whenever the step  $s_k$  does not give sufficient decrease in  $f$  (i.e., iteration  $k$  is unsuccessful), namely when

$$(2.30) \quad \rho_k \stackrel{\text{def}}{=} \frac{f_k - f(x_k + s_k)}{f_k - m_k(s_k)} \leq \eta_1,$$

where  $\eta_1 \in (0, 1)$  and

$$(2.31) \quad m_k(s) = f_k + g_k^T s + \frac{1}{2} s^T H_k s$$

is the model (2.6) with  $\beta_k = 0$ . If  $\rho_k > \eta_1$ , then  $\omega_{k+1} \leq \omega_k$  and  $x_{k+1}$  is constructed as in (2.1). Note that the choice (2.29) implies that (2.4) holds, provided  $\omega_k$  is uniformly bounded above. We show that the latter, as well as (2.5), hold for functions in  $A.\alpha$ .

**Lemma 2.6.** *Let  $f$  satisfy A. $\alpha$  with  $\alpha \in (0, 1]$ . Consider minimizing  $f$  by applying a GQT method that sets  $\lambda_k$  in (2.2) according to (2.29), measures progress according to (2.30), and chooses the parameter  $\omega_k$  and the residual  $r_k$  to satisfy, for  $k \geq 0$ ,*

$$(2.32) \quad \omega_k \geq \omega_{\min} \quad k \geq 0. \quad \text{and} \quad r_k^T s_k \leq 0.$$

*Then (2.5) and (2.4) hold, and so the GQT method belongs to  $\mathfrak{M}.\alpha$ .*

Note that the second part of (2.32) merely requires that  $s_k$  is not longer that the line minimum of the regularized model along the direction  $s_k$ , that is  $1 \leq \arg \min_{\tau \geq 0} m_k(\tau s_k)$ .

*Proof.* Let us first show (2.5). Since  $\omega_k > 0$ , and  $g_k + r_k \neq 0$  until termination, the choice of  $\lambda_k$  in (2.29) implies that  $\lambda_k + \lambda_{\min}(H_k) > 0$ , for all  $k$ , and so (2.2) provides

$$(2.33) \quad s_k = -(H_k + \lambda_k I)^{-1}(g_k + r_k),$$

and hence,

$$(2.34) \quad \|s_k\| \leq \|(H_k + \lambda_k I)^{-1}\| \cdot \|g_k + r_k\| = \frac{(1 + \kappa_{rg})\|g_k\|}{\lambda_k + \lambda_{\min}(H_k)}, \quad k \geq 0.$$

It follows from (2.29) and the first part of (2.32) that, for all  $k \geq 0$ ,

$$(2.35) \quad \lambda_k + \lambda_{\min}(H_k) \geq \omega_k \|g_k\|^{\frac{\alpha}{1+\alpha}} \geq \omega_{\min} \|g_k\|^{\frac{\alpha}{1+\alpha}},$$

This and (2.34) further give

$$(2.36) \quad \|s_k\| \leq \frac{(1 + \kappa_{rg})\|g_k\|^{\frac{1}{1+\alpha}}}{\omega_{\min}}, \quad k \geq 0.$$

As global convergence assumptions are satisfied when  $f$  in A. $\alpha$  Conn, Gould, and Toint [2000] and Goldfeld, Quandt, and Trotter [1966], we have  $\|g_k\| \rightarrow 0$  as  $k \rightarrow \infty$  (in fact, we only need the gradients  $\{g_k\}$  to be bounded). Thus (2.36) implies (2.5).

Due to (2.29), (2.4) holds if we show that  $\{\omega_k\}$  is uniformly bounded above. For this, we first need to estimate the model decrease. Taking the inner product of (2.2) with  $s_k$ , we obtain that

$$-g_k^T s_k = s_k^T H_k s_k + \lambda_k \|s_k\|^2 - r_k^T s_k.$$

Substituting this into the model decrease, we deduce also from (2.6) with  $\beta_k = 0$  that

$$\begin{aligned} f_k - m_k(s_k) &= -g_k^T s_k - \frac{1}{2} s_k^T H_k s_k = \frac{1}{2} s_k^T H_k s_k + \lambda_k \|s_k\|^2 - r_k^T s_k \\ &\geq \left(\frac{1}{2} \lambda_{\min}(H_k) + \lambda_k\right) \|s_k\|^2. \end{aligned}$$

where we used the second part of (2.32) to obtain the last inequality. It is straightforward to check that this and (2.35) now imply

$$(2.37) \quad f_k - m_k(s_k) \geq \frac{1}{2} \omega_k \|g_k\|^{\frac{\alpha}{1+\alpha}} \cdot \|s_k\|^2.$$

We show next that iteration  $k$  is successful for  $\omega_k$  sufficiently large. From (2.30) and second-order Taylor expansion of  $f(x_k + s_k)$ , we deduce

$$|\rho_k - 1| = \left| \frac{f(x_k + s_k) - m_k(s_k)}{f_k - m_k(s_k)} \right| \leq \frac{|H_k - H(\xi_k)| \cdot \|s_k\|^2}{2(f_k - m_k(s_k))} \leq \frac{L_{H,\alpha} \|s_k\|^{2+\alpha}}{2(f_k - m_k(s_k))}.$$

This and (2.37) now give

$$(2.38) \quad |\rho_k - 1| \leq \frac{L_{H,\alpha} \|s_k\|^\alpha}{\omega_k \|g_k\|^{\frac{\alpha}{1+\alpha}}} \leq \frac{L_{H,\alpha}}{\omega_{\min}^\alpha \omega_k},$$

where to obtain the last inequality, we used (2.36). Due to (2.30), iteration  $k$  is successful when  $|\rho_k - 1| \leq 1 - \eta_1$ , which from (2.38) is guaranteed to hold whenever  $\omega_k \geq \frac{L_{H,\alpha}}{\omega_{\min}^\alpha (1 - \eta_1)}$ . As on each successful iteration we set  $\omega_{k+1} \leq \omega_k$ , it follows that

$$(2.39) \quad \omega_k \leq \bar{\omega} \stackrel{\text{def}}{=} \max \left\{ \omega_0, \frac{\gamma_1 L_{H,\alpha}}{\omega_{\min}^\alpha (1 - \eta_1)} \right\}, \quad k \geq 0,$$

where the max term addresses the situation at the starting point and the  $\gamma_1$  factor is included in case an iteration was unsuccessful and close to the bound. This concludes proving (2.4).  $\square$

**Trust-region algorithms** Conn, Gould, and Toint [2000]. These methods compute the correction  $s_k$  as the global solution of the subproblem

$$(2.40) \quad \text{minimize } f_k + g_k^T s + \frac{1}{2} s^T H_k s \quad \text{subject to } \|s\| \leq \Delta_k,$$

where  $\Delta_k$  is an evolving trust-region radius that is chosen to ensure sufficient decrease of  $f$  at  $x_k + s_k$ . The resulting global minimizer satisfies (2.2)–(2.3) Conn, Gould, and Toint [ibid., Corollary 7.2.2] with  $M_k = \lambda_k I$  (or  $M_k = \lambda_k N_k$  if scaling is considered) and  $r_k = 0$ . The scalar  $\lambda_k$  is the Lagrange multiplier of the trust-region constraint, satisfies

$$(2.41) \quad \lambda_k \geq \max\{0, -\lambda_{\min}(H_k)\}$$

and is such that  $\lambda_k = 0$  whenever  $\|s_k\| < \Delta_k$  (and then,  $s_k$  is the Newton step) or calculated using (2.2) to ensure that  $\|s_k\| = \Delta_k$ . The scalar  $\beta_k = 0$  in (2.6). The iterates are defined by (2.1) whenever sufficient progress can be made in some relative function

decrease (so-called *successful iterations*), and they remain unchanged otherwise (*unsuccessful iterations*) while  $\Delta_k$  is adjusted to improve the model (decreased on unsuccessful iterations, possibly increased on successful ones). The total number of unsuccessful iterations is bounded above by a constant multiple of the successful ones plus a (negligible) term in  $\log \epsilon$  [Gratton, Sartenaer, and Toint \[2008, page 23\]](#) provided  $\Delta_k$  is not increased too fast on successful iterations. One successful iteration requires one gradient and one function evaluation while an unsuccessful one only evaluates the objective.

The property (2.5) of  $\mathfrak{M}.\alpha$  methods can be easily shown for trust-region methods, see [Lemma 2.7](#). It is unclear however, whether conditions (2.4) or (2.11) can be guaranteed in general for functions in  $A.\alpha$ . The next lemma gives conditions ensuring a uniform upper bound on the multiplier  $\lambda_k$ , which still falls short of (2.4) in general.

**Lemma 2.7.** *Let  $f$  satisfy assumptions A.0. Consider minimizing  $f$  by applying a trust-region method as described in [Conn, Gould, and Toint \[2000, Algorithm 6.1.1\]](#), where the trust-region subproblem is minimized globally to compute  $s_k$  and where the trust-region radius is chosen to ensure that*

$$(2.42) \quad \Delta_k \leq \Delta_{\max}, \quad k \geq 0,$$

for some  $\Delta_{\max} > 0$ . Then (2.5) holds. Additionally, if

$$(2.43) \quad \|g_{k+1}\| \leq \|g_k\|, \quad \text{for all } k \text{ sufficiently large,}$$

then  $\lambda_k \leq \lambda_{\max}$  for all  $k$  and some  $\lambda_{\max} > 0$ , and  $\lambda_{\min}(M_k)$  is bounded.

*Proof.* Consider the basic trust-region algorithm as described in [Conn, Gould, and Toint \[ibid., Algorithm 6.1.1\]](#), using the same notation. Since the global minimizer  $s_k$  of the trust-region subproblem is feasible with respect to the trust-region constraint, we have  $\|s_k\| \leq \Delta_k$ , and so (2.5) follows trivially from (2.42).

Clearly, the upper bound on  $\lambda_k$  holds whenever  $\lambda_k = 0$  or  $\lambda_k = -\lambda_{\min}(H_k) \leq L_g$ . Thus it is sufficient to consider the case when  $\lambda_k > 0$  and  $H_k + \lambda_k I \succ 0$ . The first condition implies that the trust-region constraint is active, namely  $\|s_k\| = \Delta_k$  [Conn, Gould, and Toint \[ibid., Corollary 7.2.2\]](#). The second condition together with (2.2) implies, as in the proof of [Lemma 2.2](#), that (2.12) holds. Thus we deduce

$$\Delta_k \leq \frac{\|g_k\|}{\lambda_k + \lambda_{\min}(H_k)},$$

or equivalently,

$$(2.44) \quad \lambda_k \leq \frac{\|g_k\|}{\Delta_k} - \lambda_{\min}(H_k) \leq \frac{\|g_k\|}{\Delta_k} + L_g, \quad k \geq 0.$$

It remains to show that

$$(2.45) \quad \{\|g_k\|/\Delta_k\} \text{ is bounded above independently of } k.$$

By [Conn, Gould, and Toint \[2000, Theorem 6.4.2\]](#), we have that there exists  $c \in (0, 1)$  such that the implication holds

$$(2.46) \quad \Delta_k \leq c\|g_k\| \implies \Delta_{k+1} \geq \Delta_k, \text{ i.e., } k \text{ is successful.}$$

(Observe that the Cauchy model decrease condition [Conn, Gould, and Toint \[ibid., Theorem 6.3.3\]](#) is sufficient to obtain the above implication.) Let  $\gamma_1 \in (0, 1)$  denote the largest factor we allow  $\Delta_k$  to be decreased by (during unsuccessful iterations). Using a similar argument to that of [Conn, Gould, and Toint \[ibid., Theorem 6.4.3\]](#), we let  $k \geq k_0$  be the first iterate such that

$$(2.47) \quad \Delta_{k+1} < c\gamma_1\|g_{k+1}\|,$$

where  $k_0$  is the iteration from which onwards (2.43) holds. Then since  $\Delta_{k+1} \geq \gamma_1\Delta_k$  and from (2.43) we have that  $\Delta_k < c\|g_k\|$ . This and (2.46) give

$$\Delta_{k+1} \geq \Delta_k \geq c\gamma_1\|g_k\| \geq c\gamma_1\|g_{k+1}\|,$$

where to obtain the second and third inequalities, we used the hypothesis and (2.43), respectively. We have reached a contradiction with our assumption that  $k+1$  is the first iteration greater than  $k_0$  such that (2.47) holds. Hence there is no such  $k$  and we deduce that

$$(2.48) \quad \Delta_k \geq \min\{\Delta_{k_0}, c\gamma_1\|g_k\|\} \text{ for all } k \geq k_0.$$

Note that since  $g_k$  remains unchanged on unsuccessful iterations, (2.43) trivially holds on such iterations. Since the assumptions of [Conn, Gould, and Toint \[ibid., Theorem 6.4.6\]](#) are satisfied, we have that  $\|g_k\| \rightarrow 0$ , as  $k \rightarrow \infty$ . This and (2.48) imply (2.45). The desired conclusion then follows from (2.44).  $\square$

Note that if (2.19) holds for some  $\alpha \in [0, 1]$ , then (2.43) is satisfied, and so [Lemma 2.7](#) shows that if (2.19) holds, then (2.18) is satisfied. It follows from [Lemma 2.4](#) that fast convergence of trust-region methods for functions in  $\mathbf{A}.\alpha$  alone is sufficient to ensure (2.16), which in turn is connected to our definition of the class  $\mathfrak{M}.\alpha$ . However, the properties of the multipliers (in the sense of (2.4) for any  $\alpha \in [0, 1]$  or even (2.16)) remain unclear in the absence of fast convergence of the method. Based on our experience, we are inclined to believe that generally, the multipliers  $\lambda_k$  are at best guaranteed to be uniformly bounded

above, even for specialized, potentially computationally expensive, rules of choosing the trust-region radius.

As the Newton step is taken in the trust-region framework satisfying (2.2) whenever it is within the trust region and gives sufficient decrease in the presence of local convexity, the A.1- (hence A. $\alpha$ -) example of inefficient behaviour for Newton's method of worst-case evaluation complexity precisely  $\epsilon^{-2}$  can be shown to apply also to trust-region methods [Cartis, Gould, and Toint \[2010\]](#) (see also [Gratton, Sartenar, and Toint \[2008\]](#)).

**Linesearch methods** [Dennis and Schnabel \[1983\]](#) and [Nocedal and Wright \[1999\]](#). We finally consider methods using a linesearch to control improvement in the objective at each step. Such methods compute  $x_{k+1} = x_k + s_k$ ,  $k \geq 0$ , where  $s_k$  is defined via (2.2) in which  $M_k$  is chosen so that  $H_k + M_k$ , the Hessian of the selected quadratic model  $m_k(s)$ , is “sufficiently” positive definite, and  $r_k = (1 - \mu_k)g_k$ , yielding a stepsize  $\mu_k \in [1 - \kappa_{rg}, 1]$  which is calculated so as to decrease  $f$  (the *linesearch*); this is always possible for sufficiently small  $\mu_k$  (and hence sufficiently small  $\kappa_{rg}$ .) The precise definition of “sufficient decrease” depends on the particular linesearch scheme being considered, but we assume here that

$$\mu_k = 1 \text{ is acceptable whenever } m_k(s_k) = f(x_k + s_k).$$

In other words, we require the unit step to be acceptable when the model and the true objective function match at the trial point. Because the minimization of the quadratic model along the step always ensure that  $m_k(s_k) = f(x_k) + \frac{1}{2}g_k s_k$ , the above condition says that  $s_k$  must be acceptable with  $\mu_k = 1$  whenever  $f(x_k + s_k) = f(x_k) + \frac{1}{2}g_k s_k$ . This is for instance the case for the Armijo and Goldstein linesearch conditions<sup>(4)</sup>, two standard linesearch techniques. As a consequence, the corresponding linesearch variants of Newton's method and of the  $(2 + \alpha)$ -regularization methods also belong to  $\mathfrak{M}.\alpha$  (with  $\beta_k = 1$  for all  $k$ ), and the list is not exhaustive. Note that linesearch methods where the search direction is computed inexactly are also covered by setting  $r_k = g_k - \mu_k(g_k + w_k)$  for some “error vector”  $w_k$ , provided the second part of (2.2) still holds.

### 3 Examples of inefficient behaviour

After reviewing the methods in  $\mathfrak{M}.\alpha$ , we now turn to showing they can converge slowly when applied to specific functions with fixed range<sup>(5)</sup> and the relevant degree of smoothness.

<sup>(4)</sup>With reasonable algorithmic constants, see Appendix A1.

<sup>(5)</sup>At variance with the examples proposed in [Cartis, Gould, and Toint \[2010, 2011d\]](#).

**3.1 General methods in  $\mathfrak{M}.\alpha$ .** Let  $\alpha \in [0, 1]$  and  $\epsilon \in (0, 1)$  be given and consider an arbitrary method in  $\mathfrak{M}.\alpha$ . Our intent is now to construct a univariate function  $f_\epsilon^{\mathfrak{M}.\alpha}(x)$  satisfying A. $\alpha$  such that

$$(3.1) \quad f_\epsilon^{\mathfrak{M}.\alpha}(0) = 1, \quad f_\epsilon^{\mathfrak{M}.\alpha}(x) \in [a, b] \quad \text{for } x \geq 0,$$

for some constants  $a \leq b$  independent of  $\epsilon$  and  $\alpha$ , and such that the method will terminate in exactly

$$(3.2) \quad k_{\epsilon, \alpha} = \left\lceil \epsilon^{-\frac{2+\alpha}{1+\alpha}} \right\rceil$$

iterations (and evaluations of  $f$ ,  $g$  and  $H$ ).

We start by defining the sequences  $f_k$ ,  $g_k$  and  $H_k$  for  $k = 0, \dots, k_{\epsilon, \alpha}$  by

$$(3.3) \quad f_k = 1 - \frac{1}{2}k\epsilon^{\frac{2+\alpha}{1+\alpha}}, \quad g_k = -2\epsilon f_k \quad \text{and} \quad H_k = 4\epsilon^{\frac{\alpha}{1+\alpha}} f_k^2.$$

They are intended to specify the objective function, gradient and Hessian values at successive iterates generated by the chosen method in  $\mathfrak{M}.\alpha$ , according to (2.1) and (2.2) for some choice of multipliers  $\{\lambda_k\} = \{M_k\} = \{\lambda_{\min}(M_k)\}$  satisfying (2.3) and (2.4). In other words, we impose that  $f_k = f_\epsilon^{\mathfrak{M}.\alpha}(x_k)$ ,  $g_k = \nabla f_\epsilon^{\mathfrak{M}.\alpha}(x_k)$  and  $H_k = \nabla^2 f_\epsilon^{\mathfrak{M}.\alpha}(x_k)$  for  $k \in \mathcal{K} \stackrel{\text{def}}{=} \{0, \dots, k_{\epsilon, \alpha}\}$ . Note that  $f_k$ ,  $|g_k|$  and  $H_k$  are monotonically decreasing and that, using (3.2),

$$(3.4) \quad f_k \in [\tfrac{1}{2}, 1] \quad \text{for } k \in \mathcal{K}.$$

In addition, (2.3) and (2.4) impose that, for  $k \in \mathcal{K}$ ,

$$0 \leq \lambda_k + 4\epsilon^{\frac{\alpha}{1+\alpha}} f_k^2 \leq \kappa_\lambda \max[4\epsilon^{\frac{\alpha}{1+\alpha}} f_k^2, (2\epsilon f_k)^{\frac{\alpha}{1+\alpha}}] = 4\kappa_\lambda \epsilon^{\frac{\alpha}{1+\alpha}} f_k^2.$$

yielding that

$$(3.5) \quad \lambda_k \in \left[0, 4(\kappa_\lambda - 1)\epsilon^{\frac{\alpha}{1+\alpha}} f_k^2\right],$$

As a consequence, we obtain, using both parts of (2.2), that, for  $k \in \mathcal{K}$ ,

$$(3.6) \quad s_k = \theta_k \frac{\epsilon^{\frac{1}{1+\alpha}}}{2f_k} \quad \text{for some } \theta_k \in \left[\frac{1 - \kappa_{rg}}{\kappa_\lambda}, 1 + \kappa_{rg}\right].$$

Note that our construction imposes that

$$(3.7) \quad \begin{aligned} m_k(s_k) &= f_k + g_k s_k + \frac{1}{2}g_k s_k + \frac{1}{2}s_k(H_k + \beta_k \lambda_k)s_k \\ &= f_k + g_k s_k + \frac{1}{2}s_k[-g_k + r_k + (\beta_k - 1)\lambda_k s_k] \\ &\geq f_k - \frac{1}{2}|g_k|s_k - \frac{1}{2}\kappa_{rg}|g_k|s_k + \frac{1}{2}\theta_k^2(\kappa_\lambda - 1)(\beta_k - 1)\epsilon^{\frac{2+\alpha}{1+\alpha}} \\ &\geq f_k - \frac{1}{2}\theta_k \epsilon^{\frac{2+\alpha}{1+\alpha}} [1 + \kappa_{rg} + \theta_k(1 - \beta_k)(\kappa_\lambda - 1)] \\ &\geq f_k - \frac{1}{2}\epsilon^{\frac{2+\alpha}{1+\alpha}} (1 + \kappa_{rg})^2 [1 + (1 - \beta_k)(\kappa_\lambda - 1)] \\ &\geq f_k - \frac{1}{2}\epsilon^{\frac{2+\alpha}{1+\alpha}} (1 + \kappa_{rg})^2 \kappa_\lambda \end{aligned}$$

where we have used (2.2), (3.3), (3.6), (3.5) and  $\beta_k \leq 1$ . Hence, again taking (3.3) into account,

$$(3.8) \quad \frac{f_k - f_{k+1}}{f_k - m_k(s_k)} \geq \frac{\frac{1}{2}\epsilon^{\frac{2+\alpha}{1+\alpha}}}{\frac{1}{2}\epsilon^{\frac{2+\alpha}{1+\alpha}}\kappa_\lambda(1+\kappa_{rg})^2} = \frac{1}{(1+\kappa_{rg})^2\kappa_\lambda} \in (0, 1),$$

and sufficient decrease of the objective function automatically follows. Moreover, given (3.4), we deduce from (3.6) that  $|s_k| \leq 1$  for  $k \in \mathbb{K}$  and (2.5) holds with  $\kappa_s = 1$ , as requested for a method in  $\mathfrak{M}.\alpha$ . It also follows from (2.1) and (3.6) that, if  $x_0 = 0$ ,

$$(3.9) \quad s_k > 0 \quad \text{and} \quad x_k = \sum_{i=0}^{k-1} s_i, \quad k = 0, \dots, k_{\epsilon, \alpha}.$$

We therefore conclude that the sequences  $\{f_k\}_{k=0}^{k_{\epsilon, \alpha}}$ ,  $\{g_k\}_{k=0}^{k_{\epsilon, \alpha}}$ ,  $\{H_k\}_{k=0}^{k_{\epsilon, \alpha}}$ ,  $\{\lambda_k\}_{k=0}^{k_{\epsilon, \alpha}-1}$  and  $\{s_k\}_{k=0}^{k_{\epsilon, \alpha}-1}$  can be viewed as produced by our chosen method in  $\mathfrak{M}.\alpha$ , and, from (3.3), that termination occurs precisely for  $k = k_{\epsilon, \alpha}$ , as desired.

We now construct the function  $f_\epsilon^{\mathfrak{M}.\alpha}(x)$  for  $x \in [0, x_{k_{\epsilon, \alpha}}]$  using Hermite interpolation. We set

$$(3.10) \quad f_\epsilon^{\mathfrak{M}.\alpha}(x) = p_k(x - x_k) + f_{k+1} \quad \text{for } x \in [x_k, x_{k+1}] \text{ and } k = 0, \dots, k_{\epsilon, \alpha} - 1,$$

where  $p_k$  is the polynomial

$$p_k(s) = c_{0,k} + c_{1,k}s + c_{2,k}s^2 + c_{3,k}s^3 + c_{4,k}s^4 + c_{5,k}s^5,$$

with coefficients defined by the interpolation conditions

$$(3.11) \quad \begin{aligned} p_k(0) &= f_k - f_{k+1}, & p_k(s_k) &= 0; \\ p'_k(0) &= g_k, & p'_k(s_k) &= g_{k+1}; \\ p''_k(0) &= H_k, & p''_k(s_k) &= H_{k+1}, \end{aligned}$$

where  $s_k$  is defined in (3.6). These conditions yield the following values for the coefficients

$$(3.12) \quad c_{0,k} = f_k - f_{k+1}, \quad c_{1,k} = g_k, \quad c_{2,k} = \frac{1}{2}H_k;$$

with the remaining coefficients satisfying

$$\begin{pmatrix} s_k^3 & s_k^4 & s_k^5 \\ 3s_k^2 & 4s_k^3 & 5s_k^4 \\ 6s_k & 12s_k^2 & 20s_k^3 \end{pmatrix} \begin{pmatrix} c_{3,k} \\ c_{4,k} \\ c_{5,k} \end{pmatrix} = \begin{pmatrix} \Delta f_k - g_k s_k - \frac{1}{2}s_k^T H_k s_k \\ \Delta g_k - H_k s_k \\ \Delta H_k \end{pmatrix},$$



where

$$\Delta f_k = f_{k+1} - f_k, \quad \Delta g_k = g_{k+1} - g_k \quad \text{and} \quad \Delta H_k = H_{k+1} - H_k.$$

Hence we obtain after elementary calculations that

$$(3.13) \quad \begin{aligned} c_{3,k} &= 10 \frac{\Delta f_k}{s_k^3} - 4 \frac{\Delta g_k}{s_k^2} + \frac{\Delta H_k}{2s_k} - 10 \frac{g_k}{s_k^2} - \frac{H_k}{s_k}; \\ c_{4,k} &= -15 \frac{\Delta f_k}{s_k^4} + 7 \frac{\Delta g_k}{s_k^3} - \frac{\Delta H_k}{s_k^2} + 15 \frac{g_k}{s_k^3} + \frac{H_k}{2s_k^2}; \\ c_{5,k} &= 6 \frac{\Delta f_k}{s_k^5} - 3 \frac{\Delta g_k}{s_k^4} + \frac{\Delta H_k}{2s_k^3} - 6 \frac{g_k}{s_k^4}; \end{aligned}$$

The top three graphs of Figure 3.1 illustrate the global behaviour of the resulting function  $f_\epsilon^{\mathfrak{m},\alpha}(x)$  and of its first and second derivatives for  $x \in [0, x_{k_{\epsilon,\alpha}}]$ , while the bottom ones show more detail of the first 10 iterations. The figure is constructed using  $\epsilon = 5 \cdot 10^{-2}$  and  $\alpha = \frac{1}{2}$ , which then yields that  $k_{\epsilon,\alpha} = 148$ . In addition, we set  $\lambda_k = \frac{1}{10} |g_k|^{\frac{\alpha}{1+\alpha}}$  for  $k = 0, \dots, k_{\epsilon,\alpha}$ . The nonconvexity of  $f_\epsilon^{\mathfrak{m},\alpha}(x)$  is clear from the bottom graphs.

**Lemma 3.1.** *The function  $f_\epsilon^{\mathfrak{m},\alpha}$  defined above on the interval  $[0, x_{k_{\epsilon,\alpha}}]$  can be extended to a function from  $\mathbb{R}$  to  $\mathbb{R}$  satisfying A.α and whose range is bounded independently of α and ε.*

*Proof.* We start by showing that, on

$$[0, x_{k_{\epsilon,\alpha}}] = \bigcup_{k \in \mathbb{K}} [x_k, x_k + s_k],$$

$f_\epsilon^{\mathfrak{m},\alpha}$  is bounded in absolute value independently of  $\epsilon$  and  $\alpha$ , twice continuously differentiable with Lipschitz continuous gradient and  $\alpha$ -Hölder continuous Hessian. Recall first (3.10) provide that  $f_\epsilon^{\mathfrak{m},\alpha}$  is twice continuously differentiable by construction on  $[0, x_{k_{\epsilon,\alpha}}]$ . It thus remains to investigate the gradient's Lipschitz continuity and Hessian's  $\alpha$ -Hölder continuity, as well as whether  $|f_\epsilon^{\mathfrak{m},\alpha}(x)|$  is bounded on this interval.

Defining now

$$(3.14) \quad \pi_k \stackrel{\text{def}}{=} \frac{\theta_k}{2} \frac{2f_k - 1}{f_k} \in [0, \frac{1}{2}\theta_k] \quad \text{and} \quad \phi(\theta) \stackrel{\text{def}}{=} 2 - \frac{1}{\theta} \in [2 - \frac{\kappa_\lambda}{1 - \kappa_{rg}}, 1 + \kappa_{rg}]$$

(where we used (3.4) and (3.6)), we obtain from (3.2), (3.3), (3.6) and (3.13), that, for  $k \in \mathbb{K}$ ,

$$(3.15) \quad \begin{aligned} |c_{3,k}| s_k^2 &= \epsilon f_k \left( 20 - \frac{10}{\theta_k} - 2\theta_k \right) - \epsilon^{\frac{3+2\alpha}{1+\alpha}} (4 + \pi_k) \leq \epsilon \left[ 10|\phi(\theta)| + 2\theta + \frac{9}{2} \epsilon^{\frac{2+\alpha}{1+\alpha}} \right] = \mathcal{O}(\epsilon), \\ |c_{4,k}| s_k^3 &= \epsilon f_k \left( \frac{15}{\theta_k} - 30 + \theta_k \right) + \epsilon^{\frac{3+2\alpha}{1+\alpha}} (7 + 2\pi_k) \leq \epsilon \left[ 15|\phi(\theta)| + \theta + 8\epsilon^{\frac{2+\alpha}{1+\alpha}} \right] = \mathcal{O}(\epsilon), \\ |c_{5,k}| s_k^4 &= \epsilon f_k \left( 12 - \frac{6}{\theta_k} \right) - \epsilon^{\frac{3+2\alpha}{1+\alpha}} (3 + \pi_k) \leq \epsilon \left[ 6|\phi(\theta)| + \frac{7}{2} \epsilon^{\frac{2+\alpha}{1+\alpha}} \right] = \mathcal{O}(\epsilon), \end{aligned}$$

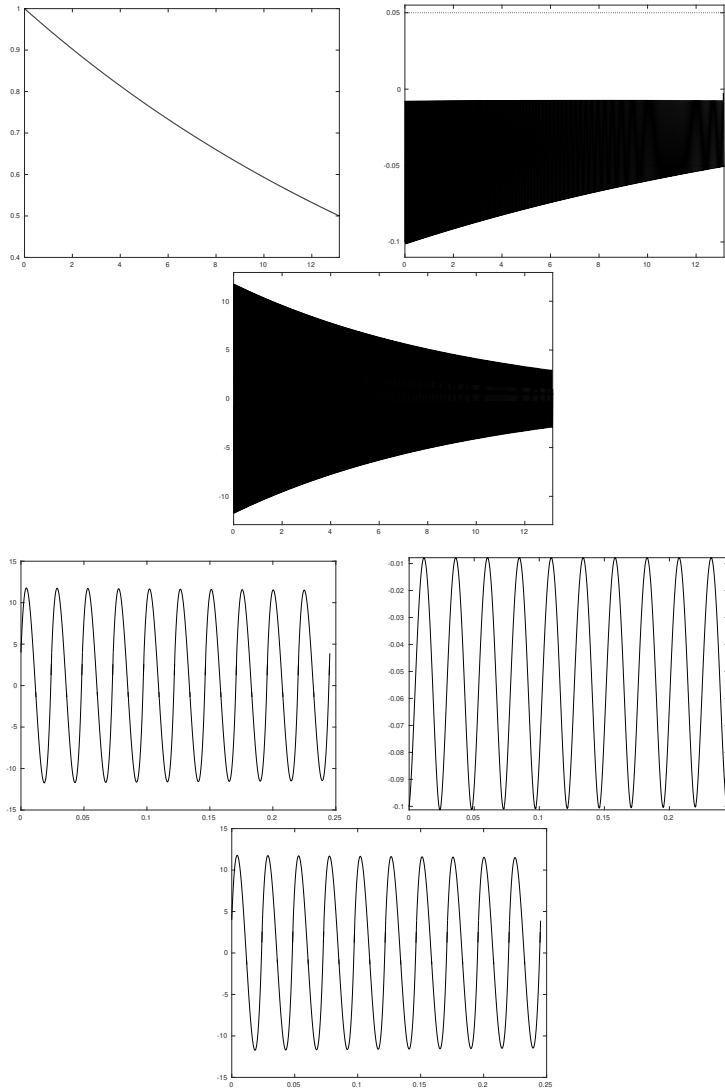


Figure 3.1:  $f_{\epsilon}^{\mathfrak{m}, \alpha}(x)$  (left) and its first (center) and second (right) derivatives as a function of  $x$  for  $\alpha = \frac{1}{2}$  and  $\epsilon = 5.10^{-2}$  (top:  $x \in [0, x_{k_{\epsilon, \alpha}}]$ ; bottom:  $x \in [0, x_{10}]$ ). Horizontal dotted lines indicate values of  $-\epsilon$  and  $\epsilon$  in the central top graph.

where we also used  $\epsilon \leq 1$  and (3.4). To show that the Hessian of  $f_\epsilon^{\mathfrak{m},\alpha}$  is globally  $\alpha$ -Hölder continuous on  $[0, x_{k_{\epsilon,\alpha}}]$ , we need to verify that (2.8) holds for all  $x, y$  in this interval. From (3.10), this is implied by

$$(3.16) \quad |p'''(s)| \leq c|s|^{-1+\alpha}, \quad \text{for all } s \in [0, s_k] \text{ and } k \in \mathcal{K},$$

for some  $c > 0$  independent of  $\epsilon$ ,  $s$  and  $k$ . We have from the expression of  $p_k$  and  $s \in [0, s_k]$  that

$$(3.17) \quad \begin{aligned} |p_k'''(s)| \cdot |s|^{1-\alpha} &\leq (6|c_{3,k}| + 24|c_{4,k}|s_k + 60|c_{5,k}|s_k^2)s_k^{1-\alpha} \\ &= (6|c_{3,k}|s_k^2 + 24|c_{4,k}|s_k^3 + 60|c_{5,k}|s_k^4)s_k^{-(1+\alpha)}. \end{aligned}$$

The boundedness of this last right-hand side on  $[0, x_{k_{\epsilon,\alpha}}]$ , and thus the  $\alpha$ -Hölder continuity of the Hessian of  $f^M$ , then follow from (3.15), (3.6) and (3.4).

Similarly, to show that the gradient of  $f^M$  is globally Lipschitz continuous in  $[0, x_{k_{\epsilon,\alpha}}]$  is equivalent to proving that  $p_k''(s)$  is uniformly bounded above on the interval  $[0, s_k]$  for  $k \in \mathcal{K}$ . Since  $s_k > 0$ , we have

$$(3.18) \quad \begin{aligned} |p_k''(s)| &\leq 2|c_{2,k}| + 6|c_{3,k}|s_k + 12|c_{4,k}|s_k^2 + 20|c_{5,k}|s_k^3 \\ &= 2|c_{2,k}| + (6|c_{3,k}|s_k^2 + 12|c_{4,k}|s_k^3 + 20|c_{5,k}|s_k^4)s_k^{-1}. \end{aligned}$$

Then the third part of (3.3) and the bounds  $\epsilon \leq 1$ , (3.15), (3.12), (3.6) and (3.4) again imply the boundedness of the last right-hand side on  $[0, x_{k_{\epsilon,\alpha}}]$ , as requested. Finally, the fact that  $|f_\epsilon^{\mathfrak{m},\alpha}|$  is bounded on  $[0, x_{k_{\epsilon,\alpha}}]$  results from the observation that, on the interval  $[0, s_k]$  with  $k \in \mathcal{K}$ ,

$$|p_k(s)| \leq f_k + |g_k||s_k| + \frac{1}{2}|H_k||s_k|^2 + (|c_{3,k}|s_k^2 + |c_{4,k}|s_k^3 + |c_{5,k}|s_k^4)s_k$$

from which a finite bound  $a$  independent from  $\alpha$  and  $\epsilon$  again follows from  $\epsilon \leq 1$ , (3.3), (3.10), (3.15), (3.12), (3.6) and (3.4). We have thus proved that  $f_\epsilon^{\mathfrak{m},\alpha}$  satisfies the desired properties on  $[0, x_{k_{\epsilon,\alpha}}]$ .

We may then smoothly prolongate  $f_\epsilon^{\mathfrak{m},\alpha}$  for  $x \in \mathbb{R}$ , for instance by defining two additional interpolation intervals  $[x_{-1}, x_0] = [-1, 0]$  and  $[x_{k_{\epsilon,\alpha}}, x_{k_{\epsilon,\alpha}} + 1]$  with end conditions

$$f_{-1} = 1, \quad f_{k_{\epsilon,\alpha}+1} = f_{k_{\epsilon,\alpha}} \quad \text{and} \quad g_{-1} = H_{-1} = g_{k_{\epsilon,\alpha}+1} = H_{k_{\epsilon,\alpha}+1} = 0,$$

and setting

$$f_\epsilon^{\mathfrak{m},\alpha}(x) = \begin{cases} 1 & \text{for } x \leq -1, \\ p_k(x - x_k) + f_{k+1} & \text{for } x \in [x_k, x_{k+1}] \text{ and } k \in \{-1, \dots, k_{\epsilon,\alpha}\}, \\ f_\epsilon^{\mathfrak{m},\alpha}(x_{k_{\epsilon,\alpha}}) & \text{for } x \geq x_{k_{\epsilon,\alpha}} + 1, \end{cases}$$

which subsumes (3.10). Using arguments similar to those used above, it is easy to verify from (3.12), (3.13) and  $s_{-1} = s_{k_{\epsilon,\alpha}} = 1$  that all desired properties are maintained.  $\square$

We formulate the results of this development in the following theorem.

**Theorem 3.2.** *For every  $\epsilon \in (0, 1)$ , every  $\alpha \in [0, 1]$  and every method in  $\mathfrak{M}.\alpha$ , a function  $f_\epsilon^{\mathfrak{M}.\alpha}$  satisfying A. $\alpha$  with values in a bounded interval independent of  $\epsilon$  and  $\alpha$  can be constructed, such, when applied to  $f_\epsilon^{\mathfrak{M}.\alpha}$ , the considered method terminates exactly at iteration*

$$k_{\epsilon,\alpha} = \left\lceil \epsilon^{-\frac{2+\alpha}{1+\alpha}} \right\rceil.$$

*with the first iterate  $x_{k_{\epsilon,\alpha}}$  such that  $\|\nabla_x f_\epsilon^{\mathfrak{M}.\alpha}(x_{k_{\epsilon,\alpha}})\| \leq \epsilon$ .*

Note that the prolongation of  $f_\epsilon^{\mathfrak{M}.\alpha}(x)$  to  $x \geq 0$  suggested as an example in the proof of [Lemma 3.1](#) admits an isolated finite global minimizer. Indeed, since the  $g_{k_{\epsilon,\alpha}} < 0$ , there must be a value lower than  $f(x_{k_{\epsilon,\alpha}})$  in  $(x_{k_{\epsilon,\alpha}}, x_{k_{\epsilon,\alpha}} + 1)$ , and thus the global minimizer must lie in one of the constructed sub-intervals in  $(-1, x_{k_{\epsilon,\alpha}+1})$ ; since  $f_\epsilon^{\mathfrak{M}.\alpha}(x)$  is quintic (and not constant) in each of these, the global minimizer must therefore be isolated.

**3.2 The inexact Newton's method.** It is interesting that the technique developed in the previous subsection can also be used to derive an  $\mathcal{O}(\epsilon^{-2})$  lower bound on worst-case evaluation complexity for an inexact Newton's method applied to a function having Lipschitz continuous Hessians on the path of iterates. This is stronger than using [Theorem 3.2](#) above for  $\alpha = 1$ , as it would result in a weaker  $\mathcal{O}(\epsilon^{-3/2})$  lower bound, or for  $\alpha = 0$  as it would then only guarantee bounded Hessians. In the spirit of [Cartis, Gould, and Toint \[2010\]](#), this new function is constructed by extending to  $\mathbb{R}^2$  the unidimensional  $f_\epsilon^{\mathfrak{M}.0}(x)$  obtained in the previous section for the specific choice  $M_k = 0$ , which then ensures that  $\theta_k \in [1 - \kappa_{rg}, 1 + \kappa_{rg}]$  for all  $k$  (see [\(3.5\)](#) and [\(3.6\)](#)). The proposed extension is of the form

$$(3.19) \quad h_\epsilon^N(x, y) \stackrel{\text{def}}{=} f_\epsilon^{\mathfrak{M}.0}(x) + u_\epsilon(y),$$

where we still have to specify the univariate function  $u_\epsilon$  such that Newton's method applied to  $u_\epsilon$  converges with large steps. In order to define it, we start by redefining

$$k_\epsilon = k_{\epsilon,0} = \lceil \epsilon^{-2} \rceil \quad \text{and} \quad \mathcal{K} = \{0, \dots, k_\epsilon\}.$$

Then we set, for  $k \in \mathcal{K}$ ,

$$(3.20) \quad u_k = 1 - \frac{1}{2}k\epsilon^2, \quad g_k^u = -2\epsilon^2 u_k, \quad H_k^u = 2|g_k^u| u_k > 0,$$

and

$$(3.21) \quad s_k^u = \frac{v_k}{2u_k} \quad \text{with} \quad v_k \in [1 - \kappa_{rg}, 1 + \kappa_{rg}] \quad \text{and} \quad u_k \in [\frac{1}{2}, 1],$$

this definition allowing for

$$H_k^u s_k^u = -g_k^u + r_k^u \quad \text{with} \quad |r_k^u| \leq \kappa_{rg} |g_k^u|.$$

(Remember that  $M_k = 0$  because we are considering Newton's method.) Note that sufficient decrease is obtained in manner similar to (3.7)-(3.8), because of (3.20), (3.21) and  $\lambda_k = 0$ , yielding that  $u_k - u_{k+1} \geq -(g_k^u s_k^u + \frac{1}{2} H_k^u (s_k^u)^2) / (1 + \kappa_{rg})$ . Setting now  $y_0 = 0$  and  $y_{k+1} = y_k + s_k^u$  for  $k \in \{1, \dots, k_\epsilon\}$ , we may then, as in Section 3.1, define

$$(3.22) \quad u_\epsilon(y) = p_k^u(y - y_k) + u_{k+1} \quad \text{for } y \in [y_k, y_{k+1}] \text{ and } k = 0, \dots, k_\epsilon - 1,$$

where  $p_k^u$  is a fifth degree polynomial interpolating the values and derivatives given by (3.20) on the interval  $[0, s_k^u]$ . We then obtain the following result.

**Theorem 3.3.** *For every  $\epsilon \in (0, 1)$ , there exists a function  $h_\epsilon^N$  with Lipschitz continuous gradient and Lipschitz continuous Hessian along the path of iterates  $\cup_{k=0}^{k_\epsilon-1} [x_j, x_{j+1}]$ , and with values in a bounded interval independent of  $\epsilon$ , such that, when applied to  $h_\epsilon^N$ , Newton's terminates exactly at iteration*

$$k_\epsilon = \lceil \epsilon^{-2} \rceil$$

with the first iterate  $x_{k_\epsilon}$  such that  $\|\nabla_x f_\epsilon^{\mathfrak{m},\alpha}(x_{k_\epsilon})\| \leq \epsilon \sqrt{1 + \epsilon^2}$ .

*Proof.* One easily verifies from (3.20), (3.21) and (3.13) that the interpolation coefficients, now denoted by  $|d_{i,k}|$ , are bounded for all  $k \in \{0, \dots, k_\epsilon - 1\}$  and  $i \in \{0, \dots, 5\}$ . This observation and (3.21) in turn guarantee that  $u_\epsilon$  and all its derivatives (including the third) remain bounded on each interval  $[0, s_k^u]$  by constants independent of  $\epsilon$ . As in Lemma 3.1, we next extend  $u_\epsilon$  to the whole of  $\mathbb{R}$  while preserving this property. We then construct  $h^N$  using (3.19). From the properties of  $f_\epsilon^{\mathfrak{m},0}$  and  $u_\epsilon$ , we deduce that  $h_\epsilon^N$  is twice continuously differentiable and has a range bounded independently of  $\epsilon$ . Moreover, it satisfies A.0. When applied on  $h_\epsilon^N(x, y)$ , Newton's generates the iterates  $(x_k, y_k)$  and its gradient at the  $k_\epsilon$ -th iterate is  $(\epsilon, \epsilon^2)$  so that  $\|\nabla h^N(x_{k_\epsilon}, y_{k_\epsilon})\| = \epsilon \sqrt{1 + \epsilon^2}$ , prompting termination. Before that, the algorithm generates the steps  $(s_k, s_k^u)$ , where, because both  $f_k$  and  $u_k$  belong to  $[\frac{1}{2}, 1]$  and because of (3.6) with  $\alpha = 0$ ,

$$(3.23) \quad s_k \in [\epsilon(1 - \kappa_{rg}), 2\epsilon(1 + \kappa_{rg})] \quad \text{and} \quad s_k^u \in [1 - \kappa_{rg}, 2(1 + \kappa_{rg})].$$

Thus the absolute value of the third derivative of  $h_\epsilon^N(x, y)$  is given, for  $(x, y)$  in the  $k$ -th segment of the path of iterates, by

$$\begin{aligned}
 (3.24) \quad & \frac{1}{\|(s_k, s_k^u)\|} \left| p_k'''(x - x_k)s_k^3 + (p_k^u)'''(y - y_k)(s_k^u)^3 \right| \\
 & \leq \frac{1}{1 - \kappa_{rg}} \left[ |p_k'''(x - x_k)|s_k^3 + |(p_k^u)'''(y - y_k)|(s_k^u)^3 \right] \\
 & = \frac{1}{1 - \kappa_{rg}} \left[ \left( 6|c_{3,k}| + 24|c_{4,k}|s_k + 60|c_{5,k}|s_k^2 \right) s_k^3 \right. \\
 & \quad \left. + \left( 6|d_{3,k}| + 24|d_{4,k}|s_k^u + 60|d_{5,k}|(s_k^u)^2 \right) (s_k^u)^3 \right] \\
 & = \frac{1}{1 - \kappa_{rg}} \left[ \left( 6|c_{3,k}|s_k^2 + 24|c_{4,k}|s_k^3 + 60|c_{5,k}|s_k^4 \right) s_k \right. \\
 & \quad \left. + 6|d_{3,k}|(s_k^u)^3 + 24|d_{4,k}|(s_k^u)^4 + 60|d_{5,k}|(s_k^u)^5 \right],
 \end{aligned}$$

where we used the fact that  $\|(s_k, s_k^u)\| \geq \|s_k^u\|$ , and (3.23). But, in view of (3.15), (3.14) with  $\theta_k \in [1 - \kappa_{rg}, 1 + \kappa_{rg}]$ , (3.23),  $\epsilon \leq 1$  and the boundedness of the  $d_{i,k}$ , the last right-hand side of (3.24) is bounded by a constant independent of  $\epsilon$ . Thus the third derivative of  $h_\epsilon^N(x, y)$  is bounded on every segment by the same constant, and, as a consequence, the Hessian of  $h_\epsilon^N(x, y)$  is Lipschitz continuous of each segment, as desired.  $\square$

Note that the same result also holds for any method in  $\mathfrak{M}.0$  with  $M_k$  small enough to guarantee that  $s_k$  is bounded away from zero for all  $k$ .

## 4 Complexity and optimality for methods in $\mathfrak{M}.\alpha$

We now consider the consequences of the examples derived in Section 3 on the evaluation complexity analysis of the various methods identified in Section 2 as belonging to  $\mathfrak{M}.\alpha$ .

**4.1 Newton's method.** First note that the third part of (3.3) ensures that  $H_k > 0$  so that the Newton iteration is well-defined for the choice (2.23). This choice corresponds to setting  $\theta_k = 1$  for all  $k \geq 0$  in the example of Section 3. So we first conclude from Theorem 3.2 that Newton's method may require  $\epsilon^{-(2+\alpha)/(1+\alpha)}$  evaluations when applied on the resulting objective function  $f_\epsilon^{\mathfrak{M}.\alpha}$  satisfying A. $\alpha$  to generate  $|g_k| \leq \epsilon$ . However, Theorem 3.3 provides the stronger result that it may in fact require  $\epsilon^{-2}$  evaluations (as a method in  $\mathfrak{M}.0$ ) for nearly the same task (we traded Lipschitz continuity of the Hessian on the whole space for that along the path of iterates). As a consequence we obtain that *Newton's method is not optimal in  $\mathfrak{M}.\alpha$  as far as worst-case evaluation complexity is concerned.*

The present results also improves on the similar bound given in Cartis, Gould, and Toint [2011d], in that the objective function on Sections 3.1 and 3.2 ensure the existence

of a lower bound  $f_{\text{low}}$  on  $f_\epsilon^{\mathfrak{M},\alpha}(x)$  such that  $f_\epsilon^{\mathfrak{M},\alpha}(x_0) - f_{\text{low}}$  is bounded, while the latter difference is unbounded in [Cartis, Gould, and Toint \[2011d\]](#) (for  $\alpha \in \{0, 1\}$ ) as the number of iterations approaches  $\epsilon^{-2}$ . We will return to the significance of this observation when discussing regularization methods.

Since the steepest-descent method is known to have a worst-case evaluation complexity of  $\mathcal{O}(\epsilon^{-2})$  when applied on functions having Lipschitz continuous gradients [Nesterov \[2004, p. 29\]](#), [Theorem 3.3](#) shows that Newton’s method may, in the worst case, converge as slowly as steepest descent in the worst case. Moreover, we show in [Appendix A1](#) that the quoted worst-case evaluation complexity bound for steepest descent is sharp, which means that steepest-descent and Newton’s method are undistinguishable from the point of view of worst-case complexity orders.

Note also that if the Hessian of the objective is unbounded, and hence, we are outside of the class A.0, the worst-case evaluation complexity of Newton’s method worsens, and in fact, it may be arbitrarily bad [Cartis, Gould, and Toint \[2010\]](#).

**4.2 Cubic and other regularizations.** Recalling our discussion of the  $(2 + \alpha)$ -regularization method in [Section 2.2](#), we first note, in the example of [Section 3.1](#), that, because of [\(2.2\)](#) and [\(2.3\)](#),  $s_k$  is a minimizer of the model [\(2.6\)](#) with  $\beta_k = \lambda_k$  at iteration  $k$ , in that

$$(4.1) \quad m_k(s_k) = f_\epsilon^{\mathfrak{M},\alpha}(x_k + s_k) = f_{k+1}$$

for  $k \in \mathcal{K}$ . Thus every iteration is successful as the objective function decrease exactly matches decrease in the model. Hence the choice  $\sigma_k = \sigma > 0$  for all  $k$  is allowed by the method, and thus  $\lambda_k = \sigma \|s_k\|^{2+\alpha}$  satisfies [\(2.3\)](#) and [\(2.4\)](#). [Theorem 3.2](#) then shows that this method may require at least  $\epsilon^{-(2+\alpha)/(1+\alpha)}$  iterations to generate an iterate with  $|g_k| \leq \epsilon$ . This is important as the *upper* bound on this number of iterations was proved<sup>(6)</sup> in [Cartis, Gould, and Toint \[2011b\]](#) to be

$$(4.2) \quad \mathcal{O}\left([f(x_0) - f_{\text{low}}]\right) \epsilon^{-\frac{2+\alpha}{1+\alpha}}$$

where  $f_{\text{low}}$  is any lower bound of  $f(x)$ . Since we have that  $f(x_0) - f_{\text{low}}$  is a fixed number independent of  $\epsilon$  for the example of [Section 3.1](#), this shows that the ratio

$$(4.3) \quad \rho_{\text{comp}} \stackrel{\text{def}}{=} \frac{\text{upper bound on the worst-case evaluation complexity}}{\text{lower bound on the worst-case evaluation complexity}}$$

for the  $(2 + \alpha)$ -regularization method is bounded independently of  $\epsilon$  and  $\alpha$ . Given that [\(4.2\)](#) involves an unspecified constant, this is the best that can be obtained as far as the

<sup>(6)</sup>As a matter of fact, [Cartis, Gould, and Toint \[2011b\]](#) contains a detailed proof of the result for  $\alpha = 1$ , as well as the statement that it generalizes for  $\alpha \in (0, 1]$ . Because of the central role of this result in the present paper, a more detailed proof of the worst-case evaluation complexity bound for  $\alpha \in (0, 1]$  is provided as [Appendix A2](#).

order in  $\epsilon$  is concerned, and yields the following important result on worst-case evaluation complexity.

**Theorem 4.1.** *When applied to a function satisfying  $A.\alpha$ , the  $(2 + \alpha)$ -regularization method may require at most (4.2) function and derivatives evaluations. Moreover this bound is sharp (in the sense that  $\rho_{\text{comp}}$  is bounded independently of  $\epsilon$  and  $\alpha$ ) and the  $(2 + \alpha)$ -regularization method is optimal in  $\mathfrak{M}.\alpha$ .*

*Proof.* The optimality of the  $(2 + \alpha)$ -regularization method within  $\mathfrak{M}.\alpha$  results from the observation that the example of Section 3 implies that no method in  $\mathfrak{M}.\alpha$  can have a worst-case evaluation complexity of a better order.  $\square$

In particular, the cubic regularization method is optimal for smooth optimization problems with Lipschitz continuous second derivatives. As we have seen above, this is in contrast with Newton's method.

Note that Theorem 4.1 as stated does *not* result from the statement in Cartis, Gould, and Toint [2011d] that the bound (4.2) is “essentially sharp”. Indeed this latter statement expresses the fact that, for any  $\tau > 0$ , there exists a function independent of  $\epsilon$ , on which the relevant method may need at least  $\epsilon^{-3/2+\tau}$  evaluations to terminate with  $|g_k| \leq \epsilon$ . But, for any fixed  $\epsilon$ , the value of  $f(x_0) - f_{\text{low}}$  tends to infinity when, in the example of that paper, the number of iterations to termination approaches  $\epsilon^{-3/2}$  as  $\tau$  goes to zero. As a consequence, the numerator of the ratio (4.3), that is (4.2), and  $\rho_{\text{comp}}$  itself are unbounded for that example. Theorem 4.1 thus brings a formal improvement on the conclusions of Cartis, Gould, and Toint [ibid.].

**4.3 Goldfeld-Quandt-Trotter.** Recalling (2.29), we can set  $\omega_k = \omega$  in the algorithm as every iteration is successful due to (4.1) which, with (3.3) and  $f_k \in [\frac{1}{2}, 1]$  gives that  $\lambda_k + \lambda_{\min}(H_k) \leq \omega |g_k|^{\frac{\alpha}{1+\alpha}}$ , which is in agreement with (2.5) and (2.4). Thus the lower bound of  $\epsilon^{-(2+\alpha)/(1+\alpha)}$  iterations for termination also applies to this method.

An upper bound on the worst-case evaluation complexity for the GQT method can be obtained by the following argument. We first note that, similarly to regularization methods, we can bound the total number of unsuccessful iterations as a constant multiple of the successful ones, provided  $\omega_k$  is chosen such that (2.32) holds. Moreover, since  $f$  satisfies  $A.\alpha$ , its Hessian is bounded above by (2.9). In addition, we have noted in Section 2.2 that  $\|g_k\|$  is also bounded above. In view of (2.29) and (2.39), this in turn implies that  $\|H_k + \lambda_k I\|$  is also bounded above. Hence we obtain from (2.33) that  $\|s_k\| \geq \kappa_{GQT} \|g_k\| \geq \kappa_{GQT} \epsilon$  for some  $\kappa_{GQT} > 0$ , as long as termination has not occurred. This last bound and (2.37) then give that GQT takes at most  $\mathcal{O}\left((f(x_0) - f_{\text{low}})\epsilon^{-\frac{\alpha}{1+\alpha}-2}\right)$  iterations, which is worse than (4.2) for  $\alpha > 0$ . Note that this bound improves if only



Newton steps are taken (i.e.  $\lambda_k = 0$  is chosen for all  $k \geq 0$ ), to be of the order of (4.2); however, this cannot be assumed in the worst-case for nonconvex functions. In any case, it implies that the GQT method is not optimal in  $\mathfrak{M}.\alpha$ .

**4.4 Trust-region methods.** Recall the choices (2.41) we make in this case. If  $\lambda_k = 0$ , the trust-region constraint  $\|s\| \leq \Delta_k$  is inactive at  $s_k$ , in which case,  $s_k$  is the Newton step. If we make precisely the choices we made for Newton's method above, choosing  $\Delta_0$  such that  $\Delta_0 > |s_0|$  implies that the Newton step will be taken in the first and in all subsequent iterations since each iteration is successful and then  $\Delta_k$  remains unchanged or increases while the choice (3.6) implies that  $s_k$  decreases. Thus the trust-region approach, through the Newton step, has a worst-case evaluation complexity when applied to  $f_\epsilon^{\mathfrak{M}.\alpha}$  which is at least that of the Newton's method, namely  $\epsilon^{-2}$ .

**4.5 Linesearch methods.** Because the examples of Sections 3.1 and 3.2 are valid for  $r_k = 0$  which corresponds to  $\mu_k = 1$  for all  $k$ , and because this stepsize is acceptable since  $f(x_{k+1}) = m_k(s_k)$ , we deduce that at least  $\epsilon^{-\frac{2+\alpha}{1+\alpha}}$  iterations and evaluations may be needed for the linesearch variants of any method in  $\mathfrak{M}.\alpha$  applied to a function satisfying A. $\alpha$ , and that  $\epsilon^{-2}$  evaluations may be needed for the linesearch variant of Newton's method applied on a function satisfying A.0. Thus the conclusions drawn regarding their (sub-)optimality in terms of worst-case evaluation complexity are not affected by the use of a linesearch.

## 5 The Curtis-Robinson-Samadi class

We finally consider a class of methods recently introduced in [Curtis, Robinson, and Samadi \[2017b\]](#), which we call the CRS class. This class depends on the parameters  $0 < \underline{\sigma} \leq \bar{\sigma}$ ,  $\eta \in (0, 1)$  and two non-negative accuracy thresholds  $\kappa_1$  and  $\kappa_2$ . It is defined as follows. At the start, adaptive regularization thresholds are set according to

$$(5.1) \quad \sigma_0^L = 0 \quad \text{and} \quad \sigma_0^U = \bar{\sigma}.$$

Then for each iteration  $k \geq 0$ , a step  $s_k$  from the current iterate  $x_k$  and a regularization parameter  $\lambda_k \geq 0$  are chosen to satisfy<sup>(7)</sup>

$$(5.2) \quad (H_k + \lambda_k I)s_k = -g_k + r_k,$$

---

<sup>(7)</sup>In [Curtis, Robinson, and Samadi \[2017b\]](#), further restrictions on the step are imposed in order to obtain global convergence under A.0 and bounded gradients, but are irrelevant for the worst-case complexity analysis under A.1. We thus ignore them here, but note that this analysis also ensures global convergence to first-order stationary points.

$$(5.3) \quad \sigma_k^L \|s_k\| \leq \lambda_k \leq \sigma_k^U \|s_k\|,$$

$$(5.4) \quad s_k^T r_k \leq \frac{1}{2} s_k^T (H_k + \lambda_k I) s_k + \frac{1}{2} \kappa_1 \|s_k\|^3,$$

and

$$(5.5) \quad \|r_k\| \leq \lambda_k \|s_k\| + \kappa_2 \|s_k\|^2.$$

The step is then accepted, setting  $x_{k+1} = x_k + s_k$ , if

$$(5.6) \quad \rho_{CRS} = \frac{f(x_k) - f(x_k + s_k)}{\|s_k\|^3} \geq \eta$$

or rejected otherwise. In the first case, the regularization thresholds are reset according to (5.1). If  $s_k$  is rejected,  $\sigma_k^L$  and  $\sigma_k^U$  are updated by a simple mechanism (using  $\underline{\sigma}$ ) which is irrelevant for our purpose here. The algorithm is terminated as soon as an iterate is found such that  $\|g_k\| \leq \epsilon$ .

Observe that (5.2) corresponds to inexactly minimizing the regularized model (2.6) and that (5.5) is very similar to the subproblem termination rule of [Birgin, Gardenghi, Martínez, Santos, and Toint \[2017\]](#).

An upper bound of  $\mathcal{O}(\epsilon^{-3/2})$  is proved in [Curtis, Robinson, and Samadi \[2017b, Theorem 17\]](#) for the worst-case evaluation complexity of the methods belonging to the CRS class. It is stated in [Curtis, Robinson, and Samadi \[ibid.\]](#) that both ARC [Griewank \[1981\]](#), [Weiser, Deuffhard, and Erdmann \[2007\]](#), [Nesterov and Polyak \[2006\]](#), and [Cartis, Gould, and Toint \[2011a,b\]](#) and TRACE [Curtis, Robinson, and Samadi \[2017a\]](#) belong to the class, although the details are not given.

Clearly, the CRS class is close to  $\mathfrak{M}.1$ , but yet differs from it. In particular, no requirement is made that  $H_k + \lambda_k I$  be positive semi-definite but (5.4) is required instead, there is no formal need for the step to be bounded and (5.5) combined with (5.3) is slightly more permissive than the second part of (2.2). We now define  $\text{CRS}_a$ , a sub-class of the CRS class of methods, as the set of CRS methods for which (5.5) is strengthened<sup>(8)</sup> to become

$$(5.7) \quad \|r_k\| \leq \min \left[ \kappa_{rg} \|g_k\|, \lambda_k \|s_k\| + \kappa_2 \|s_k\|^2 \right] \quad \text{with} \quad \kappa_{rg} < 1.$$

(in a manner reminiscent of the second part of (2.2)) and such that

$$(5.8) \quad 2\eta(1 + \kappa_{rg})^3 \leq 1$$

(a mild technical condition<sup>(9)</sup> whose need will become apparent below). We claim that, for any choice of method in the  $\text{CRS}_a$  class and termination threshold  $\epsilon$ , we can construct

<sup>(8)</sup>Hence the subscript  $a$ , for “accurate”.

<sup>(9)</sup>Due to the lack of scaling invariance of (5.6), at variance with (2.30).

a function satisfying A.1 such that the considered  $\text{CRS}_a$  method terminates in exactly  $\lceil \epsilon^{-3/2} \rceil$  iterations and evaluations. This achieved simply by showing that the generated sequences of iterates, function, gradient and Hessian values belong to those detailed in the example of [Section 3.1](#).

We now apply a method of the  $\text{CRS}_a$  class for a given  $\epsilon > 0$ , and first consider an iterate  $x_k$  with associated values  $f_k$ ,  $g_k$  and  $H_k$  given by (3.3) for  $\alpha = 1$ , that is

$$(5.9) \quad f_0 = 1, \quad f_k = f_0 - \frac{1}{2}k\epsilon^{3/2}, \quad g_k = -2\epsilon f_k \quad \text{and} \quad H_k = 4\epsilon^{1/2}f_k^2;$$

Suppose that

$$(5.10) \quad \sigma_k^L = 0 \quad \text{and} \quad \sigma_k^U = \bar{\sigma}$$

(as is the case by definition for  $k = 0$ ), and let

$$(5.11) \quad s_k = \theta_k \frac{\epsilon^{1/2}}{2f_k} \quad (\theta_k > 0)$$

be an acceptable step for an arbitrary method in the  $\text{CRS}_a$  class. Now, because of (5.10), (5.3) reduces to

$$(5.12) \quad \lambda_k \in [0, \bar{\sigma}|s_k|] = \left[ 0, \bar{\sigma}\theta_k \frac{\epsilon^{1/2}}{2f_k} \right]$$

and, given that  $H_k > 0$  because of (5.9), this in turn implies that  $H_k + \lambda_k > 0$ . Condition (5.7) requires that

$$(5.13) \quad |g_k + (H_k + \lambda_k)s_k| = |r_k| \leq \kappa_{rg}|g_k| = 2\kappa_{rg}\epsilon f_k < 2\epsilon,$$

where we used the fact that  $f_k \leq 1$  because of (5.9) and  $\kappa_{rg} < 1$  because of (5.7). Moreover, (5.13) and (5.12) imply that

$$(5.14) \quad \frac{2(1 - \kappa_{rg})\epsilon f_k}{4\epsilon^{1/2}f_k^2 + \bar{\sigma}s_k} \leq \frac{|g_k|(1 - \kappa_{rg})}{H_k + \lambda_k} \leq s_k \leq \frac{|g_k|(1 + \kappa_{rg})}{H_k + \lambda_k} \leq \frac{(1 + \kappa_{rg})\epsilon^{1/2}}{2f_k}.$$

Thus, using (5.11) and the right-most part of these inequalities, we obtain that  $\theta_k \leq 1 + \kappa_{rg}$ , which in turn ensures that  $s_k \leq (1 + \kappa_{rg})\epsilon^{1/2}/(2f_k)$ . Substituting this latter bound in the denominator of the left-most part of (5.14) and using (5.11) again with the fact that  $f_k \geq \frac{1}{2}$  before termination, we obtain that

$$(5.15) \quad \theta_k \in \left[ \frac{1 - \kappa_{rg}}{1 + \bar{\sigma}(1 + \kappa_{rg})}, 1 + \kappa_{rg} \right]$$

(note that this is (3.6) with  $\kappa_\lambda = 1 + \bar{\sigma}(1 + \kappa_{rg})$ ). We immediately note that  $\pi_k$  and  $\phi(\theta_k)$  are then both guaranteed to be bounded above and below as in (3.14). (Since this is enough for our purpose, we ignore the additional restriction on  $\theta_k$  which might result from (5.4).) Using the definitions (5.9) for  $k + 1$ , we may then construct the objective function  $f_\epsilon^{CRS}$  on the interval  $[x_k, x_k + s_k]$  by Hermite interpolation, as in Section 3.1. Moreover, using (5.6), (5.9), (5.11), (5.15),  $f_k \in [\frac{1}{2}, 1]$  and the condition (5.8), we obtain that

$$\rho_k = \frac{\epsilon^{3/2}}{2} \left( \frac{2f_k}{\theta_k \epsilon^{1/2}} \right)^3 = \frac{4f_k^3}{\theta_k^3} \geq \frac{1}{2(1 + \kappa_{rg})^3} \geq \eta.$$

Thus iteration  $k$  is successful,  $x_{k+1} = x_k + s_k$ ,  $\sigma_{k+1}^L = \sigma_k^L = 0$ ,  $\sigma_{k+1}^U = \sigma_k^U = \bar{\sigma}$ , and all subsequent iterations of the  $CRS_a$  method up to termination follow the same pattern in accordance with (5.9). As in Section 3.1, we may construct  $f_\epsilon^{CRS}$  on the whole of  $\mathbb{R}$  which satisfies A.1 and such that, the considered  $CRS_a$  method applied to  $f_\epsilon^{CRS}$  will terminate in exactly  $\lceil \epsilon^{-3/2} \rceil$  iterations and evaluations. This and the  $\mathcal{O}(\epsilon^{-3/2})$  upper bound on the worst-case evaluation complexity of CRS methods allow stating the following theorem.

**Theorem 5.1.** *For every  $\epsilon \in (0, 1)$  and every method in the  $CRS_a$  class, a function  $f_\epsilon^{CRS}$  satisfying A.1 with values in a bounded interval independent of  $\epsilon$  can be constructed, such that the considered method terminates exactly at iteration*

$$k_\epsilon = \left\lceil \epsilon^{-3/2} \right\rceil$$

*with the first iterate  $x_{k_\epsilon}$  such that  $\|\nabla_x f_\epsilon^{CRS}(x_{k_\epsilon})\| \leq \epsilon$ . As a consequence, methods in  $CRS_a$  are optimal within the CRS class and their worst-case evaluation complexity is, in order, also optimal with respect to that of methods in  $\mathfrak{M}.1$ .*

$CRS_a$  then constitutes a kernel of optimal methods (from the worst-case evaluation complexity point of view) within CRS and  $\mathfrak{M}.1$ . Methods in CRS but not in  $CRS_a$  correspond to very inaccurate minimization of the regularized model, which makes it unlikely that their worst-case evaluation complexity surpasses that of methods in  $CRS_a$ . Finally note that, since we did not use (5.4) to construct our example, it effectively applies to a class larger than  $CRS_a$  where this condition is not imposed.

## 6 The algorithm of Royer and Wright

We finally consider the linesearch algorithm proposed in Royer and Wright [2017, Algorithm 1], which is reminiscent of the double linesearch algorithm of Gould, Lucidi, Roma, and Toint [1998] and Conn, Gould, and Toint [2000, Section 10.3.1]. From a given iterate  $x_k$ , this algorithm computes a search direction  $d_k$  whose nature depends on the curvature

of the (unregularized) quadratic model along the negative gradient, and possibly computes the left-most eigenpair of the Hessian if this curvature is negative or if the gradient's norm is small enough to declare first-order stationarity. A linesearch along  $d_k$  is then performed by reducing the steplength  $\alpha_k$  from  $\alpha_k = 1$  until

$$(6.1) \quad f(x_k + \alpha_k d_k) \leq f(x_k) - \frac{\eta}{6} \alpha_k^3 \|d_k\|^3$$

for some  $\eta > 0$ . The algorithm uses  $\epsilon_g$  and  $\epsilon_H$ , two different accuracy thresholds for first- and second-order approximate criticality, respectively.

Our objective is now to show that, when applied to the function  $f_{\epsilon_g}^{\mathfrak{m}.1}$  of Section 3.1 with  $\epsilon = \epsilon_g$ , this algorithm, which we call the RW algorithm, takes exactly  $k_{\epsilon_g,1} = \lceil \epsilon_g^{-3/2} \rceil$  iterations and evaluations to terminate with  $\|g_k\| \leq \epsilon_g$ .

We first note that (3.3) guarantees that  $H_k$  is positive definite and, using (3.4), that

$$\frac{g_k^T H_k g_k}{\|g_k\|^2} = 4\epsilon_g^{1/2} f_k^2 > \epsilon_g$$

for  $k \in \{0, \dots, k_{\epsilon_g,1}\}$ . Then, provided

$$(6.2) \quad \epsilon_H \leq \sqrt{\epsilon_g},$$

and because  $\lambda_{\min}(H_k) = 4\epsilon_g^{1/2} f_k^2 > \epsilon_H$  (using (3.4) again), the RW algorithm defines the search direction from Newton's equation  $H_k d_k = -g_k$  (which corresponds, as we have already seen, to taking  $M_k = 0 = r_k$  and thus  $\theta_k = 1$  in the example of Section 3.1). The RW algorithm is therefore, on that example, identical to a linesearch variant of Newton's method with the specific linesearch condition (6.1). Moreover, using (3.4) once more,

$$f(x_k) - f(x_k + d_k) = \frac{1}{2} \epsilon_g^{3/2} \geq \frac{\eta}{6} \left( \frac{\epsilon_g^{1/2}}{2f_k} \right)^3 \geq \frac{\eta}{6} \epsilon_g^{3/2}$$

whenever  $\eta \leq 3$ , an extremely weak condition<sup>(10)</sup>. Thus (6.1) holds<sup>(11)</sup> with  $\alpha_k = 1$ . We have thus proved that the RW algorithm generates the same sequence of iterates as Newton's method when applied to  $f_{\epsilon_g}^{\mathfrak{m}.1}$ . The fact that an upper bound of  $\mathcal{O}(\epsilon_g^{-3/2})$  iterations and evaluations was proved to hold in Royer and Wright [2017, Theorem 5] then leads us to stating the following result.

**Theorem 6.1.** *Assume that  $\eta \in (0, 3]$ . Then, for every  $\epsilon_g \in (0, 1)$  and  $\epsilon_H$  satisfying (6.2), a function  $f_{\epsilon_g}^{\mathfrak{m}.1}$  satisfying A.1 with values in a bounded interval (independent of  $\epsilon_g$*

<sup>(10)</sup>In practice,  $\eta$  is most likely to belong to  $(0, 1)$  and even be reasonably close to zero.

<sup>(11)</sup>But fails for the example of Section 3.2 as  $\|s_k\| = 1$ .

and  $\epsilon_H$ ) can be constructed, such that the Royer-Wright algorithm terminates exactly at iteration

$$k_{\epsilon_g} = \left\lceil \epsilon_g^{-3/2} \right\rceil$$

with the first iterate  $x_{k_{\epsilon_g}}$  such that  $\|\nabla_x f_{\epsilon_g}^{\mathfrak{M}.1}(x_{k_{\epsilon_g}})\| \leq \epsilon_g$ . As a consequence and under assumption (6.2), the first-order worst-case evaluation complexity order of  $\mathcal{O}(\epsilon_g^{-3/2})$  for this algorithm is sharp and it is (in order of  $\epsilon_g$ ), also optimal with respect to that of algorithms in the  $\mathfrak{M}.1$  and CRS classes.

## 7 Conclusions

We have provided lower bounds on the worst-case evaluation complexity of a wide class of second-order methods for reaching approximate first-order critical points of nonconvex, adequately smooth unconstrained optimization problems. This has been achieved by providing improved examples of slow convergence on functions with bounded range independent of  $\epsilon$ . We have found that regularization algorithms, methods belonging to a subclass of that proposed in [Curtis, Robinson, and Samadi \[2017b\]](#) and the linesearch algorithm of [Royer and Wright \[2017\]](#) are optimal from a worst-case complexity point of view within a very wide class of second-order methods, in that their upper complexity bounds match in order the lower bound we have shown for relevant, sufficiently smooth objectives satisfying  $A.\alpha$ . At this point, the question of whether all known optimal second-order methods share enough design concepts to be made members of a single class remains open.

Note that every iteration complexity bound discussed above is of the order  $\epsilon^{-p}$  (for various values of  $p > 0$ ) for driving the objective's gradient below  $\epsilon$ ; thus the methods we have addressed may require an exponential number of iterations  $10^{p \cdot k}$  to generate  $k$  correct digits in the solution. Also, as our examples are one-dimensional, they fail to capture the problem-dimension dependence of the upper complexity bounds. Indeed, besides the accuracy tolerance  $\epsilon$ , existing upper bounds depend on the distance to the solution set, that is  $f(x_0) - f_{\text{low}}$ , and the gradient's and Hessian's Lipschitz or Hölder constants, all of which may dependent on the problem dimension. Some recent developments in this respect can be found in [Jarre \[2013\]](#), [Agarwal, Allen-Zhu, Bullins, Hazan, and T. Ma \[2016\]](#), [B. Jiang, Lin, S. Ma, and S. Zhang \[2016\]](#), and [Royer and Wright \[2017\]](#).

Here we have solely addressed the evaluation complexity of generating first-order critical points, but it is common to require second-order methods for nonconvex problems to achieve second-order criticality. Indeed, upper worst-case complexity bounds are known in this case for cubic regularization and trust-region methods [Nesterov and Polyak \[2006\]](#)

and Cartis, Gould, and Toint [2011b, 2012b], which are essentially sharp in some cases Cartis, Gould, and Toint [2012b]. A lower bound on the whole class of second order methods for achieving second-order optimality remains to be established, especially when different accuracy is requested in the first- and second-order criticality conditions.

Regarding the worst-case evaluation complexity of constrained optimization problems, we have shown Cartis, Gould, and Toint [2012a, 2011c, 2014] that the presence of constraints does not change the order of the bound, so that the unconstrained upper bound for some first- or second-order methods carries over to the constrained case; note that this does not include the cost of solving the constrained subproblems as the latter does not require additional problem evaluations. Since constrained problems are at least as difficult as unconstrained ones, these bounds are also sharp. It remains an open question whether a unified treatment such as the one given here can be provided for the worst-case evaluation complexity of methods for constrained problems.

## A1. An example of slow convergence of the steepest-descent method

We show in this paragraph that the steepest-descent method may need at least  $\epsilon^{-2}$  iteration to terminate on a function whose range is fixed and independent of  $\epsilon$ .

We once again follow the methodology used in Section 3.1 and build a unidimensional function  $f_\epsilon^{SD}$  by Hermite interpolation, such that the steepest-descent method applied to this function takes exactly  $k_\epsilon = \lceil \epsilon^{-2} \rceil$  iterations and function evaluations to terminate with an iterate  $x_k$  such that  $|g(x_k)| \leq \epsilon$ . Note that, for the sequence of function values to be interpretable as the result of applying the steepest-descent method (using a Goldstein linesearch), we require that, for all  $k$ ,

$$(A.1) \quad f(x_k) + \mu_1 g_k^T s_k \leq f(x_k - \mu_k g_k) \leq f(x_k) + \mu_2 g_k^T s_k \quad \text{for constants } 0 < \mu_2 < \mu_1 < 1$$

where, as above,  $s_k = x_{k+1} - x_k$ . Keeping this in mind, we define the sequences  $f_k, g_k, H_k$  and  $s_k$  for  $k \in \{0, \dots, k_\epsilon - 1\}$  by

$$f_k = 1 - \frac{1}{2} k \epsilon^2 \quad g_k = -2\epsilon f_k, \quad H_k = 0, \quad r_k = 0 \quad \text{and} \quad \mu_k = \frac{1}{4 f_k^2} \in [\tfrac{1}{4}, 1].$$

Note that this last definition ensures that (A.1) holds provided  $0 < \mu_2 < \frac{1}{2} < \mu_1 < 1$ . It also gives that  $s_k = \epsilon / (2 f_k) \leq \epsilon < 1$ . Using these values, it can also be verified that termination occurs for  $k = k_\epsilon$ , that  $f_\epsilon^{SD}$  defined by (3.10) and Hermite interpolation is twice continuously differentiable on  $[0, x_{k_\epsilon}]$  and that (3.12) again holds. Since  $|g_k| \leq \epsilon$ , we also obtain that, for  $k \in \{0, \dots, k_\epsilon - 1\}$ ,

$$\left| \frac{\Delta f_k}{s_k^2} \right| = 2 f_k^2 \leq 1, \quad \left| \frac{\Delta g_k}{s_k} \right| = 2\epsilon^2 f_k \leq 2 \quad \text{and} \quad \left| \frac{g_k}{s_k} \right| = 4 f_k^2 \leq 4.$$

These bounds,  $H_k = \Delta H_k = 0$ , the first equality of (3.18) and (3.13) then imply that the Hessian of  $f_\epsilon^{SD}$  is bounded above by a constant independent of  $\epsilon$ .  $f_\epsilon^{SD}$  thus satisfies A.0 and therefore has Lipschitz continuous gradient. Moreover, since  $s_k \leq 1$ , we also obtain, as in Section 3.1 and 3.2, that  $|f_\epsilon^{SD}|$  is bounded by a constant independent of  $\epsilon$  on  $[0, x_{k_\epsilon}]$ . As above we then extend  $f_\epsilon^{SD}$  to the whole of  $\mathbb{R}$  while preserving A.0.

**Theorem A.1.** *For every  $\epsilon \in (0, 1)$ , a function  $f_\epsilon^{SD}$  satisfying A.0 (and thus having Lipschitz continuous gradient) with values in a bounded interval independent of  $\epsilon$  can be constructed, such that the steepest-descent method terminates exactly at iteration*

$$k_\epsilon = \lceil \epsilon^{-2} \rceil$$

with the first iterate  $x_{k_\epsilon}$  such that  $\|\nabla_x f_\epsilon^{SD}(x_{k_\epsilon})\| \leq \epsilon$ .

As a consequence, the  $\mathcal{O}(\epsilon^{-2})$  order of worst-case evaluation complexity is sharp for the steepest-descent method in the sense that the complexity ratio  $\rho_{\text{comp}}$  is bounded above independently of  $\epsilon$ , which improves on the conclusion proposed in Cartis, Gould, and Toint [2010] for the steepest-descent method.

The top three graphs of Figure A.2 illustrate the global behaviour of the resulting function  $f_\epsilon^N(x)$  and of its first and second derivatives for  $x \in [0, x_{k_\epsilon}]$ , while the bottom ones show more detail of the first 10 iterations. The figure is once more constructed using  $\epsilon = 5.10^{-2}$  ( $k_\epsilon = 400$ ).

## A2. Upper complexity bound for the $(2 + \alpha)$ -regularization method

The purpose of this paragraph is to provide some of the missing details in the proof of Lemma 2.5, as well as making explicit the statement made at the end of Section 5.1 in Cartis, Gould, and Toint [2011b] that the  $(2 + \alpha)$ -regularization method needs at most (4.2) iterations (and function/derivatives evaluations) to obtain and iterate  $x_k$  such that  $|g_k| \leq \epsilon$ .

We start by proving (2.27) following the reasoning of Cartis, Gould, and Toint [2011a, Lem. 2.2]. Consider

$$\begin{aligned} m_k(s) - f(x_k) &= g_k^T s + \frac{1}{2} s^T H_k s + \frac{1}{2 + \alpha} \sigma_k \|s\|^{2+\alpha} \\ &\geq -\|g_k\| \|s\| - \frac{1}{2} \|s\|^2 \|H_k\| + \frac{1}{2 + \alpha} \sigma_k \|s\|^{2+\alpha} \\ &\geq \left( \frac{1}{3(2 + \alpha)} \sigma_k \|s\|^{2+\alpha} - \|g_k\| \|s\| \right) \\ &\quad + \left( \frac{2}{3(2 + \alpha)} \sigma_k \|s\|^{2+\alpha} - \frac{1}{2} \|s\|^2 \|H_k\| \right) \end{aligned}$$



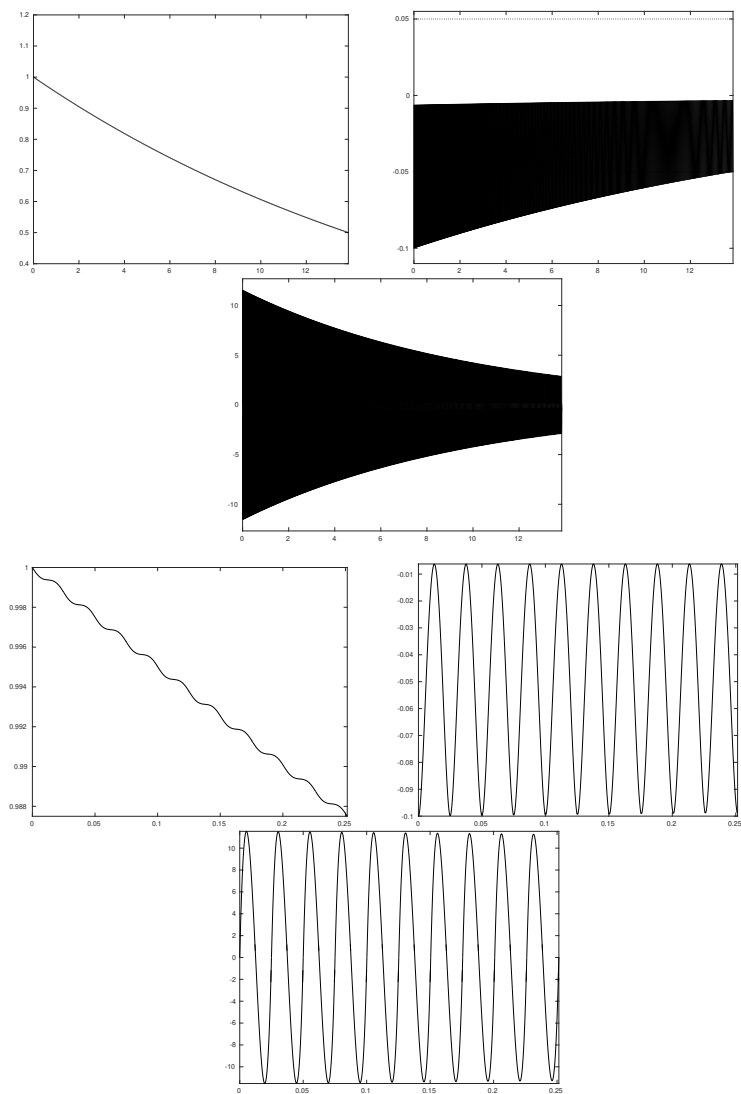


Figure A.2:  $f_{\epsilon}^{SD}(x)$  (left) and its first (center) and second (right) derivatives as a function of  $x$  for  $\epsilon = 5 \cdot 10^{-2}$  (top:  $x \in [0, x_{k_{\epsilon, \alpha}}]$ ; bottom:  $x \in [0, x_{10}]$ ). Horizontal dotted lines indicate values of  $-\epsilon$  and  $\epsilon$  in the central top graph.

But then  $\frac{2}{3(2+\alpha)}\sigma_k \|s\|^{2+\alpha} - \|H_k\| \|s\|^2 > 0$  if  $\|s_k\| < (3(2+\alpha)\|H_k\|/(4\sigma_k))^{\frac{1}{\alpha}}$  while  $\frac{1}{3(2+\alpha)}\sigma_k \|s\|^{2+\alpha} - \|g_k\| \|s\| > 0$  if  $\|s_k\| < (3(2+\alpha)\|g_k\|/\sigma_k)^{\frac{1}{1+\alpha}}$ . Hence, since  $m_k(s_k) < f(x_k)$ , we have that

$$\|s_k\| \leq \max \left[ \left( \frac{3(2+\alpha)\|H_k\|}{4\sigma_k} \right)^{\frac{1}{\alpha}}, \left( \frac{3(2+\alpha)\|g_k\|}{\sigma_k} \right)^{\frac{1}{1+\alpha}} \right]$$

which yields (2.27) because  $\|H_k\| \leq L_g$ .

We next explicit the worst-case evaluation complexity bound of Section 5.1 in [Cartis, Gould, and Toint \[2011b\]](#). Following [Cartis, Gould, and Toint \[2011a, Lemma 5.2\]](#), we start by proving that

$$(A.1) \quad \sigma_{\max} \stackrel{\text{def}}{=} c_{\sigma} \max(\sigma_0, L_{H,\alpha})$$

for some constant  $c_{\sigma}$  only dependent on  $\alpha$  and algorithm's parameters. To show this inequality, we deduce from Taylor's theorem that, for each  $k \geq 0$  and some  $\xi_k$  belonging the segment  $[x_k, x_k + s_k]$ ,

$$\begin{aligned} f(x_k + s_k) - m_k(s_k) &\leq \frac{1}{2} \|H(\xi_k) - H(x_k)\| \cdot \|s_k\|^2 - \frac{\sigma_k}{2+\alpha} \|s_k\|^{2+\alpha} \\ &\leq \left( \frac{L_{H,\alpha}}{2} - \frac{\sigma_k}{2+\alpha} \right) \|s_k\|^{2+\alpha}, \end{aligned}$$

where, to obtain the second inequality, we employed (2.8) in A.α and  $\|\xi_k - x_k\| \leq \|s_k\|$ . Thus  $f(x_k + s_k) < m_k(s_k)$  whenever  $\sigma_k > \frac{1}{2}(2+\alpha)L_{H,\alpha}$ , providing sufficient descent and ensuring that  $\sigma_{k+1} \leq \sigma_k$ . Taking into account the (possibly large) choice of the regularization parameter at startup then yields (A.1).

We next note that, because of (2.25) and (A.1), (2.11) holds. Moreover,  $\kappa(M_k) = \kappa(\sigma_k \|s_k\|^{\alpha} I) = 1$ . [Lemma 2.3](#) then ensures that (2.16) also holds.

We finally follow [Cartis, Gould, and Toint \[ibid., Corollary 5.3\]](#) to prove the final upper bound on the number of successful iterations (and hence on the number of function and derivatives evaluations). Let  $\mathcal{S}_k^{\epsilon}$  index the subset of the first  $k$  iterations that are successful and such that  $\min[\|g_k\|, \|g_{k+1}\|] > \epsilon$ , and let  $|\mathcal{S}_k^{\epsilon}|$  denote its cardinality. It follows from this definition, (2.11), (2.26) and the fact that sufficient decrease is obtained at successful iterations that, for all  $k$  before termination,

$$(A.2) \quad f(x_j) - m_k(s_j) \geq \alpha_S \epsilon^{\frac{2+\alpha}{1+\alpha}}, \quad \text{for all } j \in \mathcal{S}_k^{\epsilon},$$

for some positive constant  $\alpha_S$  independent of  $\epsilon$ . Now, if  $f_{\text{low}} > -\infty$  is a lower bound on  $f(x)$ , we have, using the monotonically decreasing nature of  $\{f(x_k)\}$ , that

$$\begin{aligned} f(x_0) - f_{\text{low}} &\geq f(x_0) - f(x_{k+1}) = \sum_{j \in \mathcal{S}_k^\epsilon} [f(x_j) - f(x_{j+1})] \\ &\geq \eta_1 \sum_{j \in \mathcal{S}_k^\epsilon} [f(x_j) - m_k(s_j)] \geq |\mathcal{S}_k^\epsilon| \eta_1 \alpha_S \epsilon^{\frac{2+\alpha}{1+\alpha}}, \end{aligned}$$

where the constant  $\eta_1 \in (0, 1)$  defines sufficient decrease. Hence, for all  $k \geq 0$ ,

$$|\mathcal{S}_k^\epsilon| \leq \frac{f(x_0) - f_{\text{low}}}{\eta_1 \alpha_S} \epsilon^{-\frac{2+\alpha}{1+\alpha}}.$$

As a consequence, the  $(2 + \alpha)$ -regularization method needs at most (4.2) successful iterations to terminate. Since it is known that, for regularization methods,  $k \leq \kappa_S |\mathcal{S}_k^\epsilon|$  for some constant  $\kappa_S$  [Cartis, Gould, and Toint 2011b, Theorem 2.1] and because every iteration involves a single evaluation, we conclude that the  $(2 + \alpha)$ -regularization method needs at most (4.2) function and derivatives evaluations to produce an iterate  $x_k$  such that  $\|g_k\| \leq \epsilon$  when applied to an objective function satisfying A.α.

We finally observe that the statement (made in the proof of Lemma 2.5) that  $\|g_k\|$  is bounded above immediately follows from this worst-case evaluation complexity bound.

## References

- Naman Agarwal, Zeyuan Allen-Zhu, Brian Bullins, Elad Hazan, and Tengyu Ma (Nov. 2016). “Finding Approximate Local Minima Faster than Gradient Descent”. arXiv: 1611.01146 (cit. on p. 3761).
- Anima Anandkumar and Rong Ge (Feb. 2016). “Efficient approaches for escaping higher order saddle points in non-convex optimization”. arXiv: 1602.05908.
- E. Bergou, Y. Diouane, and S. Gratton (2017). “On the use of the energy norm in trust-region and adaptive cubic regularization subproblems”. *Comput. Optim. Appl.* 68.3, pp. 533–554. MR: 3722090.
- Wei Bian and Xiaojun Chen (2013). “Worst-case complexity of smoothing quadratic regularization methods for non-Lipschitzian optimization”. *SIAM J. Optim.* 23.3, pp. 1718–1741. MR: 3093871.
- (2015). “Linearly constrained non-Lipschitz optimization for image restoration”. *SIAM J. Imaging Sci.* 8.4, pp. 2294–2322. MR: 3413588.
- Wei Bian, Xiaojun Chen, and Yinyu Ye (2015). “Complexity analysis of interior point algorithms for non-Lipschitz and nonconvex minimization”. *Math. Program.* 149.1-2, Ser. A, pp. 301–327. MR: 3300465.

- Tommaso Bianconcini, Giampaolo Liuzzi, Benedetta Morini, and Marco Sciadrone (2015). “On the use of iterative methods in cubic regularization for unconstrained optimization”. *Comput. Optim. Appl.* 60.1, pp. 35–57. MR: [3297888](#).
- Tommaso Bianconcini and Marco Sciadrone (2016). “A cubic regularization algorithm for unconstrained optimization using line search and nonmonotone techniques”. *Optim. Methods Softw.* 31.5, pp. 1008–1035. MR: [3534567](#).
- E. G. Birgin, J. L. Gardenghi, J. M. Martinez, S. A. Santos, and Philippe L. Toint (2016). “Evaluation complexity for nonlinear constrained optimization using unscaled KKT conditions and high-order models”. *SIAM J. Optim.* 26.2, pp. 951–967. MR: [3484405](#).
- (2017). “Worst-case evaluation complexity for unconstrained nonlinear optimization using high-order regularized models”. *Math. Program.* 163.1-2, Ser. A, pp. 359–368. MR: [3632983](#) (cit. on p. [3757](#)).
- E. G. Birgin and J. M. Martinez (2017). *On regularization and active-set methods with complexity for constrained optimization*.
- Nicolas Boumal, P. -A. Absil, and Coralia Cartis (May 2016). “Global rates of convergence for nonconvex optimization on manifolds”. arXiv: [1605.08101](#).
- Yair Carmon and John C. Duchi (Dec. 2016). “Gradient Descent Efficiently Finds the Cubic-Regularized Non-Convex Newton Step”. arXiv: [1612.00547](#).
- Yair Carmon, Oliver Hinder, John C. Duchi, and Aaron Sidford (May 2017). ““Convex Until Proven Guilty”: Dimension-Free Acceleration of Gradient Descent on Non-Convex Functions”. arXiv: [1705.02766](#).
- Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint (2010). “On the complexity of steepest descent, Newton’s and regularized Newton’s methods for nonconvex unconstrained optimization problems”. *SIAM J. Optim.* 20.6, pp. 2833–2852. MR: [2721157](#) (cit. on pp. [3730](#), [3731](#), [3733](#), [3745](#), [3751](#), [3754](#), [3763](#)).
- (2011a). “Adaptive cubic regularisation methods for unconstrained optimization. Part I: motivation, convergence and numerical results”. *Math. Program.* 127.2, Ser. A, pp. 245–295. MR: [2776701](#) (cit. on pp. [3730](#), [3731](#), [3738–3740](#), [3757](#), [3763](#), [3765](#)).
- (2011b). “Adaptive cubic regularisation methods for unconstrained optimization. Part II: worst-case function- and derivative-evaluation complexity”. *Math. Program.* 130.2, Ser. A, pp. 295–319. MR: [2855872](#) (cit. on pp. [3733](#), [3738–3740](#), [3754](#), [3757](#), [3761](#), [3763](#), [3765](#), [3766](#)).
- (2011c). “On the evaluation complexity of composite function minimization with applications to nonconvex nonlinear programming”. *SIAM J. Optim.* 21.4, pp. 1721–1739. MR: [2869514](#) (cit. on p. [3762](#)).
- (2011d). *Optimal Newton-type methods for nonconvex optimization*. Tech. rep. Technical Report naXys-17-2011, Namur Centre for Complex Systems (naXys), FUNDP-University of Namur, Namur, Belgium (cit. on pp. [3731](#), [3733](#), [3745](#), [3753–3755](#)).

- Coralia Cartis, Nicholas I. M. Gould, and Philippe L. Toint (2012a). “An adaptive cubic regularization algorithm for nonconvex optimization with convex constraints and its function-evaluation complexity”. *IMA J. Numer. Anal.* 32.4, pp. 1662–1695. MR: [2991841](#) (cit. on p. 3762).
- (2012b). “Complexity bounds for second-order optimality in unconstrained optimization”. *J. Complexity* 28.1, pp. 93–108. MR: [2871787](#) (cit. on pp. 3761, 3762).
  - (2012c). *On the complexity of the steepest-descent with exact linesearches*. Tech. rep. Technical Report naXys-16-2012, Namur Center for Complex Systems (naXys), University of Namur, Namur, Belgium (cit. on p. 3730).
  - (2013). “On the evaluation complexity of cubic regularization methods for potentially rank-deficient nonlinear least-squares problems and its relevance to constrained nonlinear optimization”. *SIAM J. Optim.* 23.3, pp. 1553–1574. MR: [3084179](#).
  - (2014). “On the complexity of finding first-order critical points in constrained nonlinear optimization”. *Math. Program.* 144.1-2, Ser. A, pp. 93–106. MR: [3179956](#) (cit. on p. 3762).
  - (2015a). *Improved worst-case evaluation complexity for potentially rank-deficient nonlinear least-Euclidean-norm problems using higher-order regularized models*. Tech. rep. Technical Report naXys-12-2015, Namur Center for Complex Systems (naXys), University of Namur, Namur, Belgium.
  - (2015b). “On the evaluation complexity of constrained nonlinear least-squares and general constrained nonlinear optimization using second-order methods”. *SIAM J. Numer. Anal.* 53.2, pp. 836–851. MR: [3325759](#).
  - (Aug. 2017a). “Improved second-order evaluation complexity for unconstrained nonlinear optimization using high-order regularized models”. arXiv: [1708.04044](#).
  - (2017b). “Second-Order Optimality and Beyond: Characterization and Evaluation Complexity in Convexly Constrained Nonlinear Optimization”. *Foundations of Computational Mathematics*, pp. 1–35.
  - (2017c). “Universal regularization methods—varying the power, the smoothness and the accuracy”. To appear in *Optimization Methods and Software*.
- Coralia Cartis, Ph. R. Sampaio, and Philippe L. Toint (2015). “Worst-case evaluation complexity of non-monotone gradient-related algorithms for unconstrained optimization”. *Optimization* 64.5, pp. 1349–1361. MR: [3316806](#).
- Coralia Cartis and Katya Scheinberg (2017). “Global convergence rate analysis of unconstrained optimization methods based on probabilistic models”. *Mathematical Programming, Series A*, pp. 1–39.
- Xiaojun Chen, Philippe L. Toint, and Hong Wang (Apr. 2017). “Partially separable convexly-constrained optimization with non-Lipschitzian singularities and its complexity”. arXiv: [1704.06919](#).

- Andrew R. Conn, Nicholas I. M. Gould, and Philippe L. Toint (2000). *Trust-region methods*. MPS/SIAM Series on Optimization. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA, pp. xx+959. MR: [1774899](#) (cit. on pp. [3730](#), [3741–3744](#), [3759](#)).
- Frank E. Curtis, Daniel P. Robinson, and Mohammadreza Samadi (2017a). “A trust region algorithm with a worst-case iteration complexity of  $\mathcal{O}(\epsilon^{-3/2})$  for nonconvex optimization”. *Math. Program.* 162.1-2, Ser. A, pp. 1–32. MR: [3612930](#) (cit. on p. [3757](#)).
- (Aug. 2017b). “An Inexact Regularized Newton Framework with a Worst-Case Iteration Complexity of  $\mathcal{O}(\epsilon^{-3/2})$  for Nonconvex Optimization”. arXiv: [1708.00475](#) (cit. on pp. [3730](#), [3732](#), [3756](#), [3757](#), [3761](#)).
- (2017c). *Complexity analysis of a trust funnel algorithm for equality constrained optimization*. Tech. rep. Technical Report 16T-03, ISE/CORL, LeHigh University, Bethlehem, PA, USA.
- John E. Dennis Jr. and Robert B. Schnabel (1983). *Numerical methods for unconstrained optimization and nonlinear equations*. Prentice Hall Series in Computational Mathematics. Prentice Hall, Inc., Englewood Cliffs, NJ, pp. xiii+378. MR: [702023](#) (cit. on pp. [3729](#), [3738](#), [3745](#)).
- M. Dodangeh, L. N. Vicente, and Z. Zhang (2016). “On the optimal order of worst case complexity of direct search”. *Optim. Lett.* 10.4, pp. 699–708. MR: [3477371](#).
- Jean-Pierre Dussault (2015). *Simple unified convergence proofs for the trust-region and a new ARC variant*. Tech. rep. Technical report, University of Sherbrooke, Sherbrooke, Canada.
- Jean-Pierre Dussault and Dominique Orban (2017). *Scalable adaptive cubic regularization methods*. Tech. rep. Technical Report G-2015-109, GERAD, Montréal.
- Francisco Facchinei, Vyacheslav Kungurtsev, Lorenzo Lampariello, and Gesualdo Scutari (Sept. 2017). “Ghost Penalties in Nonconvex Constrained Optimization: Diminishing Stepsizes and Iteration Complexity”. arXiv: [1709.03384](#).
- R. Garmanjani, D. Júdice, and L. N. Vicente (2016). “Trust-region methods without using derivatives: worst case complexity and the nonsmooth case”. *SIAM J. Optim.* 26.4, pp. 1987–2011. MR: [3554884](#).
- Dongdong Ge, Xiaoye Jiang, and Yinyu Ye (2011). “A note on the complexity of  $L_p$  minimization”. *Math. Program.* 129.2, Ser. B, pp. 285–299. MR: [2837883](#).
- Saeed Ghadimi and Guanghui Lan (2016). “Accelerated gradient methods for nonconvex nonlinear and stochastic programming”. *Math. Program.* 156.1-2, Ser. A, pp. 59–99. MR: [3459195](#).
- Stephen M. Goldfeld, Richard E. Quandt, and Hale F. Trotter (1966). “Maximization by quadratic hill-climbing”. *Econometrica* 34, pp. 541–551. MR: [0216735](#) (cit. on pp. [3731](#), [3740](#), [3741](#)).

- Nicholas I. M. Gould, Stefano Lucidi, Massimo Roma, and Philippe L. Toint (1998). “[A linesearch algorithm with memory for unconstrained optimization](#)”. In: *High performance algorithms and software in nonlinear optimization (Ischia, 1997)*. Vol. 24. Appl. Optim. Kluwer Acad. Publ., Dordrecht, pp. 207–223. MR: [1789717](#) (cit. on p. [3759](#)).
- Nicholas I. M. Gould, M. Porcelli, and Philippe L. Toint (2012). “[Updating the regularization parameter in the adaptive cubic regularization algorithm](#)”. *Comput. Optim. Appl.* 53.1, pp. 1–22. MR: [2964833](#).
- Geovani N. Grapiglia, Jinyun Yuan, and Ya-xiang Yuan (2015). “[On the convergence and worst-case complexity of trust-region and regularization methods for unconstrained optimization](#)”. *Math. Program.* 152.1-2, Ser. A, pp. 491–520. MR: [3369490](#).
- Geovani Nunes Grapiglia, Yurii Nesterov, et al. (2016). *Globally Convergent Second-order Schemes for Minimizing Twice-differentiable Functions*. Tech. rep. Technical Report CORE Discussion paper 2016/28, CORE, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).
- Geovani Nunes Grapiglia, Jinyun Yuan, and Ya-xiang Yuan (2016). “[Nonlinear stepsize control algorithms: complexity bounds for first- and second-order optimality](#)”. *J. Optim. Theory Appl.* 171.3, pp. 980–997. MR: [3575654](#).
- S. Gratton, Clément W. Royer, and L. N. Vicente (2017). *A decoupled first/second-order steps technique for nonconvex nonlinear unconstrained optimization with improved complexity bounds*. Tech. rep. Technical Report TR 17-21, Department of Mathematics, University of Coimbra, Coimbra, Portugal.
- S. Gratton, Clément W. Royer, L. N. Vicente, and Z. Zhang (2015). “[Direct search based on probabilistic descent](#)”. *SIAM J. Optim.* 25.3, pp. 1515–1541. MR: [3376788](#).
- Serge Gratton, Annick Sartenaer, and Philippe L. Toint (2008). “[Recursive trust-region methods for multiscale nonlinear optimization](#)”. *SIAM J. Optim.* 19.1, pp. 414–444. MR: [2403039](#) (cit. on pp. [3730](#), [3733](#), [3743](#), [3745](#)).
- Andreas Griewank (1981). *The modification of Newton’s method for unconstrained optimization by bounding cubic terms*. Tech. rep. Technical Report NA/12, Department of Applied Mathematics and Theoretical Physics, University of Cambridge, Cambridge, United Kingdom (cit. on pp. [3730](#), [3738](#), [3739](#), [3757](#)).
- Mingyi Hong (Apr. 2016). “[Decomposing Linearly Constrained Nonconvex Problems by a Proximal Primal Dual Approach: Algorithms, Convergence, and Applications](#)”. arXiv: [1604.00543](#).
- Florian Jarre (2013). “[On Nesterov’s smooth Chebyshev-Rosenbrock function](#)”. *Optim. Methods Softw.* 28.3, pp. 478–484. MR: [3060944](#) (cit. on p. [3761](#)).
- Bo Jiang, Tianyi Lin, Shiqian Ma, and Shuzhong Zhang (May 2016). *Structured Non-convex and Nonsmooth Optimization: Algorithms and Iteration Complexity Analysis*. arXiv: [1605.02408](#) (cit. on p. [3761](#)).

- Sha Lu, Zengxin Wei, and Lue Li (2012). “A trust region algorithm with adaptive cubic regularization methods for nonsmooth convex minimization”. *Comput. Optim. Appl.* 51.2, pp. 551–573. MR: [2891907](#).
- José Mario Martínez (2017). “On high-order model regularization for constrained optimization”. *SIAM J. Optim.* 27.4, pp. 2447–2458. MR: [3735301](#).
- José Mario Martínez and Marcos Raydan (2017). “Cubic-regularization counterpart of a variable-norm trust-region method for unconstrained minimization”. *Journal of Global Optimization* 68.2, pp. 367–385.
- Yurii Nesterov (2004). “Introductory lectures on convex optimization: A basic course” (cit. on pp. [3730](#), [3734](#), [3738](#), [3754](#)).
- Yurii Nesterov and B. T. Polyak (2006). “Cubic regularization of Newton method and its global performance”. *Math. Program.* 108.1, Ser. A, pp. 177–205. MR: [2229459](#) (cit. on pp. [3730](#), [3731](#), [3739](#), [3757](#), [3761](#)).
- Jorge Nocedal and Stephen J. Wright (1999). *Numerical optimization*. Springer Series in Operations Research. Springer-Verlag, New York, pp. xxii+636. MR: [1713114](#) (cit. on pp. [3730](#), [3739](#), [3745](#)).
- Clément W. Royer and Stephen J. Wright (2017). “Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization”. Technical report, University of Wisconsin, Madison, USA. arXiv: [1706.03131](#) (cit. on pp. [3731](#), [3732](#), [3759](#)–[3761](#)).
- Katya Scheinberg and Xiaocheng Tang (2013). *Complexity of inexact proximal Newton methods*. Tech. rep. Technical report, Lehigh University, Bethlehem, USA.
- (2016). “Practical inexact proximal quasi-Newton method with global complexity analysis”. *Math. Program.* 160.1-2, Ser. A, pp. 495–529. MR: [3555397](#).
- Kenji Ueda and Nobuo Yamashita (2010a). “Convergence properties of the regularized Newton method for the unconstrained nonconvex optimization”. *Appl. Math. Optim.* 62.1, pp. 27–46. MR: [2653894](#).
- (2010b). “On a global complexity bound of the Levenberg-Marquardt method”. *J. Optim. Theory Appl.* 147.3, pp. 443–453. MR: [2733986](#).
- Stephen A. Vavasis (1993). “Black-box complexity of local minimization”. *SIAM J. Optim.* 3.1, pp. 60–80. MR: [1202002](#) (cit. on p. [3730](#)).
- Luís Nunes Vicente (2013). “Worst case complexity of direct search”. *EURO Journal on Computational Optimization* 1, pp. 143–153.
- Martin Weiser, Peter Deuffhard, and Bodo Erdmann (2007). “Affine conjugate adaptive Newton methods for nonlinear elastomechanics”. *Optim. Methods Softw.* 22.3, pp. 413–431. MR: [2319241](#) (cit. on pp. [3739](#), [3757](#)).
- Peter Whittle (1996). *Optimal control*. Wiley-Interscience Series in Systems and Optimization. Basics and beyond. John Wiley & Sons, Ltd., Chichester, pp. x+464. MR: [1416566](#).



Peng Xu, Farbod Roosta-Khorasani, and Michael W. Mahoney (Aug. 2017). “[Newton-Type Methods for Non-Convex Optimization Under Inexact Hessian Information](#)”. arXiv: 1708.07164.

Received 2017-09-20.

CORALIA CARTIS  
MATHEMATICAL INSTITUTE  
OXFORD UNIVERSITY  
OXFORD OX2 6GG  
ENGLAND  
UNITED KINGDOM  
[coralia.cartis@maths.ox.ac.uk](mailto:coralia.cartis@maths.ox.ac.uk)

NICHOLAS I. M. GOULD  
SCIENTIFIC COMPUTING DEPARTMENT  
STFC-RUTHERFORD APPLETON LABORATORY  
CHILTON, OXFORDSHIRE, OX11 0QX  
ENGLAND  
UNITED KINGDOM  
[nick.gould@stfc.ac.uk](mailto:nick.gould@stfc.ac.uk)

PHILIPPE L. TOINT  
DEPARTMENT OF MATHEMATICS  
UNIVERSITY OF NAMUR  
61, RUE DE BRUXELLES  
B-5000, NAMUR  
BELGIUM  
[philippe.toint@unamur.be](mailto:philippe.toint@unamur.be)

# INVERSE PROBLEMS FOR LINEAR AND NON-LINEAR HYPERBOLIC EQUATIONS

MATTI LASSAS

## Abstract

We consider inverse problems for hyperbolic equations and systems and the solutions of these problems based on the focusing of waves. Several inverse problems for linear equations can be solved using control theory. When the coefficients of the modelling equation are unknown, the construction of the point sources requires solving blind control problems. For non-linear equations we consider a new artificial point source method that applies the non-linear interaction of waves to create microlocal points sources inside the unknown medium. The novel feature of this method is that it utilizes the non-linearity as a tool in imaging, instead of considering it as a difficult perturbation of the system. To demonstrate the method, we consider the non-linear wave equation and the coupled Einstein and scalar field equations.

## 1 Introduction

One of the simplest models for waves is the linear hyperbolic equation

$$\partial_t^2 u(t, x) - c(x)^2 \Delta u(t, x) = 0 \quad \text{in } \mathbb{R} \times \Omega$$

where  $\Omega \subset \mathbb{R}^n$  and  $c(x)$  is the wave speed. This equation models e.g. acoustic waves. In inverse problems one has access to measurements of waves (the solutions  $u(t, x)$ ) on the boundary, or in a subset of the domain  $\Omega$ , and one aims to determine unknown coefficients (e.g.,  $c(x)$ ) in the interior of the domain.

In particular, we will consider *anisotropic* materials, where the wave speed depends on the direction of propagation. This means that the scalar wave speed  $c(x)$ , where  $x =$

---

The author was partly supported by Academy of Finland.

MSC2010: primary 35R30; secondary 35Q91, 49J20, 83C35, 53C50.

Keywords: Inverse problems, non-linear hyperbolic equations, point sources.

$(x^1, x^2, \dots, x^n) \in \Omega$ , is replaced by a positive definite symmetric matrix  $(g^{jk}(x))_{j,k=1}^n$ , and the wave equation takes for example the form

$$(1) \quad \frac{\partial^2}{\partial t^2} u(t, x) - \sum_{j,k=1}^n g^{jk}(x) \frac{\partial^2 u}{\partial x^j \partial x^k}(t, x) = 0.$$

Anisotropic materials appear frequently in applications such as in seismic imaging, where one wishes to determine the interior structure of the Earth by making various measurements of waves on its surface.

It is convenient to interpret the anisotropic wave speed  $(g^{jk})$  as the inverse of a Riemannian metric, thus modelling the medium as a *Riemannian manifold*. This is due to fact that if  $\Psi : \Omega \rightarrow \Omega$  is a diffeomorphism such that  $\Psi|_{\partial\Omega} = Id$  (and for an equation of the form (1) it is also assumed to be volume preserving in  $\Omega \subset \mathbb{R}^n$ ), then all boundary measurements for the metric  $g$  and the pull-forward metric  $\Psi^*g$  coincide. Thus to prove uniqueness results for inverse problems, one has to consider properties that are invariant in diffeomorphisms and try to reconstruct those uniquely, for example, to show that an underlying manifold structure can be uniquely determined. In practice, the inverse problem in a subset of the Euclidean space is solved in two steps. The first is to reconstruct the underlying manifold structure. The second step is to find an embedding of the constructed manifold in the Euclidean space using additional a priori information. In this paper we conspase-timeate on the first step.

## 2 Inverse problems for linear equations

In this section we review the classical results for Gel'fand inverse problems [Gelfand \[1954\]](#) for linear scalar wave equations. Note that these results require that the coefficients of the equation, or at least the leading order coefficients, are time independent. In addition, it is required that the associated operator is selfadjoint or that it satisfies strong geometrical assumptions, for example that all geodesics exit the domain at a given time. In [Section 4](#) we show how these results can be obtained using a focusing of waves that produces point sources inside the unknown medium. In [Sections 3](#) and [5](#) we consider inverse problems non-linear hyperbolic equations and systems, and consider the recently developed artificial point source method based on the non-linear interaction of waves.

Let  $(N, g)$  be an  $n$ -dimensional Riemannian manifold and consider the wave equation

$$(2) \quad \begin{aligned} \partial_t^2 u(t, x) - \Delta_g u(t, x) &= 0 \quad \text{in } (0, \infty) \times N, \\ \partial_\nu u|_{\mathbb{R}_+ \times \partial N} &= f, \quad u|_{t=0} = 0, \quad \partial_t u|_{t=0} = 0, \end{aligned}$$

where  $\Delta_g$  is the Laplace–Beltrami operator corresponding to a smooth time-independent Riemannian metric  $g$  on  $N$ . In coordinates  $(x_j)_{j=1}^n$  this operator has the representation

$$\Delta_g u = \sum_{j,k=1}^n \det(g)^{-1/2} \frac{\partial}{\partial x^j} \left( \det(g)^{1/2} g^{jk} \frac{\partial}{\partial x^k} u \right),$$

where  $g(x) = [g_{jk}(x)]_{j,k=1}^n$ ,  $\det(g) = \det(g_{jk}(x))$  and  $[g^{jk}]_{j,k=1}^n = g(x)^{-1}$ .

The solution of (2), corresponding to the boundary value  $f$  (which is interpreted as a boundary source), is denoted by  $u^f = u^f(t, x)$ .

Let us assume that the boundary  $\partial N$  is known. The inverse problem is to reconstruct the manifold  $N$  and the metric  $g$  when we are given the set

$$\{(f, u^f|_{\mathbb{R}_+ \times \partial N}) : f \in C_0^\infty(\mathbb{R}_+ \times \partial N)\},$$

that is, the Cauchy data of solutions corresponding to all possible boundary sources  $f \in C_0^\infty(\mathbb{R}_+ \times \partial N)$ . This data is equivalent to the *response operator*

$$(3) \quad \Lambda_{N,g} : f \mapsto u^f|_{\mathbb{R}_+ \times \partial N},$$

which is also called the *Neumann-to-Dirichlet map*. Physically,  $\Lambda_{N,g} f$  describes the measurement of the medium response to any applied boundary source  $f$ . In 1990s, the combination of Belishev's and Kurylev's boundary control method [Belishev and Y. V. Kurylev \[1992\]](#) and Tataru's unique continuation theorem [Tataru \[1995\]](#) gave a solution to the inverse problem of determining the isometry type of a Riemannian manifold  $(N, g)$  with given boundary  $\partial N$  and the Neumann-to-Dirichlet map  $\Lambda_{N,g}$ .

**Theorem 2.1** ([Belishev and Y. V. Kurylev \[1992\]](#) and [Tataru \[1995\]](#)). *Let  $(N_1, g_1)$  and  $(N_2, g_2)$  be compact smooth Riemannian manifolds with boundary. Assume that there is a diffeomorphism  $\Phi : \partial N_1 \rightarrow \partial N_2$  such that*

$$(4) \quad \Phi^*(\Lambda_{N_1, g_1} f) = \Lambda_{N_2, g_2}(\Phi^* f), \quad \text{for all } f \in C_0^\infty(\mathbb{R}_+ \times \partial N_1).$$

*Then  $(N_1, g_1)$  and  $(N_2, g_2)$  are isometric Riemannian manifolds.*

Above,  $\Phi^* f$  is the pull-back of  $f$  in  $\Phi$ . [Theorem 2.1](#) can be used to prove the uniqueness of other inverse problems. Katchalov, Kurylev, Mandache, and the author showed in [Katchalov, Y. Kurylev, Lassas, and Mandache \[2004\]](#) the equivalence of spectral inverse problems with several different measurements, that in particular implies the following result.

**Theorem 2.2** ([Katchalov, Y. Kurylev, Lassas, and Mandache \[ibid.\]](#)). *Let  $\partial N$  be given. Then the Neumann-to-Dirichlet map  $\Lambda : \partial_\nu u|_{\mathbb{R}_+ \times \partial N} \mapsto u|_{\mathbb{R}_+ \times \partial N}$ , for heat equation  $(\partial_t -$*

$\Delta_g)u = 0$ , or for the Schrödinger equation  $(i\partial_t - \Delta_g)u = 0$ , with vanishing initial data  $u|_{t=0} = 0$  determine the Neumann-to-Dirichlet map for the wave equation, and therefore, the manifold  $(N, g)$  up to an isometry.

The stability of the solutions of the above inverse problems have been analyzed in Anderson, Katsuda, Y. Kurylev, Lassas, and Taylor [2004], Bao and Zhang [2014], Bosi, Y. Kurylev, and Lassas [2017], and P. Stefanov and G. Uhlmann [2005].

Without making strong assumptions about the geometry of the manifold, the existing uniqueness results for linear hyperbolic equations with vanishing initial data are limited to equations whose coefficients are time independent or real analytic in time (see e.g. Anderson, Katsuda, Y. Kurylev, Lassas, and Taylor [2004], Belishev and Y. V. Kurylev [1992], Eskin [2017], Katchalov, Y. Kurylev, and Lassas [2001], Y. Kurylev, Oksanen, and Paternain [n.d.], and Oksanen [2013]). The reason for this is that these results are based on Tataru's unique continuation theorem Tataru [1995]. This sharp unique continuation result does not work for general wave equations whose coefficients are not real analytic in time, as shown by Alinhac [1983]. Alternatively, one can study inverse problems for hyperbolic equations by using the Fourier transform in the time variable and reducing the problem to an inverse boundary spectral problem for an elliptic equation. Note that this also requires that the coefficients are time independent. The obtained inverse spectral problems (see e.g. A. Nachman, Sylvester, and G. Uhlmann [1988]) can be solved using the complex geometrical optics introduced in Sylvester and G. Uhlmann [1987].

**Open Problem 1:** Do the boundary  $\partial N$  and the Neumann-to-Dirichlet map for a wave equation  $\square_g u = 0$  determine the coefficient  $g^{jk}(t, x)$  that depends on variables  $t$  and  $x$ ?

In many applications, waves can not be detected on the part of the boundary where sources are applied, that is, one is given only a restricted Neumann-to-Dirichlet map. Next we consider such problems.

We say that (2) is exactly controllable from  $\Gamma_1 \subset \partial N$  if there is  $T > 0$  such that the map

$$(5) \quad \begin{aligned} \mathcal{U} : L^2((0, T) \times \Gamma_1) &\rightarrow L^2(N) \times H^{-1}(N), \\ \mathcal{U}(f) &= (u^f(T), \partial_t u^f(T)) \end{aligned}$$

is surjective. In 1992, Bardos, Lebeau, and Rauch gave a sufficient geometric condition for exact controllability and showed that this condition is also close to being necessary Bardos, Lebeau, and Rauch [1992]. Roughly speaking, this geometric controllability condition requires that all geodesics (that reflect from the boundary) in the domain  $N$  intersect transversally to the set  $\Gamma_1$  before time  $T$ .

Under the geometric controllability condition the inverse problem with a restricted Neumann-to-Dirichlet map can be solved using exact controllability results [Lassas and Oksanen \[2014\]](#). However, in the general setting the following problem is open.

**Open Problem 2:** Assume that we are given open subsets  $\Gamma_1, \Gamma_2 \subset \partial N$ , such that  $\bar{\Gamma}_1 \cap \bar{\Gamma}_2 = \emptyset$ , and the restricted Neumann-to-Dirichlet map  $\Lambda_{\Gamma_1, \Gamma_2} : f \mapsto u^f|_{\mathbb{R}_+ \times \Gamma_2}$  defined for functions  $f \in C_0^\infty(\mathbb{R}_+ \times \Gamma_1)$ . Do these data determine  $(N, g)$  up to an isometry?

Similarly, systems having terms causing energy absorption can be considered when the geometric controllability condition is valid [Y. Kurylev and Lassas \[2000\]](#), but the following problem is open.

**Open Problem 3:** Consider [Equation \(2\)](#) where the Laplace operator  $\Delta_g$  is replaced by a non-selfadjoint operator, for example, the wave equation of the form  $(\partial_t^2 - \Delta_g + q(x))u(t, x) = 0$ , where  $q(x)$  is complex valued. Do the boundary  $\partial N$  and Neumann-to-Dirichlet map  $\Lambda_{N, g, q}$  for this equation determine  $(N, g)$  and  $q(x)$  up to an isometry?

The boundary control method has been used to solve inverse problems for some hyperbolic systems of equations, e.g. for Maxwell and Dirac equations, see [Y. Kurylev, Lassas, and Somersalo \[2006\]](#) in the special cases when the wave velocity is independent of polarization.

**Open Problem 4:** Consider a hyperbolic system of equations where the velocity of waves depends on the polarisation, such as elastic equations or Maxwell's equations in anisotropic medium. Do  $\partial N$  and the response operator defined on the boundary determine the system up to a diffeomorphism?

### 3 Inverse problems for non-linear equations

The present theory of inverse problems has largely been confined to the case of linear equations. For the few existing results on non-linear equations (e.g. [Isakov \[1993\]](#), [Salo and Zhong \[2012\]](#), and [Sun and G. Uhlmann \[1997\]](#)) the non-linearity is an obstruction rather than a helpful feature.

Below, we consider inverse problems for non-linear hyperbolic equations and use non-linearity as a tool to solve the problems. This enables us to solve inverse problems for non-linear equations for which the corresponding problems for linear equations are still unsolved (e.g. when the coefficients depend on the time variable or are complex valued, cf. Open Problems 1 and 2). Below, we will first consider scalar wave-equation with simple quadratic non-linearity. Later we consider inverse problems for the Einstein equations that can be solved using the non-linear interaction of gravitational waves and matter field waves. The inverse problems for the Einstein equations (in particular the passive problems

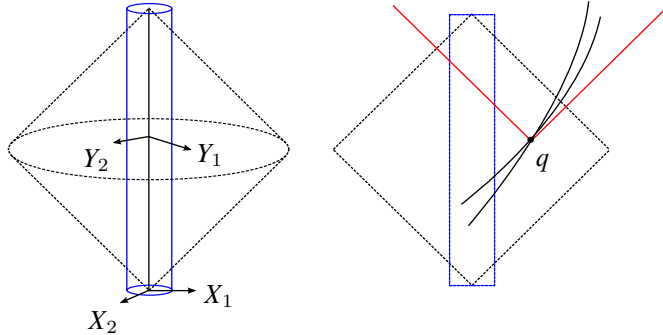


Figure 1: *Left.* The setting of Theorems 3.1 and 5.1. The solid black line depicts the time-like geodesic  $\mu$  and the blue cylinder is its neighbourhood where measurements are made. The dashed double cone is the set  $I(p^-, p^+)$  which properties are reconstructed from the data. In Theorem 5.1 we use the frame  $Y_1, Y_2, Y_3$  moving along  $\mu$  to define the Fermi coordinates in the blue cylinder (the third direction is suppressed in the picture). *Right.* A schematic picture of the proof of Theorem 3.1. Geodesics, depicted as black curves, that are sent from the neighbourhood of  $\mu$  intersect at the point  $q \in I(p^-, p^+)$ . We consider (four) distorted plane waves that propagate near the geodesics that interact at the point  $q$  and produce propagating singularities (in red), analogous to those generated by a point source at  $q$ .

considered below) could be applied in the gravitational astronomy initiated by the direct detection of gravitational waves B. P. Abbott et al. [2016].

**3.1 Notation.** Let  $(M, g)$  be a  $(1 + 3)$ -dimensional time-oriented Lorentzian manifold of signature  $(-, +, +, +)$ . Let  $q \in M$ . The set of future pointing *light-like* vectors at  $q$  is defined by

$$L_q^+ M = \{\theta \in T_q M \setminus 0 : g(\theta, \theta) = 0, \theta \text{ is future-pointing}\}.$$

A vector  $\theta \in T_q M$  is *time-like* if  $g(\theta, \theta) < 0$  and *space-like* if  $g(\theta, \theta) > 0$ . *Causal vectors* are the collection of time-like and light-like vectors, and a curve  $\gamma$  is time-like (light-like, causal, future-pointing) if the tangent vectors  $\dot{\gamma}$  are time-like (light-like, causal, future-pointing).

For  $p, q \in M$ , the notation  $p \ll q$  means that  $p, q$  can be joined by a future-pointing time-like curve. The *chronological future* and *past* of  $p \in M$  are

$$I^+(p) = \{q \in M : p \ll q\}, \quad I^-(p) = \{q \in M : q \ll p\}.$$

To emphasise the Lorentzian structure of  $(M, g)$  we sometimes write  $I_{M,g}^{\pm}(p) = I^{\pm}(p)$ . We will denote throughout the paper

$$(6) \quad I(p, q) = I^{+}(p) \cap I^{-}(q).$$

A time-oriented Lorentzian manifold  $(M, g)$  is *globally hyperbolic* if there are no closed causal paths in  $M$ , and for any  $p, q \in M$  the set  $J(p, q)$  is compact. The set  $J(p, q)$  is defined analogously to  $I(p, q)$  but with the partial order  $p \ll q$  replaced by  $p \leq q$ , meaning that  $p$  and  $q$  can be joined by a future-pointing causal curve or  $p = q$ . According to [Bernal and Sánchez \[2005\]](#), a globally hyperbolic manifold is isometric to the product manifold  $\mathbb{R} \times N$  with the Lorentzian metric given by

$$(7) \quad g = -\beta(t, y)dt^2 + \kappa(t, y),$$

where  $\beta : \mathbb{R} \times N \rightarrow \mathbb{R}_{+}$  and  $\kappa$  is a Riemannian metric on  $N$  depending on  $t$ .

**3.2 Active measurements.** Let  $(M, g)$  be a 4-dimensional globally hyperbolic Lorentzian manifold and assume, without loss of generality, that  $M = \mathbb{R} \times N$  with a metric of the form (7). Let  $t_0 > 0$  and consider the semilinear wave equation

$$(8) \quad \square_g u(x) + a(x)u(x)^2 = f(x), \quad \text{for } x \in (-\infty, t_0) \times N,$$

$$(9) \quad u = 0, \quad f = 0, \quad \text{in } (-\infty, 0) \times N.$$

Here  $a \in C^{\infty}(M)$  is a nowhere vanishing function that may be complex valued, and

$$\square_g u = \sum_{j,k=0}^n |\det(g)|^{-1/2} \frac{\partial}{\partial x^j} \left( |\det(g)|^{1/2} g^{jk} \frac{\partial}{\partial x^k} u \right).$$

Let  $\mu \subset (0, t_0) \times N$  be a time-like curve and  $V$  be its open neighbourhood. The solution of (8)–(9) exists when the source  $f$  is supported in  $V$  and satisfies  $\|f\|_{C^k(\overline{V})} < \varepsilon$ , where  $k \in \mathbb{Z}_{+}$  is sufficiently large and  $\varepsilon > 0$  is sufficiently small. For such sources  $f$  we define the measurement operator

$$(10) \quad L_V : f \mapsto u|_V.$$

Note  $L_V$  is equivalent to its graph that is given by the data set

$$(11) \quad \mathfrak{D}_{L_V} = \{(u|_V, f) : u \text{ and } f \text{ satisfy (8),(9), } f \in C_0^k(V), \|f\|_{C^k(\overline{V})} < \varepsilon\}.$$

**Theorem 3.1** ([Y. Kurylev, Lassas, and G. Uhlmann \[2014\]](#)). *Let  $(M, g)$  be a globally hyperbolic 4-dimensional Lorentzian manifold. Let  $\mu$  be a time-like path containing  $p^{+}$*



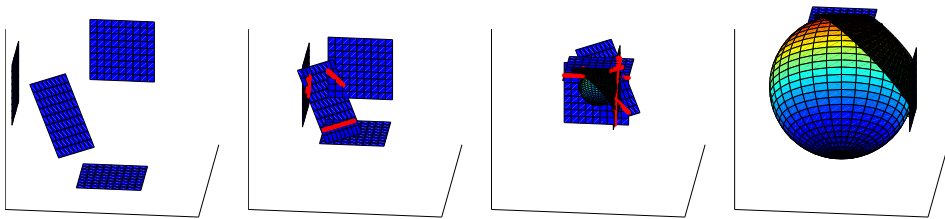


Figure 2: Four plane waves propagate in space. When the planes intersect, the non-linearity of the hyperbolic system produces new waves. *Left*: Plane waves before interacting. *Middle left*: The two-wave interactions (red line segments) appear but do not cause singularities propagating to new directions. *Middle right and Right*: All plane waves have intersected and new waves have appeared. The three-wave interactions cause conic waves (the black surface). Only one such wave is shown in the figure. The interaction of four waves causes a microlocal point source that sends a spherical wave in all future light-like directions.

and  $p^-$ . Let  $V \subset M$  be a neighborhood of  $\mu$  and let  $a : M \rightarrow \mathbb{R}$  be a nowhere vanishing  $C^\infty$ -smooth function. Then  $(V, g|_V)$  and the measurement operator  $L_V$  determine the topology, differentiable structure and the conformal class of the metric  $g$  in the double cone  $I_{M,g}(p^-, p^+)$ .

When  $M$  has a significant Ricci-flat part, [Theorem 3.1](#) can be strengthened.

**Corollary 3.2.** *Assume that  $(M, g)$  and  $V$  satisfy the conditions of [Theorem 3.1](#). Moreover, assume that  $W \subset I_{M,g}(p^-, p^+)$  is Ricci-flat and all topological components of  $W$  intersect  $V$ . Then the metric tensor  $g$  is determined in  $W$  uniquely.*

The proof of [Theorem 3.1](#) uses the results on the inverse problem for passive measurements for point sources, described below, and the non-linear interaction of waves having conormal singularities. There are many results on such non-linear interaction, starting with the studies of [Bony \[1986\]](#), [R. Melrose and Ritter \[1985\]](#), [Holt \[1995\]](#). However, these studies differ from the proof of [Theorem 3.1](#) in that they assumed that the geometrical setting of the interacting singularities, and in particular the locations and types of caustics, is known a priori. In inverse problems we study waves on an unknown manifold, so we do not know the underlying geometry and, therefore, the location of the singularities of the waves. For example, the waves can have caustics that may even be of an unstable type.

[Theorem 3.1](#) only concerns the recovery of the conformal type of the metric. The recovery of all coefficients up to a natural gauge transformation has in some special cases been

considered (in [Lassas, G. Uhlmann, and Wang \[n.d.\]](#) and [Wang and Zhou \[2016\]](#)), but for general equations both the complete recovery of all coefficients and the stable solvability of the inverse problem are open questions.

**Open Problem 5 (Recovery of all coefficients for non-linear wave equation):** Assume that we are given a time-like path  $\mu$ , its neighborhood  $V \subset M$ , and the map  $L_V : f \mapsto u|_V$  for the non-linear equation  $\square_g u + B(x, D)u + a(x)u(x)^2 = f$ , defined for small sources  $f$  supported in  $V$ , where  $B(x, D)$  is a first order differential operator. Is it possible to construct the metric tensor  $g$  and the operator  $B(x, D)$  in  $I(p^-, p^+)$  up to a local gauge transformation?

**Open Problem 6 (Stability of the inverse problem for non-linear wave equation):** Assume that we are given a time-like curve  $\mu$ , its neighborhood  $V \subset M$ , the map  $L_V$  with an error, and  $p^-, p^+ \in \mu$ . Is it possible to construct the set  $I(p^-, p^+)$  and the metric  $g$  in  $I(p^-, p^+)$  with an error that can be estimated in terms of the geometric bounds for  $M$  and the error in the given data?

For certain inverse problems for linear wave equations the essential features of several measurements can be packed in a single measurement [Helin, Lassas, and Oksanen \[2014\]](#) and [Helin, Lassas, Oksanen, and Saksala \[2016\]](#). The corresponding problem for non-linear equations is open.

**Open Problem 7 (Single measurement inverse problem for non-linear wave equations):** Can we construct a source  $f$  such that the set  $V$  and the measurement  $L_V f$  uniquely determine  $I(p^-, p^+)$  and the metric  $g$  on  $I(p^-, p^+)$ ?

**3.3 Passive measurements.** The earliest light observation set is an idealized notion of measurements of light coming from a point source.

**Definition 3.3.** Let  $M$  be a Lorentzian manifold,  $V \subset M$  be open, and  $q \in M$ . The light observation set of  $q \in M$  in  $V$  is

$$\mathcal{P}_V(q) = \{\gamma_{q,\xi}(t) \in M : t \geq 0, \xi \in L_q^+ M\} \cap V,$$

where  $\gamma_{q,\xi}$  denotes the geodesic emanating from  $q$  to the direction  $\xi$ . The earliest light observation set of  $q \in M$  in  $V$  is

$$\mathcal{E}_V(q) = \{x \in \mathcal{P}_V(q) : \text{there are no } y \in \mathcal{P}_V(q) \text{ such that } y \ll x \text{ in } (V, g)\}.$$

The set  $\mathcal{P}_V(q)$  can be viewed as a model of a measurement where light emitted by a point source at  $q$  is recorded in  $V$ . As gravitational wave packets propagate at the speed of light,  $\mathcal{P}_V(q)$  could also correspond to an observation where a gravitational wave is generated at  $q$  and detected in  $V$ . The set  $\mathcal{E}_V(q)$  is related to the distance difference functions used in Riemannian geometry.

**Definition 3.4.** Let  $N$  be a Riemannian manifold with the distance function  $\text{dist}_N(x, y)$  and let  $U \subset N$  be an open set. The distance difference function in the observation set  $U$  corresponding to a point  $x \in N$  is

$$(12) \quad D_x : U \times U \rightarrow \mathbb{R}, \quad D_x(z_1, z_2) := \text{dist}_N(z_1, x) - \text{dist}_N(z_2, x).$$

Consider a Riemannian manifold where the distance between two points is the travel time of waves between these points. When a *spontaneous point source* produces a wave at some unknown point  $x \in N$ , at some unknown time  $t \in \mathbb{R}$ , the produced wave is observed at the point  $z \in U$  at time  $T_{t,x}(z) = \text{dist}_N(z, x) + t$ . These observation times at two points  $z_1, z_2 \in U$  determine the distance difference function by

$$D_x(z_1, z_2) = T_{t,x}(z_1) - T_{t,x}(z_2) = \text{dist}_N(z_1, x) - \text{dist}_N(z_2, x).$$

Physically, this function corresponds to the difference in times when the wave produced by a point source at  $(t, x)$  is observed at  $z_1$  and  $z_2$ .

When  $M = \mathbb{R} \times N$  is the Lorentzian manifold given by the product metric of  $N$  and  $(\mathbb{R}, -dt^2)$ , the earliest light observation set corresponding to a point  $q = (t_0, x_0)$  and  $V = \mathbb{R} \times U$ , where  $x \in N$  and  $t_0 \in \mathbb{R}$ , is given by

$$\mathcal{E}_V(q) = \{(t, y) \in \mathbb{R} \times U : \text{dist}_N(y, x_0) = t - t_0\}.$$

Similarly, the earliest light observation set  $\mathcal{E}_V(q)$  corresponding to  $q = (t_0, x_0)$  determines the distance difference function  $D_{x_0}$  by

$$(13) \quad D_{x_0}(z_1, z_2) = t_1 - t_2, \quad \text{if } \exists t_1, t_2 \in \mathbb{R} \text{ such that } (t_1, z_1), (t_2, z_2) \in \mathcal{E}_V(q).$$

The following theorem says, roughly speaking, that observations of a large number of point sources in a region  $W$  determine the structure of the spacetime in  $W$ , up to a conformal factor.

**Theorem 3.5** (Y. Kurylev, Lassas, and G. Uhlmann [2014]). Let  $(M_j, g_j)$ , where  $j = 1, 2$ , be two open globally hyperbolic Lorentzian manifolds of dimension  $1 + n$ ,  $n \geq 2$ . Let  $\mu_j : [0, 1] \rightarrow M_j$  be a future-pointing time-like path, let  $V_j \subset M_j$  be a neighbourhood of  $\mu_j([0, 1])$ , and let  $W_j \subset I_{M_j, g_j}^-(\mu_j(1)) \setminus I_{M_j, g_j}^-(\mu_j(0))$  be open and relatively compact,  $j = 1, 2$ . Assume that there is a conformal diffeomorphism  $\phi : V_1 \rightarrow V_2$  such that  $\phi(\mu_1(s)) = \mu_2(s)$ ,  $s \in [0, 1]$ , and

$$\{\phi(\mathcal{E}_{V_1}(q)) : q \in W_1\} = \{\mathcal{E}_{V_2}(q) : q \in W_2\}.$$

Then there is a diffeomorphism  $\Psi : W_1 \rightarrow W_2$  and a strictly positive function  $\alpha \in C^\infty(W_1)$  such that  $\Psi^*g_2 = \alpha g_1$  and  $\Psi|_{V_1 \cap W_1} = \phi$ .

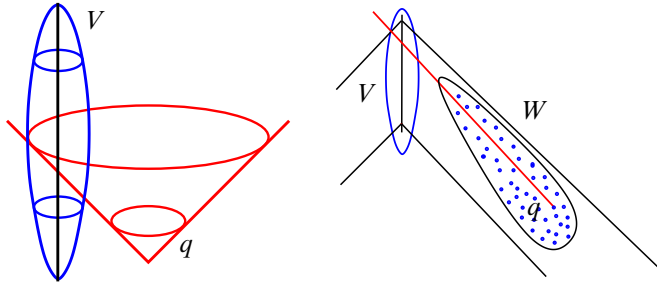


Figure 3: *Left.* When there are no cut points, the earliest light observation set  $\mathcal{E}_V(q)$  is the intersection of the cone and the open set  $V$ . The cone is the union of future-pointing light-like geodesics from  $q$ , and the ellipsoid depicts  $V$ . *Right.* The setting of Theorem 3.5. The domain  $W \subset M$  (with a black boundary) contains several points sources and a light ray from the point  $q \in W$  reaches the observation set  $V$  (with a blue boundary).

In the Riemannian case, the whole metric can be determined under conditions described in the following theorem.

**Theorem 3.6** (Lassas and Saksala [2015]). *Let  $(N, g)$  be a connected Riemannian manifold without boundary, that is either complete or compact, of the dimension  $n \geq 2$ . Let  $W \subset N$  be a compact set with non-empty complement  $U = M \setminus W$ . Then the pair  $(U, g|_U)$  and the distance difference functions  $\{D_x \in C(U \times U) : x \in W\}$  uniquely determine the manifold  $(N, g)$  up to an isometry.*

A classical distance function representation of a compact Riemannian manifold  $N$  is the Kuratowskii embedding,  $\mathcal{K} : x \mapsto \text{dist}_N(x, \cdot)$ , from  $N$  to the space of the continuous functions  $C(N)$  on it. The mapping  $\mathcal{K} : N \rightarrow C(N)$  is an isometry so that  $\mathcal{K}(N)$  is an isometric representation of  $N$  in a vector space  $C(N)$ . Next we consider a similar embedding that is applicable for inverse problems.

Let  $x \in N$  and define a function  $D_x : \overline{U} \times \overline{U} \rightarrow \mathbb{R}$  by formula (12). Let  $\mathfrak{D} : N \rightarrow C(\overline{U} \times \overline{U})$  be given by  $\mathfrak{D}(x) = D_x$ . Theorem 3.6 implies that the set  $\mathfrak{D}(N) = \{D_x : x \in N\}$  can be considered as an embedded image of the manifold  $(N, g)$  in the space  $C(\overline{U} \times \overline{U})$  in the embedding  $x \mapsto D_x$ . Thus,  $\mathfrak{D}(N)$  can be considered as a representation of the manifold  $N$ , given in terms of the distance difference functions, and we call it the *distance difference representation* of the manifold of  $N$  in  $C(\overline{U} \times \overline{U})$ .

The embedding  $\mathfrak{D}$  is different to the above embedding  $\mathcal{K}$  in the following way that makes it important for inverse problems: With  $\mathfrak{D}$  one does not need to know a priori the set  $N$  in order to consider the function space  $C(\overline{U} \times \overline{U})$  where we can embed  $N$ . Indeed, when the observation set  $U$  is given, we can determine the topological properties of  $N$

by constructing the set  $\mathfrak{D}(N)$  that is homeomorphic to  $N$ , and then consider  $\mathfrak{D}(N)$  as a “copy” of the unknown manifold  $N$  embedded in the known function space  $C(\overline{U} \times \overline{U})$ .

## 4 Ideas for proofs and reconstruction methods

**4.1 The focusing of waves for linear equations.** Let  $u^f(t, x)$  denote the solution of the hyperbolic Equation (2), let  $\Lambda = \Lambda_{N,g}$  be the Neumann-to-Dirichlet map for the Equation (2), and let  $dS_g$  denote the Riemannian volume measure on the manifold  $(\partial N, g_{\partial N})$ . We start with the Blagovestchenskii identity Blagoveščenskii [1969] (see also Katchalov, Y. Kurylev, and Lassas [2001]) which states that the inner product of waves at any time can be computed from boundary data.

**Lemma 4.1.** *Let  $f, h \in C_0^\infty(\mathbb{R}_+ \times \partial N)$  and  $T > 0$ . Then*

$$(14) \quad \langle u^f(T), u^h(T) \rangle_{L^2(N)} = \int_N u^f(T, x) u^h(T, x) dV_g(x) = \\ = \frac{1}{2} \int_L \int_{\partial M} (f(t, x)(\Lambda h)(s, x) - (\Lambda f)(t, x)h(s, x)) dS_g(x) dt ds,$$

where  $dV_g$  is the volume measure on the Riemannian manifold  $(N, g)$  and  $L = \{(s, t) \in (\mathbb{R}_+)^2 : 0 \leq t + s \leq 2T, t < s\}$ . A similar formula can be written to compute  $\langle u^f(T), 1 \rangle_{L^2(N)}$  in terms of  $f$ ,  $(\partial N, dS_g)$ , and  $\Lambda$ .

We also need an approximate controllability result that is based on the following fundamental unique continuation theorem of Tataru [1995].

**Theorem 4.2.** *Let  $u(t, x)$  solve the wave equation  $\partial_t^2 u - \Delta_g u = 0$  in  $N \times \mathbb{R}$  and  $u|_{(0, 2T_1) \times \Gamma} = 0$  and  $\partial_\nu u|_{(0, 2T_1) \times \Gamma} = 0$ , where  $\Gamma \subset \partial N$  is open and non-empty. Then  $u(t, x) = 0$  in  $K_{\Gamma, T_1}$ , where*

$$K_{\Gamma, T_1} = \{(t, x) \in \mathbb{R} \times N : \text{dist}_N(x, \Gamma) < T_1 - |t - T_1|\}$$

is the double cone of influence.

The quantitative stability results for Tataru-type unique continuation have recently been obtained by Bosi, Kurylev, and the author, Bosi, Y. Kurylev, and Lassas [2016], and by Laurent and Léautaud Laurent and Léautaud [2015]. Theorem 4.2 gives rise to the following approximate controllability result:

**Corollary 4.3.** *For any open  $\Gamma \subset \partial N$  and  $T_1 > 0$ ,*

$$cl_{L^2(N)}\{u^f(T_1, \cdot) : f \in C_0^\infty((0, T_1) \times \Gamma)\} = L^2(N(\Gamma, T_1)).$$

Here  $N(\Gamma, T_1) = \{x \in N : \text{dist}_N(x, \Gamma) < T_1\}$  is the domain of influence of  $\Gamma$  at time  $T_1$ ,  $cl$  denotes the closure, and  $L^2(N(\Gamma, T_1)) = \{v \in L^2(N) : \text{supp}(v) \subset cl(N(\Gamma, T_1))\}$ .

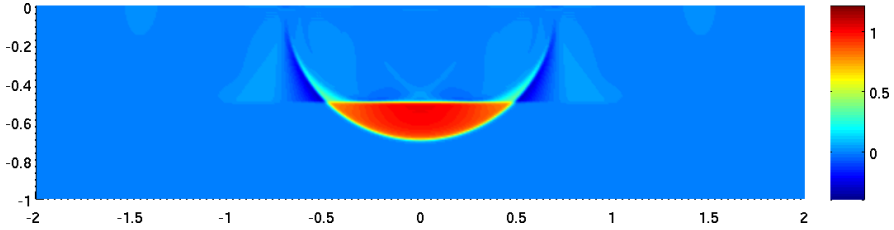


Figure 4: The numerical simulation on family of waves  $u^{f_{\varepsilon}, \alpha}$  that focus to a point as  $\alpha \rightarrow 0$  and  $\varepsilon \rightarrow 0$ , by [de Hoop, Kopley, and Oksanen \[2016\]](#). The figure shows the wave  $u^{f_{\varepsilon}, \alpha}(x, T)$  at the time  $t = T$  in the rectangle  $x \in [-2, 0] \times [-2, 2]$  that is concentrated in the neighborhood  $A'_\varepsilon \setminus A''_\varepsilon$  of the point  $x_1$ . Waves are controlled by a boundary source supported on the top of the rectangle and  $f_{\varepsilon}, \alpha$  is constructed using the local Neumann-to-Dirichlet map.

**4.1.1 Blind control problems.** The inverse problem for the linear wave [Equation \(2\)](#) can be solved by using blind control problems. To consider this approach, we consider first an example of such a control problem.

**Example 1: The blind deconvolution problem.** The problem is to determine unknown functions  $f$  and  $g$  when the convolution  $m = g * f$  is given. Naturally, this problem has no unique solution. In practical settings when a priori assumptions about  $f$  and  $g$  are given, one can approach the problem by solving a regularised problem, for example finding  $(f, g)$  minimizing  $\|f * g - m\|_{L^2(\mathbb{R})}^2 + R(f, g)$  where  $R(f, g) = \alpha(\|f\|_X^2 + \|g\|_Y^2)$  and  $X$  and  $Y$  are suitable Banach spaces, e.g. Sobolev spaces, and  $\alpha > 0$  is a regularization parameter (see e.g. [Mueller and Siltanen \[2012\]](#)).

Below we consider a blind control problem for a wave equation on a compact manifold  $N$ . Our aim is to find a boundary source  $f$  that produces a wave  $u^f(t, x)$  solving the wave equation with metric  $g$  such that at time  $t = T$  the value of the wave,  $u^f(T, x)$ , is close to a function  $m(x)$ . When the domain  $N$  and the metric  $g$  on it are known, this is a traditional control problem. We consider a blind control problem when the metric  $g$  is unknown and we only know  $\partial N$  and the map  $\Lambda$ . Below we are particularly interested in the case when  $m(x) = \chi_A(x)$  is the indicator function of a set  $A = A(z_1, z_2, \dots, z_J; T_0, T_1, T_2, \dots, T_J) \subset N$ ,

$$(15) \quad A = \{x \in N : \text{dist}_N(x, \partial N) < T_0\} \cup \bigcup_{j=1}^J B_N(z_j, T_j),$$

where points  $z_j \in \partial N$ ,  $j = 1, 2, \dots, J$ , and values  $T_j \in (0, T)$  are given and  $B_N(z_j, T_j)$  are the balls of the manifold  $N$  with the centre  $z_j$  and radius  $T_j$ . We consider the minimization problem

$$(16) \quad \min_{f \in Y_A} \|u^f(T) - 1\|_{L^2(N)}^2 + \alpha \|f\|_{L^2([0, T] \times \partial N)}^2$$

where  $Y_A \subset L^2([0, T] \times \partial N)$  is the space of functions  $f(t, x)$ , supported in the union of the sets  $[T - T_0, T] \times \partial N$  and  $\bigcup_{j=1}^J \{(t, z) \in [0, T] \times \partial N : t > T_j - \text{dist}_{\partial M}(z, z_j)\}$ , and  $\alpha > 0$  is a small regularisation parameter.

When  $\alpha \rightarrow 0$ , it follows from [Bingham, Y. Kurylev, Lassas, and Siltanen \[2008\]](#) and [de Hoop, Kepley, and Oksanen \[2016\]](#) that the solutions  $f_\alpha$  of the minimization [Equation \(16\)](#) satisfy

$$(17) \quad \lim_{\alpha \rightarrow 0} u^{f_\alpha}(T) = \chi_A \quad \text{in } L^2(N).$$

Moreover, a modification of the minimization [Equation \(16\)](#) (see [Dahl, Kirpichnikova, and Lassas \[2009\]](#)) has the solution  $\tilde{f}_\alpha$  such that

$$(18) \quad \lim_{\alpha \rightarrow 0} u^{\tilde{f}_\alpha}(T) = \chi_A \quad \text{and} \quad \lim_{\alpha \rightarrow 0} \partial_t u^{\tilde{f}_\alpha}(T) = 0,$$

where limits take place in  $L^2(N)$ .

Using [Lemma 4.1](#) we can solve the above minimization [Equation \(16\)](#) when we do not know the metric  $g$  in the manifold  $N$  but only the boundary measurements given in terms of the Dirichlet-to-Neumann map  $\Lambda_{N,g}$ . By [\(17\)](#), this means that the solutions of the minimization [Equation \(16\)](#) are approximate solutions for a blind control problem. We emphasise that one does not need to assume that the wave equation has an exact controllability property to consider this control problem.

Let  $z_1 \in \partial N$  and  $\nu$  be the unit interior normal of  $\partial N$ , and define the cut-locus function as

$$\tau_{\partial N}(z_1) = \sup\{s > 0 : \text{dist}_N(\gamma_{z_1, \nu}(T_1), \partial N) = s\}.$$

When  $T_1 \in (0, T)$  satisfies  $T_1 < \tau_{\partial N}(z_1)$ , the geodesic  $\gamma_{z_1, \nu}([0, T_1])$  is the shortest curve connecting  $x_1 = \gamma_{z_1, \nu}(T_1)$  to the boundary  $\partial N$ . For  $\varepsilon > 0$ , let

$$\begin{aligned} A'_\varepsilon &= \{x \in N : \text{dist}_N(x, \partial N) < T_1 - \varepsilon\} \cup B_N(z_1, T_1 + \varepsilon), \\ A''_\varepsilon &= \{x \in N : \text{dist}_N(x, \partial N) < T_1 - \varepsilon\} \end{aligned}$$

be sets of the form [\(15\)](#). Then the interior of  $A'_\varepsilon \setminus A''_\varepsilon$  is a small neighbourhood of  $x_1$ . Let  $f'_{\varepsilon, \alpha}$  and  $f''_{\varepsilon, \alpha}$  be the solutions of the minimization problems [\(16\)](#) with objective functions  $\chi_{A'_\varepsilon}$  and  $\chi_{A''_\varepsilon}$ , respectively. When  $\alpha > 0$  is small, [\(17\)](#) implies that the boundary source

$f_{\varepsilon,\alpha} = f'_{\varepsilon,\alpha} - f''_{\varepsilon,\alpha}$  produces a wave  $u^{f_{\varepsilon,\alpha}}(t, x)$  such that  $u^{f_{\varepsilon,\alpha}}(T, x)$  is concentrated in the set  $A'_\varepsilon \setminus A''_\varepsilon$ . Further, when  $\varepsilon \rightarrow 0$ , the set  $A'_\varepsilon \setminus A''_\varepsilon$  tends to the point  $x_1$ .

Numerical methods to constructing the family of focused waves,  $u^{f_{\varepsilon,\alpha}}(T, x)$ , by solving blind control problems similar to (17) have been developed by M. de Hoop, P. Kepley, and L. Oksanen [de Hoop, Kepley, and Oksanen \[2016\]](#) (see Fig. 4).

As discussed above, the minimization [Equation \(17\)](#) can be modified –see [Dahl, Kirpichnikova, and Lassas \[2009\]](#) and (18)– so that their solutions are boundary sources  $\tilde{f}_{\varepsilon,\alpha} \in L^2([0, T] \times \partial N)$  that produce waves  $u^{\tilde{f}_{\varepsilon,\alpha}}(t, x)$  for which the pair  $(u^{\tilde{f}_{\varepsilon,\alpha}}(T, x), \partial_t u^{\tilde{f}_{\varepsilon,\alpha}}(T, x))$  is concentrated near point  $x_1$ . Moreover, when the sources are multiplied by a factor  $c_\varepsilon = 1/\text{vol}(A'_\varepsilon \setminus A''_\varepsilon)$ , we have, in sense of distributions,

$$\lim_{\varepsilon \rightarrow 0} \lim_{\alpha \rightarrow 0} (u^{c_\varepsilon \tilde{f}_{\varepsilon,\alpha}}(T, x), \partial_t u^{c_\varepsilon \tilde{f}_{\varepsilon,\alpha}}(T, x)) = (\delta_{x_1}, 0),$$

where  $\delta_{x_1} \in \mathfrak{D}'(N)$  is the delta distribution supported at  $x_1$ . This implies that the wave  $u^{\tilde{f}_{\varepsilon,\alpha}}(t, x)$  is at times  $t > T$  close to the time derivative of Green's function  $G(t, x; T, x_1)$  corresponding to the point source  $\delta_{x_1}(x)\delta(t - T)$  at  $(T, x_1)$ . Furthermore, the boundary observations of the time derivative  $\partial_t G(t, x; T, x_1)$  determine the boundary values of Green's function  $G(t, x; T, x_1)$ .

For convex manifolds the boundary observations of the above Green's function determine the distance difference function  $D_{x_1}$  corresponding to the point  $x_1$ , see (13). For general manifolds, the distance difference function  $D_{x_1}$  can be constructed by computing the  $L^2$ -norms of the waves  $u^{f_\alpha}(T, x)$ , where  $f_\alpha$  solve the minimization [Equation \(16\)](#) with different sets  $A$  of the form (15), see [Bingham, Y. Kurylev, Lassas, and Siltanen \[2008\]](#). When  $T > \text{diam}(M)/2$ , the above focusing of waves, that creates a point source, can be replicated for arbitrary point  $x_1 \in N$ . Assuming that manifold  $N$  is a subset of a compact or closed manifold  $\tilde{N}$  and that we know the exterior  $\tilde{N} \setminus N$  and the metric on this set, [Theorem 3.6](#) implies that the collection of the distance difference functions  $\mathfrak{D}(N) = \{D_{x_1} : x_1 \in N\}$  determine the isometry type of the Riemannian manifold  $(N, g)$ . A similar construction of manifold  $(N, g)$  can also be made when we are not given the exterior  $\tilde{N} \setminus N$  but when we are given only  $\partial N$  and  $\Lambda$  (see [Katchalov, Y. Kurylev, and Lassas \[2001\]](#)).

**4.2 Non-linear equations and artificial point sources.** Below we consider the non-linear wave and the main ideas used to prove [Theorem 3.1](#).

Let  $f = \epsilon h$ ,  $\epsilon > 0$ , and write an asymptotic expansion of the solution  $u$  of (8),

$$u = \epsilon w_1 + \epsilon^2 w_2 + \epsilon^3 w_3 + \epsilon^4 w_4 + \mathcal{O}(\epsilon^5),$$



where

$$(19) \quad \begin{aligned} w_1 &= \square_g^{-1} h, & w_2 &= -\square_g^{-1}(w_1 \cdot w_1), & w_3 &= -2\square_g^{-1}(w_1 \cdot w_2), \\ w_4 &= -\square_g^{-1}(aw_2 \cdot w_2) - 2\square_g^{-1}(aw_1 \cdot w_3). \end{aligned}$$

We say, for example, that  $w_3$  results from the interaction of  $w_1$  and  $w_2$ , and consider such interactions in general.

Let us consider for the moment  $\mathbb{R}^4$  with the Minkowski metric  $g$ . We can choose in  $\mathbb{R}^4$  coordinates  $x^j$ ,  $j = 1, 2, 3, 4$ , such that the hyperplanes  $K_j = \{x^j = 0\}$  are light-like, that is,  $T_p K_j$  contains a light-like vector for all  $p \in \mathbb{R}^4$ . The plane waves  $u_j(x) = (x^j)_+^m$ , where  $m > 0$ , are solutions to the wave equation  $\square_g u = 0$ . They are singular on the hyperplanes  $K_j$ , in fact, they are conormal distributions in  $I^{-m-1}(N^* K_j)$  (see [Greenleaf and G. Uhlmann \[1993\]](#) and [R. B. Melrose and G. A. Uhlmann \[1979\]](#)).

The proof of [Theorem 3.1](#) is based on an analysis of the interaction of four waves. Analogously to (19), the derivative  $u^{(4)} = \partial_{\epsilon_1} \partial_{\epsilon_2} \partial_{\epsilon_3} \partial_{\epsilon_4} u_{\vec{\epsilon}}|_{\vec{\epsilon}=0}$  of the solution  $u_{\vec{\epsilon}}$  of (8)–(9) with the source

$$f_{\vec{\epsilon}}(x) = \sum_{j=1}^4 \epsilon_j f_j(x), \quad \vec{\epsilon} = (\epsilon_1, \epsilon_2, \epsilon_3, \epsilon_4),$$

is a linear combination of terms such as

$$(20) \quad \widetilde{w}_4 = \square_g^{-1}(S_{1234}), \quad S_{1234} = u_4 \square_g^{-1}(u_3 \square_g^{-1}(u_2 u_1)).$$

Moreover, a suitable choice of  $f_j$ ,  $j = 1, 2, 3, 4$ , guarantees that the term (20) dominates the other terms in  $u^{(4)}$ . For example, two waves  $u_1$  and  $u_2$  are singular on hyperplanes  $K_1$  and  $K_2$ , respectively, and these singularities interact on  $K_1 \cap K_2$ . The interaction of the three waves  $u_1$ ,  $u_2$ , and  $u_3$  happens on the intersection  $K_{123} = K_1 \cap K_2 \cap K_3$  which is a line. As  $N^* K_{123}$  contains light-like directions that are not in union,  $N^* K_1 \cup N^* K_2 \cup N^* K_3$ , this interaction produces interesting singularities that start to propagate. These singularities correspond to the black conic wave in Fig. 3. Finally, singularities of all four waves  $u_j$ ,  $j = 1, 2, 3, 4$  interact at the point  $\{q\} = \bigcap_{j=1}^4 K_j$ . The singularities from the point  $q$  propagate along the light cone emanating from this point and with suitably chosen sources  $f_j$  the wave  $u^{(4)}$  is singular on the light cone  $\mathcal{L}(q)$ . Thus  $S_{1234}$  can be considered as a microlocal point source that sends similar singularities in all directions as a point source located at the point  $q$ , and these singularities are observed in the set  $V$ . The singularities caused by the interactions of three waves produce artefacts that need be removed from the analysis. In this way, we see that the non-linear interaction of waves gives us the intersection of the light cone  $\mathcal{L}(q)$  and the observation domain  $V$ . The above-described microlocal point source  $S_{1234}$  can be produced at an arbitrary point  $q$  in the

future of the set  $V$ , and hence we can determine the earliest light observation sets  $\mathcal{E}_V(q)$  for any such point. After letting  $q$  vary in  $I(p^-, p^+)$ , we apply [Theorem 3.5](#) to recover the topology, differentiable structure, and the conformal class of  $g$  in  $I(p^-, p^+)$ .

## 5 Einstein-matter field equations

Einstein's equations for a Lorentzian metric  $g = (g_{jk})$  are

$$\text{Ein}(g) = T,$$

where  $\text{Ein}_{jk}(g) = \text{Ric}_{jk}(g) - \frac{1}{2}(g^{pq}\text{Ric}_{pq}(g))g_{jk}$ . Here  $\text{Ric}$  denotes the Ricci tensor,  $g^{-1} = (g^{pq})$  and  $T = (T_{jk})$  is the stress-energy tensor. In vacuum  $T = 0$ . Einstein's equations coupled with scalar fields  $\phi = (\phi_l)$ ,  $l = 1, 2, \dots, L$ , and a source  $\mathfrak{F} = (\mathfrak{F}^1, \mathfrak{F}^2)$  are

$$(21) \quad \text{Ein}(g) = T, \quad T = \mathbb{T}(g, \phi) + \mathfrak{F}^1,$$

$$(22) \quad \square_g \phi_l - \partial_{\phi_l} \mathcal{V}_l(x, \phi) = \mathfrak{F}_l^2, \quad l = 1, 2, \dots, L.$$

Here  $\mathfrak{F} = (\mathfrak{F}^1, \mathfrak{F}_1^2, \dots, \mathfrak{F}_L^2)$  models a source in active measurements, see [Section 5.1](#). The standard coupling  $\mathbb{T} = (\mathbb{T}_{jk})$  of  $g$  and  $\phi$  is given by

$$\mathbb{T}_{jk}(g, \phi) = \sum_{l=1}^L \left( \partial_j \phi_l \partial_k \phi_l - \frac{1}{2} g_{jk} g^{pq} \partial_p \phi_l \partial_q \phi_l - \mathcal{V}_l(x, \phi) g_{jk} \right),$$

the potentials  $\mathcal{V}_l$  are smooth functions  $M \times \mathbb{R}^L \rightarrow \mathbb{R}$ .

Below, we consider the case when  $M$  is 4-dimensional. We say that  $(M, \widehat{g})$  and  $\widehat{\phi}$  are the background spacetime and scalar fields if they are  $C^\infty$ -smooth, satisfy (21)–(22) with  $\mathfrak{F} = 0$  and  $(M, \widehat{g})$  is globally hyperbolic. Again, we write  $M$  in the form  $M = \mathbb{R} \times N$ . We will consider equations (21)–(22) with the initial conditions

$$(23) \quad g = \widehat{g}, \quad \phi = \widehat{\phi}, \quad \mathfrak{F} = 0, \quad \text{in } (-\infty, 0) \times N.$$

The source  $\mathfrak{F}$  can not be arbitrary since the Bianchi identities imply that  $\text{div}_g \text{Ein}(g) = 0$ , whence the stress energy tensor  $T$  needs to satisfy the conservation law

$$(24) \quad \text{div}_g T = 0.$$

This again implies the compatibility condition

$$(25) \quad \text{div}_g \mathfrak{F}^1 + \sum_{l=1}^L \mathfrak{F}_l^2 \nabla \phi_l = 0.$$

In local coordinates, the divergence is  $\operatorname{div}_g T = \nabla_p(g^{pj}T_{jk})$ ,  $k = 1, 2, 3, 4$ , where  $\nabla$  is the covariant derivative with respect to  $g$ . The conservation law (24) for Einstein's equations dictates, roughly speaking, that any source in the equation must take energy from some fields in order to increase energy in other fields.

Observe that in the system (21)–(22) the metric of the spacetime begins to change as soon as  $\mathcal{F}$  becomes non-zero, and that the system is invariant with respect to diffeomorphisms. We model an active measurement by factoring out the diffeomorphism invariance by using Fermi coordinates.

Let  $\widehat{g}$  and  $\widehat{\phi}$  be a background spacetime and scalar fields, and  $g$  be a close to  $\widehat{g}$ . We recall that  $M = \mathbb{R} \times N$ . Let  $p \in \{0\} \times N$ , and let  $\xi \in T_p M$  be time-like. Define  $\mu_g(s) = \gamma_{p,\xi}(s)$  to be the geodesic with respect to  $g$  satisfying  $\mu(0) = p$  and  $\dot{\mu}(0) = \xi$ . Let  $X_j$ ,  $j = 0, 1, 2, 3$ , be a basis of  $T_p M$ , with  $X_0 = \xi$ , and consider the following Fermi coordinates  $\Phi_g$ ,

$$\Phi_g(s, y^1, y^2, y^3) = \exp_{\mu_g(s)}(y^j Y_j), \quad \Phi_g : V \rightarrow M,$$

where  $Y_j$  is the parallel transport of  $X_j$  along  $\mu_g$ ,  $j = 1, 2, 3$ . Here the parallel transport and the exponential map  $\exp$  are with respect to  $g$ , and  $V = (0, 1) \times B$  where  $B$  is a ball centered at the origin in  $\mathbb{R}^3$ . We suppose that  $B$  is small enough so that the Fermi coordinates are well-defined with metric  $\widehat{g}$  in  $\overline{V}$ . Below, we denote the Fermi coordinates of  $(M, \widehat{g})$  by  $\Phi = \Phi_{\widehat{g}}$ .

Let  $t_0 > 0$  and consider a Lorentzian metric  $g$  on  $(-\infty, t_0) \times N$  such that the corresponding Fermi coordinates  $\Phi_g : V \rightarrow \mathbb{R} \times N$  are well-defined. We define the data set similar to (11),

$$\begin{aligned} \mathfrak{D} = \{(\Phi_g^* g|_V, \Phi_g^* \phi|_V, \Phi_g^* \mathcal{F}|_V) : (g, \phi, \mathcal{F}) \text{ satisfies (21),(22),(23)}, \\ \mathcal{F} \in C_0^k(\Phi_g(V)), \|\mathcal{F}\|_{C^k} < \varepsilon\}, \end{aligned}$$

where  $\Phi_g^*$  is the pullback under  $\Phi_g$ ,  $k$  is large enough, and  $\varepsilon > 0$  is small enough.

**Theorem 5.1** (Y. Kurylev, Lassas, Oksanen, and G. Uhlmann [2014]). *Let  $(M, g)$  be a globally hyperbolic 4-dimensional Lorentzian manifold. Let  $\mu_g^{\wedge}([0, 1])$  be a time-like geodesic and let  $p^- = \mu(0)$  and  $p^+ = \mu(1)$ . Suppose  $L \geq 4$ , and we have the non-degeneracy condition*

$$(26) \quad (\partial_j \widehat{\phi}_I)_{j,I=1}^4 \text{ is invertible at all points in } \overline{\Phi(V)}.$$

*Then the data set  $\mathfrak{D}$  determines the topology, differentiable structure and conformal class of the metric  $\widehat{g}$  in the double cone  $I(p^-, p^+)$  in  $(M, \widehat{g})$ .*

Analogous results for inverse problem for the Einstein-Maxwell system with vacuum background metric are considered in Lassas, G. Uhlmann, and Wang [2017].

**5.1 More on active measurements.** Recall that the source  $\mathfrak{F}$  must satisfy the compatibility condition (25). In particular, the set of allowed sources  $\mathfrak{F}$  depends on the solution  $(g, \phi)$  of the system (21)–(22). Due to this difficulty we use a construction that we call an adaptive source for the scalar fields. Consider the following special case of (21),(22),(23),

$$\begin{aligned} (27) \quad & \text{Ein}(g) = T, \quad T = F^1 + \mathbb{T}(g, \phi), \\ & \square_g \phi - \partial_\phi \mathcal{V}(\phi) = F^2 + \mathcal{S}(g, \phi, \nabla \phi, F, \nabla F), \\ & g = \widehat{g}, \quad \phi = \widehat{\phi}, \quad \text{in } (-\infty, 0) \times N. \end{aligned}$$

Here  $F = (F^1, F^2)$  are primary sources and  $\mathcal{S}(g, \phi, \nabla \phi, F, \nabla F)$  is the secondary source function that vanishes outside the support of the primary source  $F$  and adapts to values of the sources  $F$  and fields  $(g, \phi)$ . The secondary source functions can be considered as an abstract model for the measurement devices that one uses to implement the sources. When (26) is valid, functions  $\mathcal{S}$  can be constructed so that the conservation law (24) is valid for all sufficiently small  $F$  (see Y. Kurylev, Lassas, Oksanen, and G. Uhlmann [2014]).

## References

- Benjamin P Abbott et al. (2016). “Observation of gravitational waves from a binary black hole merger”. *Physical review letters* 116.6, p. 061102 (cit. on p. 3774).
- S. Alinhac (1983). “Non-unicité du problème de Cauchy”. *Ann. of Math. (2)* 117.1, pp. 77–108. MR: 683803 (cit. on p. 3772).
- Michael Anderson, Atsushi Katsuda, Yaroslav Kurylev, Matti Lassas, and Michael Taylor (2004). “Boundary regularity for the Ricci equation, geometric convergence, and Gelfand’s inverse boundary problem”. *Invent. Math.* 158.2, pp. 261–321. MR: 2096795 (cit. on p. 3772).
- Gang Bao and Hai Zhang (2014). “Sensitivity analysis of an inverse problem for the wave equation with caustics”. *J. Amer. Math. Soc.* 27.4, pp. 953–981. MR: 3230816 (cit. on p. 3772).
- Claude Bardos, Gilles Lebeau, and Jeffrey Rauch (1992). “Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary”. *SIAM J. Control Optim.* 30.5, pp. 1024–1065. MR: 1178650 (cit. on p. 3772).
- M. I. Belishev (1987). “An approach to multidimensional inverse problems for the wave equation”. *Dokl. Akad. Nauk SSSR* 297.3, pp. 524–527. MR: 924687.
- Michael I. Belishev and Yaroslav V. Kurylev (1992). “To the reconstruction of a Riemannian manifold via its spectral data (BC-method)”. *Comm. Partial Differential Equations* 17.5-6, pp. 767–804. MR: 1177292 (cit. on pp. 3771, 3772).

- Antonio N. Bernal and Miguel Sánchez (2005). “Smoothness of time functions and the metric splitting of globally hyperbolic spacetimes”. *Comm. Math. Phys.* 257.1, pp. 43–50. MR: [2163568](#) (cit. on p. [3775](#)).
- Kenrick Bingham, Yaroslav Kurylev, Matti Lassas, and Samuli Siltanen (2008). “Iterative time-reversal control for inverse problems”. *Inverse Probl. Imaging* 2.1, pp. 63–81. MR: [2375323](#) (cit. on pp. [3782](#), [3783](#)).
- A. S. Blagoveščenskiĭ (1969). “A one-dimensional inverse boundary value problem for a second order hyperbolic equation”. *Zap. Naučn. Sem. Leningrad. Otdel. Mat. Inst. Steklov. (LOMI)* 15, pp. 85–90. MR: [0282060](#) (cit. on p. [3780](#)).
- Jean-Michel Bony (1986). “Second microlocalization and propagation of singularities for semilinear hyperbolic equations”. In: *Hyperbolic equations and related topics (Katata/Kyoto, 1984)*. Academic Press, Boston, MA, pp. 11–49. MR: [925240](#) (cit. on p. [3776](#)).
- Roberta Bosi, Yaroslav Kurylev, and Matti Lassas (2016). “Stability of the unique continuation for the wave operator via Tataru inequality and applications”. *J. Differential Equations* 260.8, pp. 6451–6492. MR: [3460220](#) (cit. on p. [3780](#)).
- (Feb. 2017). “Reconstruction and stability in Gel’fand’s inverse interior spectral problem”. arXiv: [1702.07937](#) (cit. on p. [3772](#)).
- Yvonne Choquet-Bruhat (2009). *General relativity and the Einstein equations*. Oxford Mathematical Monographs. Oxford University Press, Oxford, pp. xxvi+785. MR: [2473363](#).
- Matias F. Dahl, Anna Kirpichnikova, and Matti Lassas (2009). “Focusing waves in unknown media by modified time reversal iteration”. *SIAM J. Control Optim.* 48.2, pp. 839–858. MR: [2486096](#) (cit. on pp. [3782](#), [3783](#)).
- G. Eskin (2017). “Inverse problems for general second order hyperbolic equations with time-dependent coefficients”. *Bull. Math. Sci.* 7.2, pp. 247–307. arXiv: [1503.00825](#). MR: [3671738](#) (cit. on p. [3772](#)).
- I. M. Gelfand (1954). “Some aspects of functional analysis and algebra”. In: *Proceedings of the International Congress of Mathematicians, Amsterdam*. Vol. 1, pp. 253–276 (cit. on p. [3770](#)).
- Allan Greenleaf and Gunther Uhlmann (1993). “Recovering singularities of a potential from singularities of scattering data”. *Comm. Math. Phys.* 157.3, pp. 549–572. MR: [1243710](#) (cit. on p. [3784](#)).
- Tapio Helin, Matti Lassas, and Lauri Oksanen (2014). “Inverse problem for the wave equation with a white noise source”. *Comm. Math. Phys.* 332.3, pp. 933–953. MR: [3262617](#) (cit. on p. [3777](#)).
- Tapio Helin, Matti Lassas, Lauri Oksanen, and Teemu Saksala (2016). “Correlation based passive imaging with a white noise source”. To appear in *J. Math. Pures et Appl.* arXiv: [1609.08022](#) (cit. on p. [3777](#)).
- Linda M. Holt (1995). “Singularities produced in conormal wave interactions”. *Trans. Amer. Math. Soc.* 347.1, pp. 289–315. MR: [1264146](#) (cit. on p. [3776](#)).

- Maarten V. de Hoop, Paul Kepley, and Lauri Oksanen (2016). “On the construction of virtual interior point source travel time distances from the hyperbolic Neumann-to-Dirichlet map”. *SIAM J. Appl. Math.* 76.2, pp. 805–825. MR: [3488169](#) (cit. on pp. [3781](#)–[3783](#)).
- Peter R Hoskins (2012). “Principles of ultrasound elastography”. *Ultrasound* 20.1, pp. 8–15.
- Thomas J. R. Hughes, Tosio Kato, and Jerrold E. Marsden (1976). “Well-posed quasi-linear second-order hyperbolic systems with applications to nonlinear elastodynamics and general relativity”. *Arch. Rational Mech. Anal.* 63.3, 273–294 (1977). MR: [0420024](#).
- V. Isakov (1993). “On uniqueness in inverse problems for semilinear parabolic equations”. *Arch. Rational Mech. Anal.* 124.1, pp. 1–12. MR: [1233645](#) (cit. on p. [3773](#)).
- Victor Isakov and Adrian I. Nachman (1995). “Global uniqueness for a two-dimensional semilinear elliptic inverse problem”. *Trans. Amer. Math. Soc.* 347.9, pp. 3375–3390. MR: [1311909](#).
- Kyeonbae Kang and Gen Nakamura (2002). “Identification of nonlinearity in a conductivity equation via the Dirichlet-to-Neumann map”. *Inverse Problems* 18.4, pp. 1079–1088. MR: [1929283](#).
- A. Katchalov, Y. Kurylev, M. Lassas, and N. Mandache (2004). “Equivalence of time-domain inverse problems and boundary spectral problems”. *Inverse Problems* 20.2, pp. 419–436. MR: [2065431](#) (cit. on p. [3771](#)).
- Alexander Katchalov, Yaroslav Kurylev, and Matti Lassas (2001). *Inverse boundary spectral problems*. Vol. 123. Chapman & Hall/CRC Monographs and Surveys in Pure and Applied Mathematics. Chapman & Hall/CRC, Boca Raton, FL, pp. xx+290. MR: [1889089](#) (cit. on pp. [3772](#), [3780](#), [3783](#)).
- Y. Kurylev, L. Oksanen, and G. Paternain (n.d.). “Inverse problems for the connection Laplacian”. To appear in *J. Diff. Geom.* (cit. on p. [3772](#)).
- Yaroslav Kurylev and Matti Lassas (2000). “Gelfand inverse problem for a quadratic operator pencil”. *J. Funct. Anal.* 176.2, pp. 247–263. MR: [1784415](#) (cit. on p. [3773](#)).
- Yaroslav Kurylev, Matti Lassas, Lauri Oksanen, and Gunther Uhlmann (May 2014). “Inverse problem for Einstein-scalar field equations”. arXiv: [1406.4776](#) (cit. on pp. [3786](#), [3787](#)).
- Yaroslav Kurylev, Matti Lassas, and Erkki Somersalo (2006). “Maxwell’s equations with a polarization independent wave velocity: direct and inverse problems”. *J. Math. Pures Appl. (9)* 86.3, pp. 237–270. MR: [2257731](#) (cit. on p. [3773](#)).
- Yaroslav Kurylev, Matti Lassas, and Gunther Uhlmann (2014). “Inverse problems for Lorentzian manifolds and non-linear hyperbolic equations”. To appear in *Inventiones Math.* arXiv: [1405.3386](#) (cit. on pp. [3775](#), [3778](#)).

- M. Lassas, G. Uhlmann, and Y. Wang (n.d.). “Inverse problems for semilinear wave equations on Lorentzian manifolds”. To appear in *Comm. Math. Phys.* (cit. on p. 3777).
- Matti Lassas and Lauri Oksanen (2014). “Inverse problem for the Riemannian wave equation with Dirichlet data and Neumann data on disjoint sets”. *Duke Math. J.* 163.6, pp. 1071–1103. MR: 3192525 (cit. on p. 3773).
- Matti Lassas and Teemu Saksala (2015). “Determination of a Riemannian manifold from the distance difference functions”. To appear in *Asian J. Math.* arXiv: 1510.06157 (cit. on p. 3779).
- Matti Lassas, Gunther Uhlmann, and Yiran Wang (Mar. 2017). “Determination of vacuum space-times from the Einstein-Maxwell equations”. arXiv: 1703.10704 (cit. on p. 3786).
- Camille Laurent and Matthieu Léautaud (2015). “Quantitative unique continuation for operators with partially analytic coefficients. Application to approximate control for waves”. To appear in *Journal of EMS*. arXiv: 1506.04254 (cit. on p. 3780).
- R. B. Melrose and G. A. Uhlmann (1979). “Lagrangian intersection and the Cauchy problem”. *Comm. Pure Appl. Math.* 32.4, pp. 483–519. MR: 528633 (cit. on p. 3784).
- Richard Melrose and Niles Ritter (1985). “Interaction of nonlinear progressing waves for semilinear wave equations”. *Ann. of Math. (2)* 121.1, pp. 187–213. MR: 782559 (cit. on p. 3776).
- Jennifer L. Mueller and Samuli Siltanen (2012). *Linear and nonlinear inverse problems with practical applications*. Vol. 10. Computational Science & Engineering. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, pp. xiv+351. MR: 2986262 (cit. on p. 3781).
- Adrian Nachman, John Sylvester, and Gunther Uhlmann (1988). “An  $n$ -dimensional Borg-Levinson theorem”. *Comm. Math. Phys.* 115.4, pp. 595–605. MR: 933457 (cit. on p. 3772).
- Gen Nakamura and Michiyuki Watanabe (2008). “An inverse boundary value problem for a nonlinear wave equation”. *Inverse Probl. Imaging* 2.1, pp. 121–131. MR: 2375325.
- Lauri Oksanen (2013). “Inverse obstacle problem for the non-stationary wave equation with an unknown background”. *Comm. Partial Differential Equations* 38.9, pp. 1492–1518. MR: 3169753 (cit. on p. 3772).
- Jonathan Ophir, S Kaisar Alam, Brian Garra, F Kallel, E Konofagou, T Krouskop, and T Varghese (1999). “Elastography: ultrasonic estimation and imaging of the elastic properties of tissues”. *Proceedings of the Institution of Mechanical Engineers, Part H: Journal of Engineering in Medicine* 213.3, pp. 203–233.
- Mikko Salo and Xiao Zhong (2012). “An inverse problem for the  $p$ -Laplacian: boundary determination”. *SIAM J. Math. Anal.* 44.4, pp. 2474–2495. MR: 3023384 (cit. on p. 3773).

- Plamen D. Stefanov (1989). “Uniqueness of the multi-dimensional inverse scattering problem for time dependent potentials”. *Math. Z.* 201.4, pp. 541–559. MR: [1004174](#).
- Plamen Stefanov and Gunther Uhlmann (2005). “Stable determination of generic simple metrics from the hyperbolic Dirichlet-to-Neumann map”. *Int. Math. Res. Not.* 17, pp. 1047–1061. MR: [2145709](#) (cit. on p. [3772](#)).
- Ziqi Sun and Gunther Uhlmann (1997). “Inverse problems in quasilinear anisotropic media”. *Amer. J. Math.* 119.4, pp. 771–797. MR: [1465069](#) (cit. on p. [3773](#)).
- John Sylvester and Gunther Uhlmann (1987). “A global uniqueness theorem for an inverse boundary value problem”. *Ann. of Math. (2)* 125.1, pp. 153–169. MR: [873380](#) (cit. on p. [3772](#)).
- Daniel Tataru (1995). “Unique continuation for solutions to PDE’s; between Hörmander’s theorem and Holmgren’s theorem”. *Comm. Partial Differential Equations* 20.5-6, pp. 855–884. MR: [1326909](#) (cit. on pp. [3771](#), [3772](#), [3780](#)).
- Yiran Wang and Ting Zhou (2016). “Inverse problems for quadratic derivative nonlinear wave equations”. To appear in *Comm. PDE*. arXiv: [1612.04437](#) (cit. on p. [3777](#)).

Received 2017-11-24.

MATTI LASSAS  
UNIVERSITY OF HELSINKI  
FINLAND  
[matti.lassas@helsinki.fi](mailto:matti.lassas@helsinki.fi)





# THE MOMENT-SOS HIERARCHY

JEAN B. LASSERRE

## Abstract

The Moment-SOS hierarchy initially introduced in optimization in 2000, is based on the theory of the **K**-moment problem and its dual counterpart, polynomials that are positive on **K**. It turns out that this methodology can be also applied to solve problems with positivity constraints “ $f(\mathbf{x}) \geq 0$  for all  $\mathbf{x} \in \mathbf{K}$ ” and/or linear constraints on Borel measures. Such problems can be viewed as specific instances of the “Generalized Problem of Moments” (GPM) whose list of important applications in various domains is endless. We describe this methodology and outline some of its applications in various domains.

## 1 Introduction

Consider the optimization problem:

$$(1-1) \quad \mathbf{P} : f^* = \inf_{\mathbf{x}} \{ f(\mathbf{x}) : \mathbf{x} \in \Omega \},$$

where  $f$  is a polynomial and  $\Omega \subset \mathbb{R}^n$  is a basic semi-algebraic set, that is,

$$(1-2) \quad \Omega := \{ \mathbf{x} \in \mathbb{R}^n : g_j(\mathbf{x}) \geq 0, \quad j = 1, \dots, m \},$$

for some polynomials  $g_j$ ,  $j = 1, \dots, m$ . Problem **P** is a particular case of *Non Linear Programming* (NLP) where the data  $(f, g_j, j = 1, \dots, m)$  are *algebraic*, and therefore the whole arsenal of methods of NLP can be used for solving **P**. So what is so specific about **P** in [Equation \(1-1\)](#)? The answer depends on the meaning of  $f^*$  in [Equation \(1-1\)](#).

If one is interested in a *local minimum* only then efficient NLP methods can be used for solving **P**. In such methods, the fact that  $f$  and  $g_j$ ’s are polynomials does not help

---

Research supported by the European Research Council (ERC) through ERC-Advanced Grant # 666981 for the TAMING project.

MSC2010: primary 90C26; secondary 90C22, 90C27, 65K05, 14P10, 44A60.

Keywords: K-Moment problem, positive polynomials, global optimization, semidefinite relaxations.

much, that is, this algebraic feature of  $\mathbf{P}$  is not really exploited. On the other hand if  $f^*$  in Equation (1-1) is understood as the *global minimum* of  $\mathbf{P}$  then the picture is totally different. Why? First, to eliminate any ambiguity on the meaning of  $f^*$  in Equation (1-1), rewrite Equation (1-1) as:

$$(1-3) \quad \mathbf{P} : f^* = \sup \{ \lambda : f(\mathbf{x}) - \lambda \geq 0, \quad \forall \mathbf{x} \in \Omega \}$$

because then indeed  $f^*$  is necessarily the global minimum of  $\mathbf{P}$ .

In full generality, most problems Equation (1-3) are very difficult to solve (they are labelled NP-hard in the computational complexity terminology) because:

*Given  $\lambda \in \mathbb{R}$ , checking whether “ $f(\mathbf{x}) - \lambda \geq 0$  for all  $\mathbf{x} \in \Omega$ ” is difficult.*

Indeed, by nature this positivity constraint is *global* and therefore cannot be handled by standard NLP optimization algorithms which use only local information around a current iterate  $\mathbf{x} \in \Omega$ . Therefore to compute  $f^*$  in Equation (1-3) one needs an efficient tool to handle the positivity constraint “ $f(\mathbf{x}) - \lambda \geq 0$  for all  $\mathbf{x} \in \Omega$ ”. Fortunately if the data are algebraic then:

1. Powerful *positivity certificates* from Real Algebraic Geometry (*Positivstellensätze* in german) are available.
2. Some of these positivity certificates have an efficient practical implementation via *Linear Programming* (LP) or *Semidefinite Programming* (SDP). In particular and importantly, testing whether a given polynomial is a sum of squares (SOS) simply reduces to solving a single SDP (which can be done in time polynomial in the input size of the polynomial, up to arbitrary fixed precision).

After the pioneers works of Shor [1998] and Nesterov [2000], Lasserre [2000, 2000/01] and Parrilo [2000, 2003] have been the first to provide a systematic use of these two key ingredients in Control and Optimization, with convergence guarantees. It is also worth mentioning another closely related pioneer work, namely the celebrated SDP-relaxation of Goemans and Williamson [1995] which provides a 0.878 approximation guarantee for MAXCUT, a famous problem in non-convex combinatorial optimization (and probably the simplest one). In fact it is perhaps the first famous example of such a successful application of the powerful SDP convex optimization technique to provide guaranteed good approximations to a notoriously difficult non-convex optimization problem. It turns out that this SDP relaxation is the first relaxation in the Moment-SOS hierarchy (a.k.a. Lasserre hierarchy) when applied to the MAXCUT problem. Since then, this spectacular success story of SDP relaxations has been at the origin of a flourishing research activity in combinatorial optimization and computational complexity. In particular, the study of

LP- and SDP-relaxations in hardness of approximation is at the core of a central topic in combinatorial optimization and computational complexity, namely proving/disproving Khot's famous Unique Games Conjecture<sup>1</sup> (UGC) in Theoretical Computer Science.

Finally, another “definition” of the global optimum  $f^*$  of  $\mathbf{P}$  reads:

$$(1-4) \quad f^* = \inf_{\mu} \left\{ \int_{\Omega} f \, d\mu : \mu(\Omega) = 1 \right\}$$

where the ‘inf’ is over all probability measures on  $\Omega$ . Equivalently, writing  $f$  as  $\sum_{\alpha} f_{\alpha} \mathbf{x}^{\alpha}$  in the basis of monomials (where  $\mathbf{x}^{\alpha} = x_1^{\alpha_1} \cdots x_n^{\alpha_n}$ ):

$$(1-5) \quad f^* = \inf_y \left\{ \sum_{\alpha} f_{\alpha} y_{\alpha} : \mathbf{y} \in \mathcal{M}(\Omega); \quad y_0 = 1 \right\},$$

where  $\mathcal{M}(\Omega) = \{\mathbf{y} = (y_{\alpha})_{\alpha \in \mathbb{N}^n} : \exists \mu \text{ s.t. } y_{\alpha} = \int_{\Omega} \mathbf{x}^{\alpha} \, d\mu, \forall \alpha \in \mathbb{N}^n\}$ , a convex cone. In fact Equation (1-3) is the LP dual of Equation (1-4). In other words standard LP duality between the two formulations Equation (1-4) and Equation (1-3) illustrates the duality between the “ $\Omega$ -moment problem” and “polynomials positive on  $\Omega$ ”.

Problem (1-4) is a very particular instance (and even the simplest instance) of the more general *Generalized Problem of Moments* (GPM):

$$(1-6) \quad \inf_{\mu_1, \dots, \mu_p} \left\{ \sum_{j=1}^p \int_{\Omega_j} f_j \, d\mu_j : \sum_{j=1}^p f_{ij} \, d\mu_j \geq b_i, \, i = 1, \dots, s \right\},$$

for some functions  $f_{ij} : \mathbb{R}^{n_j} \rightarrow \mathbb{R}, i = 1, \dots, s$ , and sets  $\Omega_j \subset \mathbb{R}^{n_j}, j = 1, \dots, p$ . The GPM is an infinite-dimensional LP with dual:

$$(1-7) \quad \sup_{\lambda_1, \dots, \lambda_s \geq 0} \left\{ \sum_{i=1}^s \lambda_i b_i : f_j - \sum_{i=1}^s \lambda_i f_{ij} \geq 0 \text{ on } \Omega_j, \, j = 1, \dots, p \right\}.$$

Therefore it should be of no surprise that the Moment-SOS hierarchy, initially developed for global optimization, also applies to solving the GPM. This is particularly interesting as the list of important applications of the GPM is almost endless; see e.g. Landau [1987].

## 2 The MOMENT-SOS hierarchy in optimization

**2.1 Notation, definitions and preliminaries.** Let  $\mathbb{R}[\mathbf{x}]$  denote the ring of polynomials in the variables  $\mathbf{x} = (x_1, \dots, x_n)$  and let  $\mathbb{R}[\mathbf{x}]_d$  be the vector space of polynomials of

<sup>1</sup>For this conjecture and its theoretical and practical implications, S. Khot was awarded the prestigious Nevanlinna prize at the last ICM 2014 in Seoul Khot [2014].

degree at most  $d$  (whose dimension is  $s(d) := \binom{n+d}{n}$ ). For every  $d \in \mathbb{N}$ , let  $\mathbb{N}_d^n := \{\alpha \in \mathbb{N}^n : |\alpha| (= \sum_i \alpha_i) \leq d\}$ , and let  $\mathbf{v}_d(\mathbf{x}) = (\mathbf{x}^\alpha)$ ,  $\alpha \in \mathbb{N}_d^n$ , be the vector of monomials of the canonical basis  $(\mathbf{x}^\alpha)$  of  $\mathbb{R}[\mathbf{x}]_d$ . Given a closed set  $\mathcal{X} \subseteq \mathbb{R}^n$ , let  $\mathcal{P}(\mathcal{X}) \subset \mathbb{R}[\mathbf{x}]$  (resp.  $\mathcal{P}_d(\mathcal{X}) \subset \mathbb{R}[\mathbf{x}]_d$ ) be the convex cone of polynomials (resp. polynomials of degree at most  $d$ ) that are nonnegative on  $\mathcal{X}$ . A polynomial  $f \in \mathbb{R}[\mathbf{x}]_d$  is written

$$\mathbf{x} \mapsto f(\mathbf{x}) = \sum_{\alpha \in \mathbb{N}^n} f_\alpha \mathbf{x}^\alpha,$$

with vector of coefficients  $\mathbf{f} = (f_\alpha) \in \mathbb{R}^{s(d)}$  in the canonical basis of monomials  $(\mathbf{x}^\alpha)_{\alpha \in \mathbb{N}^n}$ . For real symmetric matrices, let  $(\mathbf{B}, \mathbf{C}) := \text{trace}(\mathbf{B}\mathbf{C})$  while the notation  $\mathbf{B} \succeq 0$  stands for  $\mathbf{B}$  is positive semidefinite (psd) whereas  $\mathbf{B} \succ 0$  stands for  $\mathbf{B}$  is positive definite (pd).

**The Riesz functional.** Given a sequence  $\mathbf{y} = (y_\alpha)_{\alpha \in \mathbb{N}^n}$ , the Riesz functional is the linear mapping  $L_y : \mathbb{R}[\mathbf{x}] \rightarrow \mathbb{R}$  defined by:

$$(2-1) \quad f = \left( \sum_{\alpha} f_{\alpha} \mathbf{x}^{\alpha} \right) \mapsto L_y(f) = \sum_{\alpha \in \mathbb{N}^n} f_{\alpha} y_{\alpha}.$$

**Moment matrix.** The *moment matrix* associated with a sequence  $\mathbf{y} = (y_\alpha)$ ,  $\alpha \in \mathbb{N}^n$ , is the real symmetric matrix  $\mathbf{M}_d(\mathbf{y})$  with rows and columns indexed by  $\mathbb{N}_d^n$ , and whose entry  $(\alpha, \beta)$  is just  $y_{\alpha+\beta}$ , for every  $\alpha, \beta \in \mathbb{N}_d^n$ . Alternatively, let  $\mathbf{v}_d(\mathbf{x}) \in \mathbb{R}^{s(d)}$  be the vector  $(\mathbf{x}^\alpha)$ ,  $\alpha \in \mathbb{N}_d^n$ , and define the matrices  $(\mathbf{B}_{o,\alpha}) \subset \mathcal{S}^{s(d)}$  by

$$(2-2) \quad \mathbf{v}_d(\mathbf{x}) \mathbf{v}_d(\mathbf{x})^T = \sum_{\alpha \in \mathbb{N}_{2d}^n} \mathbf{B}_{o,\alpha} \mathbf{x}^\alpha, \quad \forall \mathbf{x} \in \mathbb{R}^n.$$

Then  $\mathbf{M}_d(\mathbf{y}) = \sum_{\alpha \in \mathbb{N}_{2d}^n} y_\alpha \mathbf{B}_{o,\alpha}$ . If  $\mathbf{y}$  has a representing measure  $\mu$  then  $\mathbf{M}_d(\mathbf{y}) \succeq 0$  because  $\langle \mathbf{f}, \mathbf{M}_d(\mathbf{y}) \mathbf{f} \rangle = \int f^2 d\mu \geq 0$ , for all  $f \in \mathbb{R}[\mathbf{x}]_d$ .

A measure whose all moments are finite, is *moment determinate* if there is no other measure with same moments. The support of a Borel measure  $\mu$  on  $\mathbb{R}^n$  (denoted  $\text{supp}(\mu)$ ) is the smallest closed set  $\Omega$  such that  $\mu(\mathbb{R}^n \setminus \Omega) = 0$ .

**Localizing matrix.** With  $\mathbf{y}$  as above and  $g \in \mathbb{R}[\mathbf{x}]$  (with  $g(\mathbf{x}) = \sum_{\gamma} g_{\gamma} \mathbf{x}^{\gamma}$ ), the *localizing matrix* associated with  $\mathbf{y}$  and  $g$  is the real symmetric matrix  $\mathbf{M}_d(g, \mathbf{y})$  with rows and columns indexed by  $\mathbb{N}_d^n$ , and whose entry  $(\alpha, \beta)$  is just  $\sum_{\gamma} g_{\gamma} y_{\alpha+\beta+\gamma}$ , for every  $\alpha, \beta \in \mathbb{N}_d^n$ . Alternatively, let  $\mathbf{B}_{g,\alpha} \in \mathcal{S}^{s(d)}$  be defined by:

$$(2-3) \quad g(\mathbf{x}) \mathbf{v}_d(\mathbf{x}) \mathbf{v}_d(\mathbf{x})^T = \sum_{\alpha \in \mathbb{N}_{2d+\deg g}^n} \mathbf{B}_{g,\alpha} \mathbf{x}^\alpha, \quad \forall \mathbf{x} \in \mathbb{R}^n.$$

Then  $\mathbf{M}_d(g \mathbf{y}) = \sum_{\alpha \in \mathbb{N}_{2d+\deg g}^n} y_\alpha \mathbf{B}_{g,\alpha}$ . If  $\mathbf{y}$  has a representing measure  $\mu$  whose support is contained in the set  $\{\mathbf{x} : g(\mathbf{x}) \geq 0\}$  then  $\mathbf{M}_d(g \mathbf{y}) \succeq 0$  for all  $d$  because  $\langle \mathbf{f}, \mathbf{M}_d(g \mathbf{y}) \mathbf{f} \rangle = \int f^2 g d\mu \geq 0$ , for all  $f \in \mathbb{R}[\mathbf{x}]_d$ .

**SOS polynomials and quadratic modules.** A polynomial  $f \in \mathbb{R}[\mathbf{x}]$  is a Sum-of-Squares (SOS) if there exist  $(f_k)_{k=1,\dots,s} \subset \mathbb{R}[\mathbf{x}]$ , such that  $f(\mathbf{x}) = \sum_{k=1}^s f_k(\mathbf{x})^2$ , for all  $\mathbf{x} \in \mathbb{R}^n$ . Denote by  $\Sigma[\mathbf{x}]$  (resp.  $\Sigma[\mathbf{x}]_d$ ) the set of SOS polynomials (resp. SOS polynomials of degree at most  $2d$ ). Of course every SOS polynomial is nonnegative whereas the converse is not true. In addition, checking whether a given polynomial  $f$  is nonnegative on  $\mathbb{R}^n$  is difficult whereas checking whether  $f$  is SOS is much easier and can be done efficiently. Indeed let  $f \in \mathbb{R}[\mathbf{x}]_{2d}$  (for  $f$  to be SOS its degree must be even),  $\mathbf{x} \mapsto f(\mathbf{x}) = \sum_{\alpha \in \mathbb{N}_{2d}^n} f_\alpha \mathbf{x}^\alpha$ . Then  $f$  is SOS if and only if there exists a real symmetric matrix  $\mathbf{X}^T = \mathbf{X}$  of size  $s(d) = \binom{n+d}{n}$ , such that:

$$(2-4) \quad \mathbf{X} \succeq 0; \quad f_\alpha = \langle \mathbf{X}, \mathbf{B}_{o,\alpha} \rangle, \quad \forall \alpha \in \mathbb{N}_{2d}^n,$$

and this can be checked by solving an SDP.

Next, let  $\mathbf{x} \mapsto g_0(\mathbf{x}) := 1$  for all  $\mathbf{x} \in \mathbb{R}^n$ . With a family  $(g_1, \dots, g_m) \subset \mathbb{R}[\mathbf{x}]$  is associated the *quadratic module*  $Q(g) (= Q(g_1, \dots, g_m)) \subset \mathbb{R}[\mathbf{x}]$ :

$$(2-5) \quad Q(g) := \left\{ \sum_{j=0}^m \sigma_j g_j : \sigma_j \in \Sigma[\mathbf{x}], j = 0, \dots, m \right\},$$

and its *truncated* version

$$(2-6) \quad Q_k(g) := \left\{ \sum_{j=0}^m \sigma_j g_j : \sigma_j \in \Sigma[\mathbf{x}]_{k-d_j}, j = 0, \dots, m \right\},$$

where  $d_j = \lceil \deg(g_j)/2 \rceil$ ,  $j = 0, \dots, m$ .

**Definition 1.** The quadratic module  $Q(g)$  associated with  $\Omega$  in Equation (1-2) is said to be *Archimedean* if there exists  $M > 0$  such that the quadratic polynomial  $\mathbf{x} \mapsto M - \|\mathbf{x}\|^2$  belongs to  $Q(g)$  (i.e., belongs to  $Q_k(g)$  for some  $k$ ).

If  $Q(g)$  is Archimedean then necessarily  $\Omega$  is compact but the reverse is not true. The Archimedean condition (which depends on the representation of  $\Omega$ ) can be seen as an *algebraic certificate* that  $\Omega$  is compact. For more details on the above notions of moment and localizing matrix, quadratic module, as well as their use in potential applications, the interested reader is referred to Lasserre [2010], Laurent [2009], Schmüdgen [2017].

**2.2 Two certificates of positivity (Positivstellensätze).** Below we describe two particular certificates of positivity which are important because they provide the theoretical justification behind the so-called SDP- and LP-relaxations for global optimization.

**Theorem 2.1 (Putinar [1993]).** *Let  $\Omega \subset \mathbb{R}^n$  be as in Equation (1-2) and assume that  $Q(g)$  is Archimedean.*

(a) *If a polynomial  $f \in \mathbb{R}[\mathbf{x}]$  is (strictly) positive on  $\Omega$  then  $f \in Q(g)$ .*

(b) *A sequence  $\mathbf{y} = (y_\alpha)_{\alpha \in \mathbb{N}^n} \subset \mathbb{R}$  has a representing Borel measure on  $\Omega$  if and only if  $L_{\mathbf{y}}(f^2 g_j) \geq 0$  for all  $f \in \mathbb{R}[\mathbf{x}]$ , and all  $j = 0, \dots, m$ . Equivalently, if and only if  $\mathbf{M}_d(\mathbf{y} g_j) \geq 0$  for all  $j = 0, \dots, m$ ,  $d \in \mathbb{N}$ .*

There exists another certificate of positivity which does not use SOS.

**Theorem 2.2 (Krivine [1964a], Krivine [1964b], and Vasilescu [2003]).** *Let  $\Omega \subset \mathbb{R}^n$  as in Equation (1-2) be compact and such that (possibly after scaling)  $0 \leq g_j(\mathbf{x}) \leq 1$  for all  $\mathbf{x} \in \Omega$ ,  $j = 1, \dots, m$ . Assume also that  $[1, g_1, \dots, g_m]$  generates  $\mathbb{R}[\mathbf{x}]$ .*

(a) *If a polynomial  $f \in \mathbb{R}[\mathbf{x}]$  is (strictly) positive on  $\Omega$  then*

$$(2-7) \quad f(\mathbf{x}) = \sum_{\alpha, \beta \in \mathbb{N}^n} c_{\alpha, \beta} \prod_{j=1}^m g_j(\mathbf{x})^{\alpha_j} (1 - g_j(\mathbf{x}))^{\beta_j},$$

*for finitely many positive coefficients  $(c_{\alpha, \beta})_{\alpha, \beta \in \mathbb{N}^m}$ .*

(b) *A sequence  $\mathbf{y} = (y_\alpha)_{\alpha \in \mathbb{N}^n} \subset \mathbb{R}$  has a representing Borel measure on  $\Omega$  if and only*

$$\text{if } L_{\mathbf{y}} \left( \prod_{j=1}^m g_j(\mathbf{x})^{\alpha_j} (1 - g_j(\mathbf{x}))^{\beta_j} \right) \geq 0 \text{ for all } \alpha, \beta \in \mathbb{N}^m.$$

The two facets (a) and (b) of Theorem 2.1 and Theorem 2.2 illustrate the duality between *polynomials positive on  $\Omega$*  (in (a)) and the  *$\Omega$ -moment problem* (in (b)). In addition to their mathematical interest, both Theorem 2.1(a) and Theorem 2.2(a) have another distinguishing feature. They both have a practical implementation. Testing whether  $f \in \mathbb{R}[\mathbf{x}]_d$  is in  $Q(g)_k$  is just solving a single SDP, whereas testing whether  $f$  can be written as in Equation (2-7) with  $\sum_{i=1}^m \alpha_i + \beta_i \leq k$ , is just solving a single Linear Program (LP).

**2.3 The Moment-SOS hierarchy.** The Moment-SOS hierarchy is a numerical scheme based on Putinar's theorem. In a nutshell it consists of replacing the intractable positivity constraint " $f(\mathbf{x}) \geq 0$  for all  $\mathbf{x} \in \Omega$ " with Putinar's positivity certificate  $f \in Q_d(g)$  of Theorem 2.1(a), i.e., with a fixed degree bound on the SOS weights  $(\sigma_j)$  in Equation (2-6). By duality, it consists of replacing the intractable constraint  $\mathbf{y} \in \mathcal{M}(\Omega)$  with the necessary conditions  $\mathbf{M}_d(g_j \mathbf{y}) \geq 0$ ,  $j = 0, \dots, m$ , of Theorem 2.1(b) for a fixed  $d$ . This results in solving an SDP which provides a lower bound on the global minimum. By allowing the

degree bound  $d$  to increase, one obtains a *hierarchy* of SDPs (of increasing size) which provides a monotone non-decreasing sequence of lower bounds. A similar strategy based on Krivine-Stengle-Vasilescu positivity certificate (Equation (2-7)) is also possible and yields a hierarchy of LP (instead of SDPs). However even though one would prefer to solve LPs rather than SDPs, the latter Moment-LP hierarchy has several serious drawbacks (some explained in e.g. Lasserre [2015a, 2002b]), and therefore we only describe the Moment-SOS hierarchy.

Recall problem  $\mathbf{P}$  in Equation (1-1) or equivalently in Equation (1-3) and Equation (1-4), where  $\Omega \subset \mathbb{R}^n$  is the basic semi-algebraic set defined in Equation (1-2).

**The Moment-SOS hierarchy.** Consider the sequence of semidefinite programs  $(\mathbf{Q}_d)_{d \in \mathbb{N}}$  with  $d \geq \hat{d} := \max[\deg(f), \max_j \deg(g_j)]$ :

$$(2-8) \quad \mathbf{Q}_d : \rho_d = \inf_{\mathbf{y}} \{ L_{\mathbf{y}}(f) : y_0 = 1; \mathbf{M}_{d-d_j}(g_j \mathbf{y}) \succeq 0, \quad 0 \leq j \leq m \}$$

(where  $\mathbf{y} = (y_\alpha)_{\alpha \in \mathbb{N}_{2d}^n}$ )<sup>2</sup>, with associated sequence of their SDP duals:

$$(2-9) \quad \mathbf{Q}_d^* : \rho_d^* = \sup_{\lambda, \sigma_j} \{ \lambda : f - \lambda = \sum_{j=0}^m \sigma_j g_j; \sigma_j \in \Sigma[\mathbf{x}]_{d-d_j}, \quad 0 \leq j \leq m \}$$

(where  $d_j = \lceil (\deg g_j)/2 \rceil$ ). By standard weak duality in optimization  $\rho_d^* \leq \rho_d$  for every  $d \geq \hat{d}$ . The sequence  $(\mathbf{Q}_d)_{d \in \mathbb{N}}$  forms a *hierarchy* of *SDP-relaxations* of  $\mathbf{P}$  because  $\rho_d \leq f^*$  and  $\rho_d \leq \rho_{d+1}$  for all  $d \geq \hat{d}$ . Indeed for each  $d \geq \hat{d}$ , the constraints of  $\mathbf{Q}_d$  consider only necessary conditions for  $\mathbf{y}$  to be the moment sequence (up to order  $2d$ ) of a probability measure on  $\Omega$  (cf. Theorem 2.1(b)) and therefore  $\mathbf{Q}_d$  is a relaxation of Equation (1-5).

By duality, the sequence  $(\mathbf{Q}_d^*)_{d \in \mathbb{N}}$  forms a *hierarchy* of *SDP-strengthenings* of Equation (1-3). Indeed in Equation (2-9) one has replaced the intractable positivity constraint of Equation (1-3) by the (stronger) Putinar's positivity certificate with degree bound  $2d - 2d_j$  on the SOS weights  $\sigma_j$ 's.

**Theorem 2.3** (Lasserre [2000, 2000/01]). *Let  $\Omega$  in Equation (1-2) be compact and assume that its associated quadratic module  $\mathcal{Q}(g)$  is Archimedean. Then:*

(i) *As  $d \rightarrow \infty$ , the monotone non-decreasing sequence  $(\rho_d)_{d \in \mathbb{N}}$  (resp.  $(\rho_d^*)_{d \in \mathbb{N}}$ ) of optimal values of the hierarchy (Equation (2-8)) (resp. Equation (2-9)) converges to the global optimum  $f^*$  of  $\mathbf{P}$ .*

<sup>2</sup>In Theoretical Computer Science,  $\mathbf{y}$  is called a sequence of “pseudo-moments”.



(ii) Moreover, let  $\mathbf{y}^d = (y_\alpha^d)_{\alpha \in \mathbb{N}_{2d}^n}$  be an optimal solution of  $\mathbf{Q}_d$  in Equation (2-8), and let  $s = \max_j d_j$  (recall that  $d_j = \lceil (\deg g_j)/2 \rceil$ ). If

$$(2-10) \quad \text{rank } \mathbf{M}_d(\mathbf{y}^d) = \text{rank } \mathbf{M}_{d-s}(\mathbf{y}^d) (= : t)$$

then  $\rho_d = f^*$  and there are  $t$  global minimizers  $\mathbf{x}_j^* \in \Omega$ ,  $j = 1, \dots, t$ , that can be “extracted” from  $\mathbf{y}^d$  by a linear algebra routine.

The sequence of SDP-relaxations  $(\mathbf{Q}_d)$ ,  $d \geq \hat{d}$ , and the rank test (Equation (2-10)) to extract global minimizers, are implemented in the GloptiPoly software [Henrion, Lasserre, and Löfberg \[2009\]](#).

**Finite convergence and a global optimality certificate.** After being introduced in [Lasserre \[2000\]](#), in many numerical experiments it was observed that typically, finite convergence takes place, that is,  $f^* = \rho_d$  for some (usually small)  $d$ . In fact there is a rationale behind this empirical observation.

**Theorem 2.4** ([Nie \[2014a\]](#)). *Let  $\mathbf{P}$  be as in Equation (1-3) where  $\Omega$  in Equation (1-2) is compact and its associated quadratic module is Archimedean. Suppose that at each global minimizer  $\mathbf{x}^* \in \Omega$ :*

- *The gradients  $(\nabla g_j(\mathbf{x}^*))_{j=1, \dots, m}$  are linearly independent. (This implies existence of nonnegative Lagrange-KKT multipliers  $\lambda_j^*$ ,  $j \leq m$ , such that  $\nabla f(\mathbf{x}^*) - \sum_{j=1}^m \lambda_j^* \nabla g_j(\mathbf{x}^*) = 0$  and  $\lambda_j^* g_j(\mathbf{x}^*) = 0$  for all  $j \leq m$ .)*
- *Strict complementarity holds, that is,  $g_j(\mathbf{x}^*) = 0 \Rightarrow \lambda_j^* > 0$ .*
- *Second-order sufficiency condition holds, i.e.,*

$$\langle \mathbf{u}, \nabla_{\mathbf{x}}^2 (f(\mathbf{x}^*) - \sum_{j=1}^m \lambda_j^* g_j(\mathbf{x}^*)) \mathbf{u} \rangle > 0,$$

for all  $0 \neq \mathbf{u} \in \nabla (f(\mathbf{x}^*) - \sum_{j=1}^m \lambda_j^* g_j(\mathbf{x}^*))^\perp$ .

Then  $f - f^* \in \mathcal{Q}(g)$ , i.e., there exists  $d^*$  and SOS multipliers  $\sigma_j^* \in \Sigma[\mathbf{x}]_{d^*-d_j}$ ,  $j = 0, \dots, m$ , such that:

$$(2-11) \quad f(\mathbf{x}) - f^* = \sigma_0^*(\mathbf{x}) + \sum_{j=1}^m \sigma_j^*(\mathbf{x}) g_j(\mathbf{x}).$$

With Equation (2-11), Theorem 2.4 provides a *certificate of global optimality* in polynomial optimization, and to the best of our knowledge, the first at this level of generality. Next, observe that  $\mathbf{x}^* \in \Omega$  is a global unconstrained minimizer of the *extended Lagrangian polynomial*  $f - f^* - \sum_{j=1}^n \sigma_j^* g_j$ , and therefore Theorem 2.4 is the analogue for *non-convex* polynomial optimization of the Karush-Kuhn-Tucker (KKT) optimality conditions

*in the convex case.* Indeed in the convex case, any local minimizer is global and is also a global unconstrained minimizer of the Lagrangian  $f - f^* - \sum_{j=1}^m \lambda_j^* g_j$ .

Also interestingly, whenever the SOS weight  $\sigma_j^*$  in Equation (2-11) is non trivial, it testifies that the constraint  $g_j(\mathbf{x}) \geq 0$  is important for  $\mathbf{P}$  even if it is not active at  $\mathbf{x}^*$  (meaning that if  $g_j \geq 0$  is deleted from  $\mathbf{P}$  then the new global optimum decreases strictly). The multiplier  $\lambda_j^*$  plays the same role in the KKT-optimality conditions *only in the convex case*. See Lasserre [2015a] for a detailed discussion.

**Finite convergence** of the Moment-SOS-hierarchies (Equations (2-8) and (2-9)) is an immediate consequence of Theorem 2.4. Indeed by Equation (2-11)  $(f^*, \sigma_0^*, \dots, \sigma_m^*)$  is a feasible solution of  $\mathbf{Q}_{d^*}^*$  with value  $f^* \leq \rho_d^* \leq f^*$  (hence  $\rho_d^* = \rho_d = f^*$ ).

**Genericity:** Importantly, as proved in Nie [2014a], the conditions in Theorem 2.4 are *generic*. By this we mean the following: Consider the class  $\mathcal{P}(t, m)$  of optimization problems  $\mathbf{P}$  with data  $(f, g_1, \dots, g_m)$  of degree bounded by  $t$ , and with nonempty compact feasible set  $\Omega$ . Such a problem  $\mathbf{P}$  is a “point” in the space  $\mathbb{R}^{(m+1)s(t)}$  of coordinates of  $(f, g_1, \dots, g_m)$ . Then the “good” problems  $\mathbf{P}$  are points in a Zariski open set. Moreover, generically the rank test (Equation (2-10)) is also satisfied at an optimal solution of Equation (2-8) (for some  $d$ ); for more details see Nie [2013].

**Computational complexity:** Each relaxation  $\mathbf{Q}_d$  in Equation (2-8) is a semidefinite program with  $s(2d) = \binom{n+2d}{n}$  variables  $(y_\alpha)$ , and a psd constraint  $\mathbf{M}_d(\mathbf{y}) \succeq 0$  of size  $s(d)$ . Therefore solving  $\mathbf{Q}_d$  in its canonical form Equation (2-8) is quite expensive in terms of computational burden, especially when using interior-point methods. Therefore its brute force application is limited to small to medium size problems.

**Exploiting sparsity:** Fortunately many large scale problems exhibit a structured sparsity pattern (e.g., each polynomial  $g_j$  is concerned with a few variables only, and the objective function  $f$  is a sum  $\sum_i f_i$  where each  $f_i$  is also concerned with a few variables only). Then Waki, Kim, Kojima, and Muramatsu [2006] have proposed a sparsity-adapted hierarchy of SDP-relaxations which can handle problems  $\mathbf{P}$  with thousands variables. In addition, if the sparsity pattern satisfies a certain condition then convergence of this sparsity-adapted hierarchy is also guaranteed like in the dense case Lasserre [2006]. Successful applications of this strategy can be found in e.g. Laumond, Mansard, and Lasserre [2017a] in Control (systems identification) and in Molzahn and Hiskens [2015] for solving (large scale) Optimum Power Flow problems (OPF is an important problem encountered in the management of energy networks).

**2.4 Discussion.** We claim that the Moment-SOS hierarchy and its rationale Theorem 2.4, unify convex, non-convex (continuous), and discrete (polynomial) Optimization. Indeed in the description of  $\mathbf{P}$  we do not pay attention to what particular class of problems  $\mathbf{P}$  belongs to. This is in sharp contrast to the usual common practice in (local) optimization

where several classes of problems have their own tailored favorite class of algorithms. For instance, problems are not treated the same if equality constraints appear, and/or if boolean (or discrete variables) are present, etc. Here a boolean variable  $x_i$  is modeled by the quadratic equality constraint  $x_i^2 = x_i$ . So it is reasonable to speculate that this lack of specialization could be a handicap for the moment-SOS hierarchy.

But this is not so. For instance for the sub-class of convex<sup>3</sup> problems  $\mathbf{P}$  where  $f$  and  $(-g_j)_{j=1,\dots,m}$  are SOS-convex<sup>4</sup> polynomials, finite convergence takes place at the first step of the hierarchy. In other words, the SOS hierarchy somehow “recognizes” this class of easy problems [Lasserre \[2015a\]](#). In the same time, for a large class of 0/1 combinatorial optimization problems on graphs, the Moment-SOS hierarchy has been shown to provide the tightest upper bounds when compared to the class of *lift-and-project* methods, and has now become a central tool to analyze hardness of approximations in combinatorial optimization. For more details the interested reader is referred to e.g. [Lasserre \[2002b\]](#), [Laurent \[2003\]](#), [Barak and Steurer \[2014\]](#), [Khot \[2010, 2014\]](#) and the many references therein.

### 3 The Moment-SOS hierarchy outside optimization

**3.1 A general framework for the Moment-SOS hierarchy.** Let  $\Omega_i \subset \mathbb{R}^{n_i}$  be a finite family of compact sets,  $\mathcal{M}(\Omega_i)$  (resp.  $\mathcal{C}(\Omega_i)$ ) be the space of finite Borel signed measures (resp. continuous functions) on  $\Omega_i$ ,  $i = 0, 1, \dots, s$ , and let  $\mathbf{T}$  be a continuous linear mapping with adjoint  $\mathbf{T}^*$ :

$$\begin{aligned} \mathbf{T} : \mathcal{M}(\Omega_1) \times \dots \times \mathcal{M}(\Omega_s) &\rightarrow \mathcal{M}(\Omega_0) \\ \mathcal{C}(\Omega_1) \times \dots \times \mathcal{C}(\Omega_s) &\leftarrow \mathcal{C}(\Omega_0) : \mathbf{T}^* \end{aligned}$$

Let  $\phi := (\phi_1, \dots, \phi_s)$  and let  $\phi_i \geq 0$  stand for  $\phi_i$  is a positive measure. Then consider the general framework:

$$(3-1) \quad \rho = \inf_{\phi \geq 0} \left\{ \sum_{i=1}^s \langle f_i, \phi_i \rangle : \mathbf{T}(\phi) = \lambda; \sum_{i=1}^s \langle f_{ij}, \phi_i \rangle \geq b_j, j \in J \right\},$$

where  $J$  is a finite or countable set,  $\mathbf{b} = (b_j)$  is given,  $\lambda \in \mathcal{M}(\Omega_0)$  is a given measure,  $(f_{ij})_{j \in J, i = 1, \dots, s}$  are given polynomials, and  $\langle \cdot, \cdot \rangle$  is the duality bracket between  $\mathcal{C}(\Omega_i)$  and  $\mathcal{M}(\Omega_i)$  ( $\langle h, \phi_i \rangle = \int_{\Omega_i} h d\phi_i$ ),  $i = 1, \dots, s$ .

<sup>3</sup>Convex problems  $\mathbf{P}$  where  $f$  and  $(-g_j)_{j=1,\dots,m}$  are convex, are considered “easy” and can be solved efficiently.

<sup>4</sup>A polynomial  $f \in \mathbb{R}[\mathbf{x}]$  is SOS-convex if its Hessian  $\nabla^2 f$  is a SOS matrix-polynomial, i.e.,  $\nabla f^2(\mathbf{x}) = \mathbf{L}(\mathbf{x})\mathbf{L}(\mathbf{x})^T$  for some matrix-polynomial  $\mathbf{L} \in \mathbb{R}[\mathbf{x}]^{n \times p}$ .

As we will see, this general framework is quite rich as it encompasses a lot of important applications in many different fields. In fact Problem (3-1) is equivalent to the Generalized Problem of Moments (GPM):

$$(3-2) \quad \rho = \inf_{\phi \geq 0} \left\{ \sum_{i=1}^s \langle f_i, \phi_i \rangle : \langle \mathbf{T}^* p_k, \phi \rangle = \langle p_k, \lambda \rangle, \quad k = 0, 1, \dots \right. \\ \left. \sum_{i=1}^s \langle f_{ij}, \phi_i \rangle \geq b_j, \quad j \in J \right\},$$

where the family  $(p_k)_{k=0,\dots}$  is dense in  $\mathcal{C}(\Omega_0)$  (e.g. a basis of  $\mathbb{R}[x_1, \dots, x_{n_0}]$ ).

The Moment-SOS hierarchy can also be applied to help solve the Generalized Problem of Moments (GPM) (Equation (3-2)) or its dual :

$$(3-3) \quad \rho^* = \sup_{(\theta_j \geq 0, \gamma)} \left\{ \sum_k \gamma_k \langle p_k, \lambda \rangle + \langle \theta, \mathbf{b} \rangle : \right. \\ \left. \text{s.t. } f_i - \sum_k \gamma_k (\mathbf{T}^* p_k)_i - \sum_{j \in J} \theta_j f_{ij} \geq 0 \text{ on } \Omega_i \text{ for all } i \right\},$$

where the unknown  $\gamma = (\gamma_k)_{k \in \mathbb{N}}$  is a finite sequence.

### 3.2 A hierarchy of SDP-relaxations. Let

$$(3-4) \quad \Omega_i := \{ \mathbf{x} \in \mathbb{R}^{n_i} : g_{i,\ell}(\mathbf{x}) \geq 0, \ell = 1, \dots, m_i \}, \quad i = 1, \dots, s,$$

for some polynomials  $(g_{i,\ell}) \subset \mathbb{R}[x_1, \dots, x_{n_i}]$ ,  $\ell = 1, \dots, m_i$ . Let  $d_{i,\ell} = \lceil \deg(g_{i,\ell})/2 \rceil$  and  $\hat{d} := \max_{i,j,\ell} [\deg(f_i), \deg(f_{ij}), \deg(g_{i,\ell})]$ . To solve Equation (3-2), define the “moment” sequences  $\mathbf{y}_i = (y_{i,\alpha})$ ,  $\alpha \in \mathbb{N}^{n_i}$ ,  $i = 1, \dots, s$ , and with  $d \in \mathbb{N}$ , define  $\Gamma_d := \{ p_k : \deg(T^* p_k)_i \leq 2d, i = 1, \dots, s \}$ . Consider the hierarchy of semidefinite programs indexed by  $\hat{d} \leq d \in \mathbb{N}$ :

$$(3-5) \quad \rho_d = \inf_{(\mathbf{y}_i)} \left\{ \sum_{i=1}^s L_{\mathbf{y}_i}(f_i) : \sum_{i=1}^s L_{\mathbf{y}_i}((T^* p_k)_i) = \langle p_k, \lambda \rangle, \quad p_k \in \Gamma_d \right. \\ \left. \sum_{i=1}^s L_{\mathbf{y}_i}(f_{ij}) \geq b_j, \quad j \in J_d \right. \\ \left. \mathbf{M}_d(\mathbf{y}_i), \mathbf{M}_{d-d_\ell}(g_{i,\ell} \mathbf{y}_i) \geq 0, \quad \ell \leq m_i; i \leq s \right\},$$

where  $J_d \subset J$  is finite  $\bigcup_{d \in \mathbb{N}} J_d = J$ . Its dual SDP-hierarchy reads:

$$(3-6) \quad \begin{aligned} \rho_d^* = & \sup_{(\theta_j \geq 0, \gamma_k)} \left\{ \sum_{p_k \in \Gamma_d} \gamma_k \langle p_k, \lambda \rangle + \langle \theta, \mathbf{b} \rangle : \right. \\ \text{s.t. } & f_i - \sum_{p_k \in \Gamma_d} \gamma_k (\mathbf{T}^* p_k)_i - \sum_{j \in J} \theta_j f_{ij} = \sum_{\ell=0}^{m_i} \sigma_{i,\ell} g_{i,\ell} \\ & \sigma_{i,\ell} \in \Sigma[x_1, \dots, x_{n_i}]_{d-d_{i,\ell}}; \ i = 1, \dots, s \}, \end{aligned}$$

As each  $\Omega_i$  is compact, for technical reasons and with no loss of generality, in the sequel we may and will assume that for every  $i = 1, \dots, s$ ,  $g_{i,0}(\mathbf{x}) = M_i - \|\mathbf{x}\|^2$ , where  $M_i > 0$  is sufficiently large.

**Theorem 3.1.** *Assume that  $\rho > -\infty$  and that for every  $i = 1, \dots, s$ ,  $f_{i0} = 1$ . Then for every  $d \geq \hat{d}$ , Equation (3-5) has an optimal solution, and  $\lim_{d \rightarrow \infty} \rho_d = \rho$ .*

### 3.3 Example In Probability and Computational Geometry.

**Bounds on measures with moment conditions.** Let  $Z$  be a random vector with values in a compact semi-algebraic set  $\Omega_1 \subset \mathbb{R}^n$ . Its distribution  $\lambda$  on  $\Omega_1$  is unknown but some of its moments  $\int \mathbf{x}^\alpha d\lambda = b_\alpha$ ,  $\alpha \in \Gamma \subset \mathbb{N}^n$ , are known ( $b_0 = 1$ ). Given a basic semi-algebraic set  $\Omega_2 \subset \Omega_1$  we want to compute (or approximate as closely as desired) the best upper bound on  $\text{Prob}(Z \in \Omega_2)$ . This problem reduces to solving the GPM:

$$(3-7) \quad \begin{aligned} \rho = & \sup_{\phi_1, \phi_2 \geq 0} \{ \langle 1, \phi_2 \rangle : \langle \mathbf{x}^\alpha, \phi_1 \rangle + \langle \mathbf{x}^\alpha, \phi_2 \rangle = b_\alpha, \alpha \in \Gamma; \\ & \phi_i \in \mathcal{M}(\Omega_i), i = 1, 2 \}, \end{aligned}$$

With  $\Omega_1$  and  $\Omega_2$  as in Equation (3-4) one may compute upper bounds on  $\rho$  by solving the Moment-SOS hierarchy (Equation (3-5)) adapted to problem (Equation (3-7)). Under the assumptions of Theorem 3.1, the resulting sequence  $(\rho_d)_{d \in \mathbb{N}}$  converges to  $\rho$  as  $d \rightarrow \infty$ ; for more details the interested reader is referred to Lasserre [2002a].

**Lebesgue & Gaussian measures of semi-algebraic sets.** Let  $\Omega_2 \subset \mathbb{R}^n$  be compact. The goal is to compute (or approximate as closely as desired) the Lebesgue measure  $\lambda(\Omega_2)$  of  $\Omega_2$ . Then take  $\Omega_1 \supset \Omega_2$  be a simple set, e.g. an ellipsoid or a box (in fact any set such that one knows all moments  $(b_\alpha)_{\alpha \in \mathbb{N}^n}$  of the Lebesgue measure on  $\Omega_1$ ). Then:

$$(3-8) \quad \begin{aligned} \lambda(\Omega_2) = & \sup_{\phi_1, \phi_2 \geq 0} \{ \langle 1, \phi_2 \rangle : \langle \mathbf{x}^\alpha, \phi_1 \rangle + \langle \mathbf{x}^\alpha, \phi_2 \rangle = b_\alpha, \alpha \in \mathbb{N}^n; \\ & \phi_i \in \mathcal{M}(\Omega_i), i = 1, 2 \}. \end{aligned}$$

Problem (3-8) is very similar to (3-7) except that we now have countably many moment constraints ( $\Gamma = \mathbb{N}^n$ ). Again, with  $\Omega_2$  and  $\Omega_2$  as in Equation (3-4) one may compute upper bounds on  $\lambda(\Omega_2)$  by solving the Moment-SOS hierarchy (Equation (3-5)) adapted to problem (3-8). Under the assumptions of Theorem 3.1, the resulting monotone non-increasing sequence  $(\rho_d)_{d \in \mathbb{N}}$  converges to  $\lambda(\Omega_2)$  from above as  $d \rightarrow \infty$ . The convergence  $\rho_d \rightarrow \lambda(\Omega_2)$  is slow because of a Gibb's phenomenon<sup>5</sup>. Indeed the semidefinite program (Equation (3-6)) reads:

$$\rho_d^* = \inf_{p \in \mathbb{R}[\mathbf{x}]_{2d}} \left\{ \int_{\Omega_1} p \, d\lambda : p \geq 1 \text{ on } \Omega_2; \quad p \geq 0 \text{ on } \Omega_1 \right\},$$

i.e., as  $d \rightarrow \infty$  one tries to approximate the discontinuous function  $\mathbf{x} \mapsto 1_{\Omega_2}(\mathbf{x})$  by polynomials of increasing degrees. Fortunately there are several ways to accelerate the convergence, e.g. as in Henrion, Lasserre, and Savorgnan [2009] (but loosing the monotonicity) or in Lasserre [2017] (preserving monotonicity) by including in Equation (3-5) additional constraints on  $\mathbf{y}_2$  coming from an application of Stokes' theorem.

For the **Gaussian measure**  $\lambda$  we need and may take  $\Omega_1 = \mathbb{R}^n$  and  $\Omega_2$  is not necessarily compact. Although both  $\Omega_1$  and  $\Omega_2$  are allowed to be non-compact, the Moment-SOS hierarchy (Equation (3-5)) still converges, i.e.,  $\rho_d \rightarrow \lambda(\Omega_2)$  as  $d \rightarrow \infty$ . This is because the moments of  $\lambda$  satisfy the generalized Carleman's condition

$$(3-9) \quad \sum_{k=1}^{\infty} \left( \int_{\mathbb{R}^n} x_i^{2k} \, d\lambda \right)^{-1/2k} = +\infty, \quad i = 1, \dots, n,$$

which imposes implicit constraints on  $\mathbf{y}_1$  and  $\mathbf{y}_2$  in Equation (3-5), strong enough to guarantee  $\rho_d \rightarrow \lambda(\Omega_2)$  as  $d \rightarrow \infty$ . For more details see Lasserre [ibid.]. This deterministic approach is computationally demanding and should be seen as complementary to brute force Monte-Carlo methods that provide only an estimate (but can handle larger size problems).

**3.4 In signal processing and interpolation.** In this application, a signal is identified with an atomic signed measure  $\phi$  supported on few atoms  $(\mathbf{x}_k)_{k=1,\dots,s} \subset \Omega$ , i.e.,  $\phi = \sum_{k=1}^s \theta_k \delta_{\mathbf{x}_k}$ , for some weights  $(\theta_k)_{k=1,\dots,s}$ .

**Super-Resolution.** The goal of Super-Resolution is to reconstruct the unknown measure  $\phi$  (the signal) from a few measurements only, when those measurements are the

<sup>5</sup>The Gibbs' phenomenon appears at a jump discontinuity when one approximates a piecewise  $C^1$  function with a continuous function, e.g., by its Fourier series.

moments  $(b_\alpha)_{\alpha \in \mathbb{N}_t^n}$  of  $\phi$ , up to order  $t$  (fixed). One way to proceed is to solve the infinite-dimensional program:

$$(3-10) \quad \rho = \inf_{\phi} \{ \|\phi\|_{TV} : \int \mathbf{x}^\alpha d\phi = b_\alpha, \quad \alpha \in \mathbb{N}_t^n \},$$

where the inf is over the finite signed Borel measures on  $\Omega$ , and  $\|\phi\|_{TV} = |\phi|(\Omega)$  (with  $|\phi|$  being the total variation of  $\phi$ ). Equivalently:

$$(3-11) \quad \rho = \inf_{\phi^+, \phi^- \geq 0} \{ \langle 1, \phi^+ + \phi^- \rangle : \langle \mathbf{x}^\alpha, \phi^+ - \phi^- \rangle = b_\alpha, \quad \alpha \in \mathbb{N}_t^n \},$$

which is an instance of the GPM with dual:

$$(3-12) \quad \rho^* = \sup_{p \in \mathbb{R}[\mathbf{x}]_t} \{ \sum_{\alpha \in \mathbb{N}_t^n} p_\alpha b_\alpha : \|p\|_\infty \leq 1 \},$$

where  $\|p\|_\infty = \sup\{|p(\mathbf{x})| : \mathbf{x} \in \Omega\}$ . In this case, the Moment-SOS hierarchy (Equation (3-5)) with  $d \geq \hat{d} := \lceil t/2 \rceil$ , reads:

$$(3-13) \quad \begin{aligned} \rho_d &= \inf_{y^+, y^-} \{ y_0^+ + y_0^- : y_\alpha^+ - y_\alpha^- = b_\alpha, \quad \alpha \in \mathbb{N}_t^n \\ &\quad \mathbf{M}_d(\mathbf{y}^\pm) \geq 0; \mathbf{M}_{d-d_\ell}(g_j \mathbf{y}^\pm) \geq 0, \quad \ell = 1, \dots, m \}, \end{aligned}$$

where  $\Omega = \{\mathbf{x} : g_\ell(\mathbf{x}) \geq 0, \ell = 1, \dots, m\}$ .

In the case where  $\Omega$  is the torus  $\mathbb{T} \subset \mathbb{C}$ , Candès and Fernandez-Granda [2014] showed that if  $\delta > 2/f_c$  (where  $\delta$  is the minimal distance between the atoms of  $\phi$ , and  $f_c$  is the number of measurements) then Equation (3-10) has a unique solution and one may recover  $\phi$  *exactly* by solving the single semidefinite program (Equation (3-10)) with  $d = \lceil t/2 \rceil$ . The dual (Equation (3-12)) has an optimal solution  $p^*$  (a trigonometric polynomial) and the support of  $\phi^+$  (resp.  $\phi^-$ ) consists of the atoms  $\mathbf{z} \in \mathbb{T}$  of  $\phi$  such that  $p^*(\mathbf{z}) = 1$  (resp.  $p^*(\mathbf{z}) = -1$ ). In addition, this procedure is more robust to noise in the measurements than Prony's method; on the other hand, the latter requires less measurements and no separation condition on the atoms.

In the general multivariate case treated in De Castro, Gamboa, Henrion, and Lasserre [2017] one now needs to solve the Moment-SOS hierarchy (Equation (3-11)) for  $d = \hat{d}, \dots$  (instead of a single SDP in the univariate case). However since the moment constraints of Equation (3-11) are finitely many, exact recovery (i.e. finite convergence of the Moment-SOS hierarchy (Equation (3-13))) is possible (usually with a few measurements only). This is indeed what has been observed in all numerical experiments of De Castro, Gamboa, Henrion, and Lasserre [ibid.], and in all cases with significantly less measurements than the theoretical bound (of a tensorized version of the univariate case).

In fact, the rank condition (Equation (2-10)) is always satisfied at an optimal solution  $(\mathbf{y}^+, \mathbf{y}^-)$  at some step  $d$  of the hierarchy (Equation (3-13)), and so the atoms of  $\phi^+$  and  $\phi^-$  are extracted via a simple linear algebra routine (as for global optimization). Nie's genericity result Nie [2013] should provide a rationale which explains why the rank condition (Equation (2-10)) is satisfied in all examples.

**Sparse interpolation.** Here the goal is to recover an unknown (black-box) polynomial  $p \in \mathbb{R}[\mathbf{x}]_t$  through a few evaluations of  $p$  only. In Josz, Lasserre, and Mourrain [2017] we have shown that this problem is in fact a particular case of Super-Resolution (and even *discrete* Super-Resolution) on the torus  $\mathbb{T}^n \subset \mathbb{C}^n$ . Indeed let  $\mathbf{z}_0 \in \mathbb{T}^n$  be fixed, arbitrary. Then with  $\beta \in \mathbb{N}^n$ , notice that

$$\begin{aligned} p(\mathbf{z}_0^\beta) &= \sum_{\alpha \in \mathbb{N}_d^n} p_\alpha (z_{01}^{\beta_1} \cdots z_{0n}^{\beta_n})^\alpha = \sum_{\alpha \in \mathbb{N}_d^n} p_\alpha (z_{01}^{\alpha_1} \cdots z_{0n}^{\alpha_n})^\beta \\ &= \int_{\mathbb{T}^n} \mathbf{z}^\beta d \left( \sum_{\alpha \in \mathbb{N}_d^n} p_\alpha \delta_{\mathbf{z}_0^\alpha} \right) = \int_{\mathbb{T}^n} \mathbf{z}^\beta d\phi. \end{aligned}$$

In other words, one may identify the polynomial  $p$  with an atomic signed Borel measure  $\phi$  on  $\mathbb{T}^n$  supported on finitely many atoms  $(\mathbf{z}_0^\alpha)_{\alpha \in \mathbb{N}_d^n}$  with associated weights  $(p_\alpha)_{\alpha \in \mathbb{N}_d^n}$ .

Therefore, if the evaluations of the black-box polynomial  $p$  are done at a few “powers”  $(\mathbf{z}_0^\beta)$ ,  $\beta \in \mathbb{N}^n$ , of an arbitrary point  $\mathbf{z}_0 \in \mathbb{T}^n$ , then the sparse interpolation problem is equivalent to recovering an unknown atomic signed Borel measure  $\phi$  on  $\mathbb{T}^n$  from knowledge of a few moments, that is, the Super-Resolution problem that we have just described above. Hence one may recover  $p$  by solving the Moment-SOS hierarchy (Equation (3-13)) for which finite convergence usually occurs fast. For more details see Josz, Lasserre, and Mourrain [ibid.].

**3.5 In Control & Optimal Control.** Consider the Optimal Control Problem (OCP) associated with a controlled dynamical system:

$$\begin{aligned} (3-14) \quad J^* &= \inf_{\mathbf{u}(t)} \int_0^T L(\mathbf{x}(t), \mathbf{u}(t)) dt : \dot{\mathbf{x}}(t) = f(\mathbf{x}(t), \mathbf{u}(t)), t \in (0, T) \\ &\quad \mathbf{x}(t) \in \mathbf{X}, \mathbf{u}(t) \in \mathbf{U}, \forall t \in (0, T) \\ &\quad \mathbf{x}(0) = \mathbf{x}_0; \mathbf{x}(T) \in \mathbf{X}_T, \end{aligned}$$

where  $L, f$  are polynomials,  $\mathbf{X}, \mathbf{X}_T \subset \mathbb{R}^n$  and  $\mathbf{U} \subset \mathbb{R}^p$  are compact basic semi-algebraic sets. In full generality the OCP problem (Equation (3-14)) is difficult to solve, especially when state constraints  $\mathbf{x}(t) \in \mathbf{X}$  are present. Given an admissible state-control trajectory



$(t, \mathbf{x}(t), \mathbf{u}(t))$ , its associated occupation measure  $\phi_1$  up to time  $T$  (resp.  $\phi_2$  at time  $T$ ) are defined by:

$$\phi_1(A \times B \times C) := \int_{[0, T] \cap C} 1_{(A, B)}((\mathbf{x}(t), \mathbf{u}(t))) dt; \quad \phi_2(D) = 1_D(\mathbf{x}(T)),$$

for all  $A \in \mathfrak{B}(\mathbf{X})$ ,  $B \in \mathfrak{B}(\mathbf{U})$ ,  $C \in \mathfrak{B}([0, T])$ ,  $D \in \mathfrak{B}(\mathbf{X}_T)$ . Then for every differentiable function  $h : \mathbf{X} \times [0, T] \rightarrow \mathbb{R}$

$$h(T, \mathbf{x}(T)) - h(0, x_0) = \int_0^T \left( \frac{\partial h(\mathbf{x}(t), \mathbf{u}(t))}{\partial t} + \frac{\partial h(\mathbf{x}(t), \mathbf{u}(t))}{\partial \mathbf{x}} f(\mathbf{x}(t), \mathbf{u}(t)) \right) dt,$$

or, equivalently, with  $\mathbf{S} := [0, T] \times \mathbf{X} \times \mathbf{U}$ :

$$\int_{\mathbf{X}_T} h(T, \mathbf{x}) d\phi_2(\mathbf{x}) = h(0, \mathbf{x}_0) + \int_{\mathbf{S}} \left( \frac{\partial h(\mathbf{x}, \mathbf{u})}{\partial t} + \frac{\partial h(\mathbf{x}, \mathbf{u})}{\partial \mathbf{x}} f(\mathbf{x}, \mathbf{u}) \right) d\phi_1(t, \mathbf{x}, \mathbf{u}).$$

Then *the weak formulation* of the OCP (Equation (3-14)) is the infinite-dimensional linear program:

$$(3-15) \quad \begin{aligned} \rho = \inf_{\phi_1, \phi_2 \geq 0} \{ & \int_{\mathbf{S}} L(\mathbf{x}, \mathbf{u}) d\phi_1 : \\ \text{s.t. } & \int_{\mathbf{X}_T} h(T, \cdot) d\phi_2 - \int_{\mathbf{S}} \left( \frac{\partial h}{\partial t} + \frac{\partial h}{\partial \mathbf{x}} f \right) d\phi_1 = h(0, \mathbf{x}_0) \\ & \forall h \in \mathbb{R}[t, \mathbf{x}] \}. \end{aligned}$$

It turns out that under some conditions the optimal values of Equations (3-14) and (3-15) are equal, i.e.,  $J^* = \rho$ . Next, if one replaces “for all  $h \in \mathbb{R}[t, \mathbf{x}, \mathbf{u}]$ ” with “for all  $t^k \mathbf{x}^\alpha \mathbf{u}^\beta$ ”,  $(t, \alpha, \beta) \in \mathbb{N}^{1+n+p}$ , then Equation (3-15) is an instance of the GPM (Equation (3-2)). Therefore one may apply the Moment-SOS hierarchy (Equation (3-5)). Under the conditions of Theorem 3.1 one obtains the asymptotic convergence  $\rho_d \rightarrow \rho = J^*$  as  $d \rightarrow \infty$ . For more details see Lasserre, Henrion, Prieur, and Trélat [2008] and the many references therein.

**Robust control.** In some applications (e.g. in robust control) one is often interested in optimizing over sets of the form:

$$\mathbf{G} := \{\mathbf{x} \in \Omega_1 : f(\mathbf{x}, \mathbf{u}) \geq 0, \forall \mathbf{u} \in \Omega_2\},$$

where  $\Omega_2 \subset \mathbb{R}^p$ , and  $\Omega_1 \subset \mathbb{R}^n$  is a simple set, in fact a compact set such that one knows all moments of the Lebesgue measure  $\lambda$  on  $\Omega_1$ .

The set  $\mathbf{G}$  is difficult to handle because of the universal quantifier. Therefore one is often satisfied with an inner approximation  $\mathbf{G}_d \subset \mathbf{G}$ , and if possible, with (i) a simple

form and (ii) some theoretical approximation guarantees. We propose to approximate  $\mathbf{G}$  from inside by sets of (simple) form  $\mathbf{G}_d = \{\mathbf{x} \in \Omega_1 : p_d(\mathbf{x}) \geq 0\}$  where  $p_d \in \mathbb{R}[\mathbf{x}]_{2d}$ .

To obtain such an inner approximation  $\mathbf{G}_d \subset \mathbf{G}$ , define  $F : \Omega_1 \rightarrow \mathbb{R}$ ,  $\mathbf{x} \mapsto F(\mathbf{x}) := \min_{\mathbf{u}} \{f(\mathbf{x}, \mathbf{u}) : \mathbf{u} \in \Omega_2\}$ . Then with  $d \in \mathbb{N}$ , fixed, solve:

$$(3-16) \quad \inf_{p \in \mathbb{R}[\mathbf{x}]_{2d}} \int_{\Omega_1} (F - p) d\lambda : f(\mathbf{x}, \mathbf{u}) - p(\mathbf{x}) \geq 0, \forall (\mathbf{x}, \mathbf{u}) \in \Omega_1 \times \Omega_2\}.$$

Any feasible solution  $p_d$  of Equation (3-16) is such that  $\mathbf{G}_d = \{\mathbf{x} : p_d(\mathbf{x}) \geq 0\} \subset \mathbf{G}$ . In Equation (3-16)  $\int_{\Omega_1} (F - p) d\lambda = \|F - p\|_1$  (with  $\|\cdot\|_1$  being the  $L_1(\Omega_1)$ -norm), and

$$\inf_p \int_{\Omega_1} (F - p) d\lambda = \underbrace{\int_{\Omega_1} F d\lambda}_{=\text{cte}} + \inf_p \int_{\Omega_1} -p d\lambda = \text{cte} - \sup_p \int_{\Omega_1} p d\lambda$$

and so in Equation (3-16) it is equivalent to maximize  $\int_{\Omega_1} p d\lambda$ . Again the Moment-SOS hierarchy can be applied. This time one replaces the difficult positivity constraint  $f(\mathbf{x}, \mathbf{u}) - p(\mathbf{x}) \geq 0$  for all  $(\mathbf{x}, \mathbf{u}) \in \Omega_1 \times \Omega_2$  with a certificate of positivity, with a degree bound on the SOS weights. That is, if  $\Omega_1 = \{\mathbf{x} : g_{1,\ell}(\mathbf{x}) \geq 0, \ell = 1, \dots, m_1\}$  and  $\Omega_2 = \{\mathbf{u} : g_{2,\ell}(\mathbf{u}) \geq 0, \ell = 1, \dots, m_2\}$ , then with  $d_{i,\ell} := \lceil (\deg(\sigma_{i,\ell})/2 \rceil$ , one solves

$$(3-17) \quad \begin{aligned} \rho_d = \sup_{p \in \mathbb{R}[\mathbf{x}]_{2d}} \int_{\Omega_1} p d\lambda : & f(\mathbf{x}, \mathbf{u}) - p(\mathbf{x}) = \sigma_0(\mathbf{x}, \mathbf{u}) \\ & + \sum_{\ell=1}^{m_1} \sigma_{1,\ell}(\mathbf{x}, \mathbf{u}) g_{1,\ell}(\mathbf{x}) + \sum_{\ell=1}^{m_2} \sigma_{2,\ell}(\mathbf{x}, \mathbf{u}) g_{2,\ell}(\mathbf{u}) \\ & \sigma_{i,\ell} \in \Sigma[\mathbf{x}, \mathbf{u}]_{d-d_{i,\ell}}, \ell = 1, \dots, m_i, i = 1, 2. \end{aligned}$$

**Theorem 3.2** (Lasserre [2015b]). *Assume that  $\Omega_1 \times \Omega_2$  is compact and its associated quadratic module is Archimedean. Let  $p_d$  be an optimal solution of Equation (3-17). If  $\lambda(\{\mathbf{x} \in \Omega_1 : F(\mathbf{x}) = 0\}) = 0$  then  $\lim_{d \rightarrow \infty} \|F - p_d\|_1 = 0$  and  $\lim_{d \rightarrow \infty} \lambda(\mathbf{G} \setminus \mathbf{G}_d) = 0$ .*

Therefore one obtains a nested sequence of inner approximations  $(\mathbf{G}_d)_{d \in \mathbb{N}} \subset \mathbf{G}$ , with the desirable property that  $\lambda(\mathbf{G} \setminus \mathbf{G}_d)$  vanishes as  $d$  increases. For more details the interested reader is referred to Lasserre [ibid.].

**Example 1.** In some robust control problems one would like to approximate as closely as desired a non-convex set  $\mathbf{G} = \{\mathbf{x} \in \Omega_1 : \lambda_{\min}(\mathbf{A}(\mathbf{x})) \geq 0\}$  for some real symmetric  $r \times r$  matrix-polynomial  $\mathbf{A}(\mathbf{x})$ , and where  $\mathbf{x} \mapsto \lambda_{\min}(\mathbf{A}(\mathbf{x}))$  denotes its smallest eigenvalue. If one rewrites

$$\mathbf{G} = \{\mathbf{x} \in \Omega_1 : \mathbf{u}^T \mathbf{A}(\mathbf{x}) \mathbf{u} \geq 0, \forall \mathbf{u} \in \Omega_2\}; \quad \Omega_2 = \{\mathbf{u} \in \mathbb{R}^r : \|\mathbf{u}\| = 1\},$$

one is faced with the problem we have just described. In applying the above methodology the polynomial  $p_d$  in [Theorem 3.2](#) approximates  $\lambda_{\min}(\mathbf{A}(\mathbf{x}))$  from below in  $\Omega_1$ , and  $\|p_d(\cdot) - \lambda_{\min}(\mathbf{A}(\cdot))\|_1 \rightarrow 0$  as  $d$  increases. For more details see [Henrion and Lasserre \[2006\]](#).

There are many other applications of the Moment-SOS hierarchy in Control, e.g. in Systems Identification [Cerone, Piga, and Regruto \[2012\]](#) and [Laumond, Mansard, and Lasserre \[2017a\]](#), Robotics [Posa, Tobenkin, and Tedrake \[2016\]](#), for computing Lyapunov functions [Parrilo \[2003\]](#), largest regions of attraction [Henrion and Korda \[2014\]](#), to cite a few.

### 3.6 Some inverse optimization problems. In particular:

**Inverse Polynomial Optimization.** Here we are given a polynomial optimization problem  $\mathbf{P}$ :  $f^* = \min\{f(\mathbf{x}) : \mathbf{x} \in \Omega\}$  with  $f \in \mathbb{R}[\mathbf{x}]_d$ , and we are interested in the following issue: Let  $\mathbf{y} \in \Omega$  be given, e.g.  $\mathbf{y}$  is the current iterate of a local minimization algorithm applied to  $\mathbf{P}$ . Find

$$(3-18) \quad g^* = \arg \min_{g \in \mathbb{R}[\mathbf{x}]_d} \{\|f - g\|_1 : g(\mathbf{x}) - g(\mathbf{y}) \geq 0, \forall \mathbf{x} \in \Omega\},$$

where  $\|h\|_1 = \sum_{\alpha} |h_{\alpha}|$  is the  $\ell_1$ -norm of coefficients of  $h \in \mathbb{R}[\mathbf{x}]_d$ . In other words, one searches for a polynomial  $g^* \in \mathbb{R}[\mathbf{x}]_d$  as close as possible to  $f$  and such that  $\mathbf{y} \in \Omega$  is a global minimizer of  $g^*$  on  $\Omega$ . Indeed if  $\|f - g^*\|_1$  is small enough then  $\mathbf{y} \in \Omega$  could be considered a satisfying solution of  $\mathbf{P}$ . Therefore given a fixed small  $\epsilon > 0$ , the test  $\|f - g^*\|_1 < \epsilon$  could be a new stopping criterion for a local optimization algorithm, with a strong theoretical justification.

Again the Moment-SOS hierarchy can be applied to solve [Equation \(3-18\)](#) as positivity certificates are perfect tools to handle the positivity constraint “ $g(\mathbf{x}) - g(\mathbf{y}) \geq 0$  for all  $\mathbf{x} \in \Omega$ ”. Namely with  $\Omega$  as in [Equation \(1-2\)](#), solve:

$$(3-19) \quad \rho_t = \min_{g \in \mathbb{R}[\mathbf{x}]_d} \{\|f - g\|_1 : g(\mathbf{x}) - g(\mathbf{y}) := \sum_{j=0}^m \sigma_j(\mathbf{x}) g_j(\mathbf{x}), \quad \forall \mathbf{x}\},$$

where  $g_0(\mathbf{x}) = 1$  for all  $\mathbf{x}$ , and  $\sigma_j \in \Sigma[\mathbf{x}]_{t-d_j}$ ,  $j = 0, \dots, m$ . Other norms are possible but for the sparsity inducing  $\ell_1$ -norm  $\|\cdot\|_1$ , it turns out that an optimal solution  $g^*$  of [Equation \(3-19\)](#) has a canonical simple form. For more details the interested reader is referred to [Lasserre \[2013\]](#).

**Inverse Optimal Control.** With the OCP (Equation (3-14)) in Section 3.5, we now consider the following issue: *Given a database of admissible trajectories  $(\mathbf{x}(t; \mathbf{x}_\tau), \mathbf{u}(t, \mathbf{x}_\tau))$ ,  $t \in [\tau, T]$ , starting in initial state  $\mathbf{x}_\tau \in \mathbf{X}$  at time  $\tau \in [0, T]$ , does there exist a Lagrangian  $(\mathbf{x}, \mathbf{u}) \mapsto L(\mathbf{x}, \mathbf{u})$  such that all these trajectories are optimal for the OCP problem (Equation (3-14))?* This problem has important applications, e.g., in Humanoid Robotics to explain human locomotion [Laumond, Mansard, and Lasserre \[2017b\]](#).

Again the Moment-SOS hierarchy can be applied because a weak version of the Hamilton-Jacobi-Bellman (HJB) optimality conditions is the perfect tool to state whether some given trajectory is  $\epsilon$ -optimal for the OCP (Equation (3-14)). Indeed given  $\epsilon > 0$  and an admissible trajectory  $(t, \mathbf{x}^*(t), \mathbf{u}^*(t))$ , let  $\varphi : [0, T] \times \mathbf{X} \rightarrow \mathbb{R}$ , and  $L : \mathbf{X} \times \mathbf{U} \rightarrow \mathbb{R}$ , be such that:

$$(3-20) \quad \varphi(T, \mathbf{x}) \leq 0, \quad \forall \mathbf{x} \in \mathbf{X}; \quad \frac{\partial \varphi(t, \mathbf{x})}{\partial t} + \frac{\partial \varphi(t, \mathbf{x})}{\partial \mathbf{x}} f(\mathbf{x}, \mathbf{u}) + L(\mathbf{x}, \mathbf{u}) \geq 0,$$

for all  $(t, \mathbf{x}, \mathbf{u}) \in [0, T] \times \mathbf{X} \times \mathbf{U}$ , and:  $\varphi(T, \mathbf{x}^*(T)) > -\epsilon$ ,

$$(3-21) \quad \frac{\partial \varphi(t, \mathbf{x}^*(t))}{\partial t} + \frac{\partial \varphi(t, \mathbf{x}^*(t))}{\partial \mathbf{x}} f(\mathbf{x}^*(t), \mathbf{u}^*(t)) + L(\mathbf{x}^*(t), \mathbf{u}^*(t)) < \epsilon,$$

for all  $t \in [0, T]$ . Then the trajectory  $(t, \mathbf{x}^*(t), \mathbf{u}^*(t))$  is an  $\epsilon$ -optimal solution of the OCP (Equation (3-14)) with  $\mathbf{x}_0 = \mathbf{x}^*(0)$  and Lagrangian  $L$ . Therefore to apply the Moment-SOS hierarchy:

- (i) The unknown functions  $\varphi$  and  $L$  are approximated by polynomials in  $\mathbb{R}[t, \mathbf{x}]_{2d}$  and  $\mathbb{R}[\mathbf{x}, \mathbf{u}]_{2d}$ , where  $d$  is the parameter in the Moment-SOS hierarchy (Equation (3-6)).
- (ii) The above positivity constraint (Equation (3-20)) on  $[0, T] \times \mathbf{X} \times \mathbf{U}$  is replaced with a positivity certificate with degree bound on the SOS weights.
- (iii) Equation (3-21) is stated for every trajectory  $(\mathbf{x}(t; \mathbf{x}_\tau), \mathbf{u}(t, \mathbf{x}_\tau))$ ,  $t \in [\tau, T]$ , in the database. Using a discretization  $\{t_1, \dots, t_N\}$  of the interval  $[0, T]$ , the positivity constraints (Equation (3-21)) then become a set of linear constraints on the coefficients of the unknown polynomials  $\varphi$  and  $L$ .
- (iv)  $\epsilon$  in Equation (3-21) is now taken as a variable and one minimizes a criterion of the form  $\|L\|_1 + \gamma \epsilon$ , where  $\gamma > 0$  is chosen to balance between the sparsity-inducing norm  $\|L\|_1$  of the Lagrangian and the error  $\epsilon$  in the weak version of the optimality conditions (Equation (3-20)). A detailed discussion and related results can be found in [Pauwels, Henrion, and Lasserre \[2016\]](#).

**3.7 Optimal design in statistics.** In designing experiments one models the responses  $z_1, \dots, z_N$  of a random *experiment* whose inputs are represented by a vector  $\mathbf{t} = (t_i) \in \mathbb{R}^n$  with respect to known *regression functions*  $\Phi = (\varphi_1, \dots, \varphi_p)$ , namely:  $z_i = \sum_{j=1}^p \theta_j \varphi_j(t_i) +$

$\varepsilon_i, i = 1, \dots, N$ , where  $\theta_1, \dots, \theta_p$  are unknown parameters that the experimenter wants to estimate,  $\varepsilon_i$  is some noise and the  $(t_i)$ 's are chosen by the experimenter in a *design space*  $\mathfrak{X} \subseteq \mathbb{R}^n$ . Assume that the inputs  $t_i, i = 1, \dots, N$ , are chosen within a set of distinct points  $\mathbf{x}_1, \dots, \mathbf{x}_\ell \in \mathfrak{X}, \ell \leq \mathbb{N}$ , and let  $n_k$  denote the number of times the particular point  $\mathbf{x}_k$  occurs among  $t_1, \dots, t_N$ . A design  $\xi$  is then defined by:

$$(3-22) \quad \xi = \begin{pmatrix} \mathbf{x}_1 & \dots & \mathbf{x}_\ell \\ \frac{n_1}{\mathbb{N}} & \dots & \frac{n_\ell}{\mathbb{N}} \end{pmatrix}.$$

The matrix  $\mathbf{M}(\xi) := \sum_{i=1}^{\ell} \frac{n_i}{\mathbb{N}} \Phi(\mathbf{x}_i) \Phi(\mathbf{x}_i)^T$  is called the information matrix of  $\xi$ . Optimal design is concerned with finding a set of points in  $\mathfrak{X}$  that optimizes a certain statistical criterion  $\phi(\mathbf{M}(\xi))$ , which must be real-valued, positively homogeneous, non constant, upper semi-continuous, isotonic w.r.t. Loewner ordering, and concave. For instance in *D-optimal design* one maximizes  $\phi(\mathbf{M}(\xi)) := \log \det(\mathbf{M}(\xi))$  over all  $\xi$  of the form (Equation (3-22)). This is a difficult problem and so far most methods have used a discretization of the design space  $\mathfrak{X}$ .

The Moment-SOS hierarchy that we describe below does not rely on any discretization and works for an arbitrary compact basic semi-algebraic design space  $\mathfrak{X}$  as defined in Equation (1-2). Instead we look for an atomic measure on  $\mathfrak{X}$  (with finite support) and we proceed in two steps:

- In the first step one solves the hierarchy of convex optimization problems indexed by  $\delta = 0, 1, \dots$

$$(3-23) \quad \begin{aligned} \rho_\delta &= \sup_{\mathbf{y}} \{ \log \det(\mathbf{M}_d(\mathbf{y})) : y_0 = 1 \\ &\quad \mathbf{M}_{d+\delta}(\mathbf{y}) \succeq 0; \mathbf{M}_{d+\delta-d_j}(g_j \mathbf{y}) \succeq 0 \}, \end{aligned}$$

where  $d$  is fixed by the number of basis functions  $\varphi_j$  considered (here the monomials  $(\mathbf{x}^\alpha)_{\alpha \in \mathbb{N}_{2d}^n}$ ). (Note that Equation (3-23) is not an SDP because the criterion is not linear in  $\mathbf{y}$ , but it is still a tractable convex problem.) This provides us with an optimal solution  $\mathbf{y}^*(\delta)$ . In practice one chooses  $\delta = 0$ .

- In a second step we extract an atomic measure  $\mu$  from the “moments”  $\mathbf{y}^*(\delta)$ , e.g. via Nie’s method Nie [2014b] which consists of solving the SDP:

$$(3-24) \quad \begin{aligned} \rho_r &= \sup_{\mathbf{y}} \{ L_{\mathbf{y}}(f_r) : y_\alpha = y_\alpha^*(\delta), \forall \alpha \in \mathbb{N}_{2d}^n \\ &\quad \mathbf{M}_{d+r}(\mathbf{y}) \succeq 0; \mathbf{M}_{d+r-d_j}(g_j \mathbf{y}) \succeq 0 \}, \end{aligned}$$

where  $f_r$  is a (randomly chosen) polynomial strictly positive on  $\mathfrak{X}$ . If  $(y_\alpha^*(\delta))_{\alpha \in \mathbb{N}_{2d}^n}$  has a representing measure then it has an atomic representing measure, and generically the rank condition (Equation (2-10)) will be satisfied. Extraction of atoms is obtained via a linear

algebra routine. We have tested this two-steps method on several non-trivial numerical experiments (in particular with highly non-convex design spaces  $\mathcal{X}$ ) and in all cases we were able to obtain a design. For more details the interested reader is referred to [De Castro, Gamboa, Henrion, Hess, and Lasserre \[2017\]](#).

**Other applications & extensions.** In this partial overview, by lack of space we have not described some impressive success stories of the Moment-SOS hierarchy, e.g. in coding [Bachoc and Vallentin \[2008\]](#), packing problems in discrete geometry [de Laat and Vallentin \[2015\]](#) and [Schürmann and Vallentin \[2006\]](#). Finally, there is also a *non-commutative* version [Pironio, Navascués, and Acín \[2010\]](#) of the Moment-SOS hierarchy based on non-commutative positivity certificates [Helton and McCullough \[2004\]](#) and with important applications in quantum information [Navascués, Pironio, and Acín \[2008\]](#).

## 4 Conclusion

The list of important applications of the GPM is almost endless and we have tried to convince the reader that the Moment-SOS hierarchy is one promising powerful tool for solving the GPM with already some success stories. However much remains to be done as its brute force application does not scale well to the problem size. One possible research direction is to exploit symmetries and/or sparsity in large scale problems. Another one is to determine alternative positivity certificates which are less expensive in terms of computational burden to avoid the size explosion of SOS-based positivity certificates.

## References

- Christine Bachoc and Frank Vallentin (2008). “[New upper bounds for kissing numbers from semidefinite programming](#)”. *J. Amer. Math. Soc.* 21.3, pp. 909–924. MR: [2393433](#) (cit. on p. [3811](#)).
- Boaz Barak and David Steurer (2014). “Sum-of-squares proofs and the quest toward optimal algorithms”. In: *Proceedings of the International Congress of Mathematicians—Seoul 2014. Vol. IV*. Kyung Moon Sa, Seoul, pp. 509–533. MR: [3727623](#) (cit. on p. [3800](#)).
- Emmanuel J. Candès and Carlos Fernandez-Granda (2014). “[Towards a mathematical theory of super-resolution](#)”. *Comm. Pure Appl. Math.* 67.6, pp. 906–956. MR: [3193963](#) (cit. on p. [3804](#)).
- Vito Cerone, Dario Piga, and Diego Regruto (2012). “[Set-membership error-in-variables identification through convex relaxation techniques](#)”. *IEEE Trans. Automat. Control* 57.2, pp. 517–522. MR: [2918760](#) (cit. on p. [3808](#)).

- Yohann De Castro, F. Gamboa, Didier Henrion, and Jean B. Lasserre (2017). “Exact solutions to super resolution on semi-algebraic domains in higher dimensions”. *IEEE Trans. Inform. Theory* 63.1, pp. 621–630. MR: [3599963](#) (cit. on p. [3804](#)).
- Yohann De Castro, Fabrice Gamboa, Didier Henrion, Roxana Hess, and Jean B. Lasserre (2017). “Approximate Optimal Designs for Multivariate Polynomial Regression”. LAAS report No 17044. 2017, Toulouse, France. To appear in *Annals of Statistics*. arXiv: [1706.04059](#) (cit. on p. [3811](#)).
- Michel X. Goemans and David P. Williamson (1995). “Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming”. *J. Assoc. Comput. Mach.* 42.6, pp. 1115–1145. MR: [1412228](#) (cit. on p. [3792](#)).
- J. William Helton and Scott A. McCullough (2004). “A Positivstellensatz for non-commutative polynomials”. *Trans. Amer. Math. Soc.* 356.9, pp. 3721–3737. MR: [2055751](#) (cit. on p. [3811](#)).
- D. Henrion, Jean B. Lasserre, and C. Savorgnan (2009). “Approximate volume and integration for basic semialgebraic sets”. *SIAM Rev.* 51.4, pp. 722–743. MR: [2563831](#) (cit. on p. [3803](#)).
- Didier Henrion and Milan Korda (2014). “Convex computation of the region of attraction of polynomial control systems”. *IEEE Trans. Automat. Control* 59.2, pp. 297–312. MR: [3164876](#) (cit. on p. [3808](#)).
- Didier Henrion and Jean B. Lasserre (2006). “Convergent relaxations of polynomial matrix inequalities and static output feedback”. *IEEE Trans. Automat. Control* 51.2, pp. 192–202. MR: [2201707](#) (cit. on p. [3808](#)).
- Didier Henrion, Jean B. Lasserre, and Johan Löfberg (2009). “GloptiPoly 3: moments, optimization and semidefinite programming”. *Optim. Methods Softw.* 24.4-5, pp. 761–779. MR: [2554910](#) (cit. on p. [3798](#)).
- Cédric Josz, Jean B. Lasserre, and Bernard Mourrain (Aug. 2017). “Sparse polynomial interpolation: compressed sensing, super resolution, or Prony?” LAAS Report no 17279. 2017, Toulouse, France. arXiv: [1708.06187](#) (cit. on p. [3805](#)).
- Subhash Khot (2010). “Inapproximability of NP-complete problems, discrete Fourier analysis, and geometry”. In: *Proceedings of the International Congress of Mathematicians. Volume IV*. Hindustan Book Agency, New Delhi, pp. 2676–2697. MR: [2827989](#) (cit. on p. [3800](#)).
- (2014). “Hardness of approximation”. In: *Proceedings of the International Congress of Mathematicians—Seoul 2014. Vol. 1*. Kyung Moon Sa, Seoul, pp. 711–728. MR: [3728489](#) (cit. on p. [3793](#), [3800](#)).
- Etienne de Klerk, Jean B. Lasserre, Monique Laurent, and Zhao Sun (2017). “Bound-constrained polynomial optimization using only elementary calculations”. *Math. Oper. Res.* 42.3, pp. 834–853. MR: [3685268](#).

- J.-L. Krivine (1964a). “Anneaux préordonnés”. *J. Analyse Math.* 12, pp. 307–326. MR: [0175937](#) (cit. on p. [3796](#)).
- Jean-Louis Krivine (1964b). “Quelques propriétés des préordres dans les anneaux commutatifs unitaires”. *C. R. Acad. Sci. Paris* 258, pp. 3417–3418. MR: [0169083](#) (cit. on p. [3796](#)).
- David de Laat and Frank Vallentin (2015). “A semidefinite programming hierarchy for packing problems in discrete geometry”. *Math. Program.* 151.2, Ser. B, pp. 529–553. MR: [3348162](#) (cit. on p. [3811](#)).
- Henry J Landau (1987). *Moments in mathematics*. Vol. 37. Proc. Sympos. Appl. Math. (cit. on p. [3793](#)).
- Jean B. Lasserre (2000). “Optimisation globale et théorie des moments”. *C. R. Acad. Sci. Paris Sér. I Math.* 331.11, pp. 929–934. MR: [1806434](#) (cit. on pp. [3792](#), [3797](#), [3798](#)).
- (2002a). “Bounds on measures satisfying moment conditions”. *Ann. Appl. Probab.* 12.3, pp. 1114–1137. MR: [1925454](#) (cit. on p. [3802](#)).
  - (2002b). “Semidefinite programming vs. LP relaxations for polynomial programming”. *Math. Oper. Res.* 27.2, pp. 347–360. MR: [1908532](#) (cit. on pp. [3797](#), [3800](#)).
  - (2006). “Convergent SDP-relaxations in polynomial optimization with sparsity”. *SIAM J. Optim.* 17.3, pp. 822–843. MR: [2257211](#) (cit. on p. [3799](#)).
  - (2010). *Moments, positive polynomials and their applications*. Vol. 1. Imperial College Press Optimization Series. Imperial College Press, London, pp. xxii+361. MR: [2589247](#) (cit. on p. [3795](#)).
  - (2013). “Inverse polynomial optimization”. *Math. Oper. Res.* 38.3, pp. 418–436. MR: [3092539](#) (cit. on p. [3808](#)).
  - (2015a). *An introduction to polynomial and semi-algebraic optimization*. Cambridge Texts in Applied Mathematics. Cambridge University Press, Cambridge, pp. xiv+339. MR: [3469431](#) (cit. on pp. [3797](#), [3799](#), [3800](#)).
  - (2015b). “Tractable approximations of sets defined with quantifiers”. *Math. Program.* 151.2, Ser. B, pp. 507–527. MR: [3348161](#) (cit. on p. [3807](#)).
  - (2017). “Computing Gaussian & exponential measures of semi-algebraic sets”. *Adv. in Appl. Math.* 91, pp. 137–163. MR: [3673583](#) (cit. on p. [3803](#)).
  - (2000/01). “Global optimization with polynomials and the problem of moments”. *SIAM J. Optim.* 11.3, pp. 796–817. MR: [1814045](#) (cit. on pp. [3792](#), [3797](#)).
- Jean B. Lasserre, Didier Henrion, Christophe Prieur, and Emmanuel Trélat (2008). “Non-linear optimal control via occupation measures and LMI-relaxations”. *SIAM J. Control Optim.* 47.4, pp. 1643–1666. MR: [2421324](#) (cit. on p. [3806](#)).
- Jean B. Lasserre, Monique Laurent, and Philipp Rostalski (2008). “Semidefinite characterization and computation of zero-dimensional real radical ideals”. *Found. Comput. Math.* 8.5, pp. 607–647. MR: [2443091](#).



- Jean-Paul Laumond, Nicolas Mansard, and Jean B. Lasserre, eds. (2017a). *Geometric and numerical foundations of movements*. Vol. 117. Springer Tracts in Advanced Robotics. Springer, Cham, pp. x+419. MR: [3642945](#) (cit. on pp. [3799](#), [3808](#)).
- eds. (2017b). *Geometric and numerical foundations of movements*. Vol. 117. Springer Tracts in Advanced Robotics. Springer, Cham, pp. x+419. MR: [3642945](#) (cit. on p. [3809](#)).
- Monique Laurent (2003). “A comparison of the Sherali-Adams, Lovász-Schrijver, and Lasserre relaxations for 0-1 programming”. *Math. Oper. Res.* 28.3, pp. 470–496. MR: [1997246](#) (cit. on p. [3800](#)).
- (2009). “Sums of squares, moment matrices and optimization over polynomials”. In: *Emerging applications of algebraic geometry*. Vol. 149. IMA Vol. Math. Appl. Springer, New York, pp. 157–270. MR: [2500468](#) (cit. on p. [3795](#)).
- Daniel K Molzahn and Ian A Hiskens (2015). “Sparsity-exploiting moment-based relaxations of the optimal power flow problem”. *IEEE Transactions on Power Systems* 30.6, pp. 3168–3180 (cit. on p. [3799](#)).
- Miguel Navascués, Stefano Pironio, and Antonio Acín (2008). “A convergent hierarchy of semidefinite programs characterizing the set of quantum correlations”. *New Journal of Physics* 10.7, p. 073013 (cit. on p. [3811](#)).
- Yurii Nesterov (2000). “Squared functional systems and optimization problems”. In: *High performance optimization*. Vol. 33. Appl. Optim. Kluwer Acad. Publ., Dordrecht, pp. 405–440. MR: [1748764](#) (cit. on p. [3792](#)).
- Jiawang Nie (2013). “Certifying convergence of Lasserre’s hierarchy via flat truncation”. *Math. Program.* 142.1-2, Ser. A, pp. 485–510. MR: [3127083](#) (cit. on pp. [3799](#), [3805](#)).
- (2014a). “Optimality conditions and finite convergence of Lasserre’s hierarchy”. *Math. Program.* 146.1-2, Ser. A, pp. 97–121. MR: [3232610](#) (cit. on pp. [3798](#), [3799](#)).
- (2014b). “The  $\mathcal{A}$ -truncated  $K$ -moment problem”. *Found. Comput. Math.* 14.6, pp. 1243–1276. MR: [3273678](#) (cit. on p. [3810](#)).
- Pablo A. Parrilo (2000). “Structured Semidefinite Programs and Semialgebraic Geometry Methods in Robustness and Optimization”. PhD thesis. California Institute of Technology (cit. on p. [3792](#)).
- (2003). “Semidefinite programming relaxations for semialgebraic problems”. *Math. Program.* 96.2, Ser. B. Algebraic and geometric methods in discrete optimization, pp. 293–320. MR: [1993050](#) (cit. on pp. [3792](#), [3808](#)).
- Edouard Pauwels, Didier Henrion, and Jean B. Lasserre (2016). “Linear conic optimization for inverse optimal control”. *SIAM J. Control Optim.* 54.3, pp. 1798–1825. MR: [3516862](#) (cit. on p. [3809](#)).
- S. Pironio, M. Navascués, and A. Acín (2010). “Convergent relaxations of polynomial optimization problems with noncommuting variables”. *SIAM J. Optim.* 20.5, pp. 2157–2180. MR: [2650843](#) (cit. on p. [3811](#)).

- Michael Posa, Mark Tobenkin, and Russ Tedrake (2016). “[Stability analysis and control of rigid-body systems with impacts and friction](#)”. *IEEE Trans. Automat. Control* 61.6, pp. 1423–1437. MR: [3508689](#) (cit. on p. [3808](#)).
- Mihai Putinar (1993). “[Positive polynomials on compact semi-algebraic sets](#)”. *Indiana Univ. Math. J.* 42.3, pp. 969–984. MR: [1254128](#) (cit. on p. [3796](#)).
- Konrad Schmüdgen (2017). *The moment problem*. Vol. 277. Graduate Texts in Mathematics. Springer, Cham, pp. xii+535. MR: [3729411](#) (cit. on p. [3795](#)).
- Achill Schürmann and Frank Vallentin (2006). “[Computational approaches to lattice packing and covering problems](#)”. *Discrete Comput. Geom.* 35.1, pp. 73–116. MR: [2183491](#) (cit. on p. [3811](#)).
- Naum Z. Shor (1998). *Nondifferentiable optimization and polynomial problems*. Vol. 24. Nonconvex Optimization and its Applications. Kluwer Academic Publishers, Dordrecht, pp. xviii+394. MR: [1620179](#) (cit. on p. [3792](#)).
- F.-H. Vasilescu (2003). “Spectral measures and moment problems”. In: *Spectral analysis and its applications*. Vol. 2. Theta Ser. Adv. Math. Theta, Bucharest, pp. 173–215. MR: [2082433](#) (cit. on p. [3796](#)).
- Hayato Waki, Sunyoung Kim, Masakazu Kojima, and Masakazu Muramatsu (2006). “[Sums of squares and semidefinite program relaxations for polynomial optimization problems with structured sparsity](#)”. *SIAM J. Optim.* 17.1, pp. 218–242. MR: [2219151](#) (cit. on p. [3799](#)).

Received 2017-11-06.

JEAN B. LASSERRE  
[lasserre@laas.fr](mailto:lasserre@laas.fr)



# A $\mathcal{UU}$ -POINT OF VIEW OF NONSMOOTH OPTIMIZATION

CLAUDIA SAGASTIZÁBAL

## Abstract

The realization that many nondifferentiable functions exhibit some form of structured nonsmoothness has been attracting the efforts of many researchers in the last decades. Identifying theoretically and computationally certain manifolds where a nonsmooth function behaves smoothly poses challenges for the nonsmooth optimization community. We review a sequence of milestones in the area that led to the development of algorithms of the bundle type that can track the region of smoothness and mimic a Newton algorithm to converge with superlinear speed. The new generation of bundle methods is sufficiently versatile to deal with structured objective functions, even when the available information is inexact.

## 1 What is at stake in nonsmooth optimization

In 2008 the American magazine *Wired* published in its Science section an article entitled *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*. The author, Chris Anderson, argued that when confronted to massive data theory was no longer meaningful because in the “Petabyte Age” the traditional approach to science –hypothesize, model, test– had become obsolete. Quoting from the publication,

There is now a better way. Petabytes allow us to say: “Correlation is enough.” We can stop looking for models. We can analyze the data without hypotheses about what it might show. We can throw the numbers into the biggest computing clusters the world has ever seen and let statistical algorithms find patterns where science cannot.

---

Partially supported by CNPq Grant 303905/2015-8 and FAPERJ, Brazil.

MSC2010: primary 90C25; secondary 65K05, 49M15, 46N10.

Keywords: Nonsmooth convex optimization, bundle methods, proximal-point mapping, Moreau-Yosida regularization,  $\mathcal{UU}$ -space decomposition.

While it is true that nowadays the Big Data concept pervades many scientific domains, in Mathematical Optimization “theory” is still thriving, alive and kicking. Rather than replacing the theoretical methodology, Big Data has in fact enriched the field with a whole new set of paradigms and technical tools necessary to deal with the “Petabyte” issues.

In Nonsmooth Optimization (NSO) there are situations in which it is not affordable to drop the theory and stop looking for models, as the article advocates. In particular, to develop sound solution algorithms statistical measures and heuristics need to be complemented with convergence proofs and optimality criteria. In this respect, *models* are crucial: they yield a constructive mechanism to estimate the distance to the solution set, certifying the quality of the solution approach and ensuring that the iterative process returns a reliable numerical solution.

Consider the problem of minimizing over the whole space a finite-valued convex function  $f : \mathbb{R}^n \rightarrow \mathbb{R}$ . Let  $x \in \mathbb{R}^n$  be a given point and recall the Convex Analysis subdifferential definition

$$\partial f(x) = \{g \in \mathbb{R}^n : f(y) \geq f(x) + \langle g, y - x \rangle \text{ for all } y \in \mathbb{R}^n\},$$

where  $\langle g, x \rangle$  denotes the Euclidean inner product between  $g$  and  $x$  in  $\mathbb{R}^n$ . Since the function  $f$  is convex, a minimizer is characterized by the generalized Fermat condition

$$(1-1) \quad 0 \in \partial f(\bar{x}).$$

For convergence analysis purposes the iterative process must ensure the inclusion is eventually satisfied. On the other hand, for algorithmic purposes a fundamental related question is:

### **What is a sound stopping test for a NSO method?**

As explained in [Section 2.2](#) such a matter is not as straightforward as it may look at first glance. The answer is achieved by building elements in a continuous *model* for the subdifferential, the  $\varepsilon$ -subdifferential, depending on a parameter  $\varepsilon \geq 0$ :

$$\partial_\varepsilon f(x) = \{g \in \mathbb{R}^n : f(y) \geq f(x) + \langle g, y - x \rangle - \varepsilon \text{ for all } y \in \mathbb{R}^n\}.$$

Continuity here is understood in a Set-Valued Analysis sense, both on  $x$  and  $\varepsilon$ .

Like with smooth optimization methods, the NSO stopping test relies on properties guaranteeing asymptotic global convergence: for any given starting point the sequence has an accumulation point satisfying the Fermat condition.

Having resolved the issue of global convergence another very important matter refers to local convergence, or *speed*. Here arises a second fundamental question:

### **Is it possible for a NSO algorithm to converge with superlinear rate?**

We review in [Section 4](#) the main elements regarding this very difficult question, whose answer required the continued efforts of many researchers for at least thirty years. As with the first question, the response was also found by defining suitable *models*, only that now the object of interest is certain *manifold of smoothness* along which Newton-like steps are possible. This, even if the function is nonsmooth and a subdifferential set (instead of just one gradient) describes its first-order behaviour.

As observed in [A. S. Lewis \[2014\]](#), Variational Analysis has come of age and its computational aspects have significantly grown in the last years. This work continues that line of thinking, reviewing advances in the area and pointing out some open problems. Rather than making an exhaustive list, which is clearly an impossible task, the focus is put on transmitting the main ideas behind selected topics that we believe reflect well the state of the art in the field.

To make the presentation palatable to non-expert readers, the emphasis is put on outlining key points, keeping a general view without entering into technicalities, to the extent that this is possible in Mathematics.

## 2 A brief history of models in NSO

Pursuing further our claim that models cannot be just dismissed when it comes to Optimization, we now describe how some of the more fundamental challenges have been addressed by a particular class of NSO algorithms, known as bundle methods, [Hiriart-Urruty and Lemaréchal \[1993\]](#), [Bonnans, Gilbert, Lemaréchal, and Sagastizábal \[2006\]](#).

**2.1 Oracle information.** Designing a NSO method means to define iterates  $x^k$  eventually solving the Fermat inclusion. There are various possibilities to ensure that (1-1) holds asymptotically, depending on how much information is available on the function to be minimized. Typically, the function  $f$  is not known in a closed form but the functional value and one subgradient can be computed for each given  $x^k \in \mathbb{R}^n$ . Such information is provided in the form of an *oracle*, a procedure coded by the user independently of the algorithm designer.

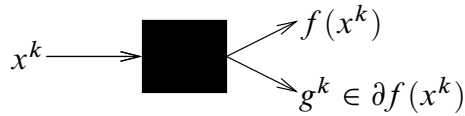
Initially oracles were assumed to yield exact values, namely  $f(x^k)$  and one subgradient  $g(x^k) \in \partial f(x^k)$ , as illustrated by [Figure 1](#).

Since there is no control on which particular subgradient in the subdifferential is provided as an output, such oracles are said to be of the black-box type<sup>1</sup>. Modern oracles, delivering inexact information, will be considered in [Section 5.1](#). We just mention here that inexact information in bundle methods was best handled after K. Kiwiel introduced

---

<sup>1</sup>Be aware that in machine learning the wording “black-box” is used to refer to an oracle making *inexact* calculations, similar to the ones described in [Section 5.1](#).

Figure 1: NSO black-box oracle



the notion of noise attenuation in [K. C. Kiwiel \[2006\]](#); see also [M. V. Solodov \[2003\]](#), [Hintermüller \[2001\]](#).

The importance of suitably modelling in NSO is demonstrated below by a couple of algorithmic highlights that represented a breakthrough in the field.

**2.2 Modelling the subdifferential.** When introduced in the late 1970s, bundle methods [Lemaréchal and Mifflin \[1978\]](#) collected the black-box output of past iterations to create a model for the subdifferential that was enriched along iterations, say  $k = 1, 2, \dots$ ,

$$\partial f(\bar{x}) \approx \text{conv} \{g^i \in \partial f(x^i), i = 1, \dots, k\},$$

where  $\text{conv } S$  stands for the convex hull of a set  $S$ . The Fermat inclusion was approximated by a quadratic programming (QP) problem, yielding the smallest element of such subdifferential model. The resulting convex combination of subgradients, denoted by  $\hat{g}^k$  and called *aggregate gradient*, gives a direction along which a line search defines the next iterate.

The forerunner subgradient methods [Shor \[1970\]](#) take an even simpler model, since they can be seen as making the crude approximation  $\partial f(\bar{x}) \approx \{g^k\}$ . With the exception of the dilation algorithms (a class including the well-known ellipsoid method by L. Khachiyan), subgradient methods keep no memory of the past; every point is as if the first one in the iterative process. This is in sheer contrast with the *bundle* of past information that is collected in the bundle methodology to define its sequence of iterates.

Regarding the first question, on how to stop iterations in a NSO method, the transportation formula introduced by [Lemaréchal \[1980\]](#),

$$g^i \in \partial f(x^i) \implies g^i \in \partial_{e^i(x^k)} f(x^k) \quad e_i(x^i) = f(x^k) - f(x^i) - \langle g^i, x^k - x^i \rangle \geq 0$$

proved fundamental. Indeed, thanks to this relation, when solving the QP problem to compute the aggregate gradient a convex combination of terms  $e_i(x^k)$ , denoted by  $\hat{e}^k$  and called *aggregate error*, is also available and the inclusion

$$(2-1) \quad \hat{g}^k \in \partial_{\hat{e}^k} f(x^k)$$

gives a certificate for approximate optimality. Namely, if the aggregate error and the norm of the aggregate gradient are both sufficiently small, the inclusion (1-1) is satisfied up to certain tolerance. Furthermore, if the aggregate error and gradient associated with some bounded subsequence of the iterates  $\{x^k\}$  converge to zero, then any accumulation point of such subsequence satisfies the Fermat condition, as desired.

Without having at hand a model for the subdifferential and without the transportation formula, designing an optimality certificate does not seem possible. Consider once more one iteration of a subgradient method and recall its “amnesic” feature. For the algorithm designer to define an optimality certificate at the  $k$ th iteration, the only relation at hand is  $g^k \in \partial f(x^k)$ . In this case satisfaction of (1-1) in the limit is only possible if  $g^k \rightarrow 0$  as  $k \rightarrow \infty$ . This is too strong of a requirement in a black-box oracle context. It has no reason to hold for the simplest nonsmooth convex function,  $f(x) = |x|$  for  $x \in \mathbb{R}$ , not even in the best of circumstances, that is when  $x^k = 0$  is the minimizer; the output of the black-box can be *any*  $g^k \in [-1, 1]$ , instead of the “right” subgradient  $g^k = 0$ .

The lack of a reliable stopping test, typical of subgradient methods, explains the development of complexity analysis for those algorithms. Such results estimate the number of iterations that takes for the method to achieve a given accuracy in the worst case. Complexity estimates are pessimistic by nature and for this reason they are not often used in practical implementations, especially if accuracy of the numerical solution is a concern.

Back to the first generation of bundle methods, in view of (2-1), an exogenous parameter  $\varepsilon^k \rightarrow 0$  was introduced in the QP problem to bound from above the convex sum of errors and drive the aggregate error to zero.

The challenge of making the aggregate gradient asymptotically null was resolved by a brilliant new idea. In the late 1970s the main source of inspiration was smooth Nonlinear Programming, a field that was in full bloom with the quasi-Newton approaches in those times. It was then natural to try to follow the nonlinear programming path, defining a descent direction first and then performing a line search. At this stage came the realization that, unlike the smooth case, even for a convex function there is no guarantee that the direction obtained from a black-box subgradient will provide descent (the subdifferential is not a continuous set-valued mapping).

One crucial innovation brought by bundle methods to NSO is precisely the introduction of a mechanism separating iterates into two subsequences, respectively called of *serious* and *null* steps. Serious steps, usually denoted by  $\hat{x}^k$ , are reference points giving sufficient reduction of the objective function. This *descent* subsequence has functional values that decrease monotonically to a minimum (if such a minimum exists, of course). Null steps are all the other iterates, they were named in this way because they do not contribute to reduce the functional value. Their role is to supply the oracle information and help building a richer model, so that eventually a new serious steps can be found.



These considerations marked a departing point from the classical convergence results in smooth optimization: splitting the iterates into two subsequences leads to a special theory of convergence because the analysis must be split accordingly. When the subsequence of serious step is infinite, the standard smooth optimization results can be mimicked (in particular, the serious aggregate gradient subsequence tends to zero). The second case, in which there is a last serious step followed by an infinite number of null steps, is more involved and typically relies on properties of the proximal point method introduced by Moreau [1965]. This issue, addressed in the next subsection, is again related to the introduction of certain *model*, now of primal nature.

**2.3 Modelling the objective function.** Calling the oracle at  $x^k$  yields a *linearization* of  $f$

$$\ell^k(x) := f(x^k) + \langle g^k, x - x^k \rangle \leq f(x),$$

which is a lower approximation of the function, tangent to  $f$  at  $x^k$  by convexity. The principle of collecting in a bundle of information the output of past oracle iterations can then be used to define a cutting-plane model for the function:

$$f(x) \approx m^k(x) \quad \text{for} \quad m^k(x) = \max\{\ell^i(x), i = 1, \dots, k\}.$$

If the next iterate minimizes the model, the corresponding optimality condition,  $0 \in \partial m^k(x^{k+1})$ , is close to the one for the QP problem in the dual bundle methods in [Section 2.2](#). This is the basis of the cutting-plane method [Cheney and Goldstein \[1959\]](#) and [Kelley \[1960\]](#) which, albeit convergent, does not distinguish between serious and null iterates, as bundle methods do. As a result, functional values start oscillating, to an extent that in some cases the whole process stalls, due to tailing-off effects and numerical errors.

In the second generation of bundle methods, developed close to the end of the 1980s, modelling switches from a dual to a primal point of view. Instead of approaching solely the subdifferential, the objective function is also modelled along iterations. The cutting-plane model is stabilized in a manner ensuring the next iterate remains sufficiently close to  $\hat{x}^k$ , the last generated serious step. In particular, the proximal variants of bundle methods [K. C. Kiwiel \[1990\]](#) prevent oscillations by solving a QP problem whose objective function is the model

$$m^k(x) + \frac{1}{2}\mu\|x - \hat{x}^k\|_2^2,$$

depending on a parameter  $\mu > 0$ . The unique minimizer of the stabilized model gives the next iterate, which becomes the next reference point only if its functional value is sufficiently smaller than a target  $\tau^k$ , as follows:

$$\hat{x}^{k+1} := x^{k+1} \quad \text{when} \quad f(x^{k+1}) \leq \tau^k := f(\hat{x}^k) - m\delta^k.$$

In these relations,  $m \in (0, 1)$  is a parameter and the expected decrease  $\delta^k \geq 0$  is computed when solving the QP problem. If the inequality is not satisfied, a null step is declared. In both cases, the next cutting-plane model is enriched with the linearization defined with the oracle information at  $x^{k+1}$ .

The name of the variant is explained by recalling that the proximal point operator of a convex function  $f$  at a given  $x \in \mathbb{R}^n$  is defined by

$$p^f(x) := \operatorname{argmin}_{p \in \mathbb{R}^n} \left\{ f(p) + \frac{1}{2} \mu \|p - x\|_2^2 \right\}.$$

The point  $x$  above is referred to as the *prox-center*. Writing down the optimality condition gives the equivalence

$$0 \in \partial f(p^f(x)) + \mu(p^f(x) - x) \iff p^f(x) = x - \frac{1}{\mu} g^p \text{ for some } g^p \in \partial f(p^f(x)),$$

where the subgradient in the identity has an implicit nature. As a result, computing the proximal point of  $f$  in the black-box oracle context is not possible because the function  $f$  is not available analytically. However, it was proved in [Correa and Lemaréchal \[1993\]](#) that

$$(2-2) \quad \lim_{k \rightarrow \infty} p^{\mathfrak{m}^k}(\hat{x}) = \lim_{k \rightarrow \infty} \operatorname{argmin}_{p \in \mathbb{R}^n} \{ \mathfrak{m}^k(p) + \frac{1}{2} \mu \|p - \hat{x}\|_2^2 \} = p^f(\hat{x}),$$

see also [Auslender \[1987\]](#), [Fukushima \[1984\]](#). This nice result shows that enriching the cutting-plane model without moving the prox-center eventually gives the proximal point for  $f$ .

The importance of (2-2) is better understood by recalling a well-known characterization for minimizers of a convex function, alternative to the Fermat condition (1-1). This is the statement that  $\bar{x}$  minimizes  $f$  if and only if it is a fixed point of the proximal point operator for any  $\mu > 0$ :

$$\bar{x} = p^f(\bar{x}).$$

By construction, in the proximal bundle method the next iterate is the proximal point of the cutting-plane model at the reference point:

$$x^{k+1} = p^{\mathfrak{m}^k}(\hat{x}^k).$$

In particular, when the iterate is declared a serious step, since  $\hat{x}^{k+1} = x^{k+1}$ , the recursion

$$\hat{x}^{k+1} = p^{\mathfrak{m}^k}(\hat{x}^k)$$

is satisfied. Because of (2-2) this is nothing but an implementable form of a Picard iterative process to find the fixed point of  $p^f(\cdot)$ . Interpreting the bundle iterations in the context

of (2-2) confirms the perception that null steps contribute to enrich the model, improving its accuracy until a new serious step is found.

It is important to be aware that any statement regarding superlinear speed of a NSO algorithm refers to the rate of convergence of the subsequence of serious steps *only*. With a black-box oracle there is no control on how fast the proximal point of the model converges to the proximal point of the function in the relations (2-2). No result is known for the speed of convergence of the null step subsequence.

**Bibliographical note.** The very first bundle methods can be tracked back to Lemaréchal [1975] and Mifflin [1977]; see also K. Kiwiel [1985]. A second wave came after more than ten years of research with the trust-region Schramm and Zowe [1992], level Lemaréchal, Nemirovskii, and Nesterov [1995], and variable metric Lemaréchal and Sagastizábal [1997b] bundle variants; see also Lukšan and Vlček [1999]. The years 2000 brought an asymptotically exact level algorithm Fábíán [2000], the spectral method Helmberg and K. Kiwiel [2002], the generalized bundle Frangioni [2002] and the limited memory algorithm Haarala, Miettinen, and Makela [2004] tailored to tackle large-scale problems.

More recently the proximal Chebyshev method Ouorou [2013], the doubly stabilized bundle method de Oliveira and M. Solodov [2016], the target radius algorithm de Oliveira [2017], and Apkarian, Noll, and Ravanbod [2016], a trust-region bundle algorithm declined for application in optimal control, are a proof of the continued interest on the subject of researchers in the NSO area.

### 3 Speeding up the method

Picard iterations can be slow to converge and the subsequence of serious steps has at best an R-linear rate Robinson [1999]. To accelerate the bundle variants the proximal parameter  $\mu$  needs to be suitably updated at each iteration (the more correct notation  $\mu^k$  was discarded in this text, to alleviate the reading).

The characterization considered below for a minimizer  $\bar{x}$  reveals useful for such purposes, in fact it was a key to defining superlinearly convergent NSO algorithms.

**3.1 A not so regular regularization.** Given a positive semi-definite matrix  $H$ , the Moreau-Yosida regularization of  $f$  has the expression

$$(3-1) \quad F(x) := \inf_{p \in \mathbb{R}^n} \left\{ f(p) + \frac{1}{2} \langle H(p - x), p - x \rangle \right\}.$$

The original regularization (Moreau [1965], Yosida [1964]) was defined for positive definite matrices. The slightly more general case considered here enjoys similar properties,

listed below and extracted from [Lemaréchal and Sagastizábal \[1994\]](#). In the theorem,  $\text{dom } f^*$  denotes the domain of the conjugate function, that is the set of  $g \in \mathbb{R}^n$  where  $\sup_{x \in \mathbb{R}^n} \{ \langle g, x \rangle - f(x) \}$  is finite.

**Theorem 3.1** (Moreau-Yosida regularization properties). *If  $\text{dom } f^* \cap \text{Im } H \neq \emptyset$  then  $F(x) > -\infty$  for all  $x$ , and  $F$  is a convex function defined on the whole of  $\mathbb{R}^n$ . Suppose, in addition, that for all  $x$  the infimum in (3-1) is attained on some nonempty set  $P(x)$ . Then the convex function  $F$  has at all  $x$  a gradient given by*

$$(3-2) \quad \nabla F(x) = H(x - p), \quad \text{for arbitrary } p \in P(x),$$

which is Lipschitzian with constant equal to  $\Lambda$ , the largest eigenvalue of  $H$ . More precisely, for all  $x, x'$ :

$$\|\nabla F(x) - \nabla F(x')\|_2^2 \leq \Lambda \langle \nabla F(x) - \nabla F(x'), x - x' \rangle.$$

Furthermore, minimizing  $f$  is equivalent to minimizing  $F$ .

Note in passing that minimizing  $F$  is as difficult, if not more, than minimizing  $f$ . Having the black-box oracle yields a cutting-plane model  $m^k$  for  $f$ , so the information available for the regularization cannot be exact and corresponds to the inexact oracle

$$F(\hat{x}^k) \approx m^k(x^{k+1}) + \frac{1}{2}\mu \|x^{k+1} - \hat{x}^k\|_2^2 \quad \text{and} \quad \nabla F(\hat{x}^k) \approx \mu(\hat{x}^k - x^{k+1}).$$

Another remark is the relation between the contraction factor of the fixed point iterations and the value of  $\Lambda$ : the smaller the maximum eigenvalue of  $H$ , the faster will be the Picard process. This observation confirms the interest of moving the Moreau-Yosida (pseudo) metric along iterations. Actually, when applied with  $H = \mu I$ , a scalar multiple of the identity matrix, the relations in (3-2) give an enlightening interpretation of the Picard iterations:

$$\begin{aligned} \hat{x}^{k+1} = p_\mu^f(\hat{x}^k) &\iff \hat{x}^{k+1} = \hat{x}^k + \left( p_\mu^f(\hat{x}^k) - \hat{x}^k \right) \\ &= \hat{x}^k - \frac{1}{\mu} \mu \left( \hat{x}^k - p_\mu^f(\hat{x}^k) \right) \\ &= \hat{x}^k - \frac{1}{\mu} \nabla F(\hat{x}^k) \\ &= \hat{x}^k - H^{-1} \nabla F(\hat{x}^k). \end{aligned}$$

The proximal point method is in fact a preconditioned gradient method to minimize the (smooth) Moreau-Yosida regularization.

This new insight brought back to life the hope from the end of the 1970s, that designing NSO quasi-Newton schemes might be possible. The hope was not vain: since  $F$  has

a Lipschitzian gradient, applying a Newton-like scheme “just” needed for the Moreau-Yosida regularization to have an invertible Hessian in a ball about a minimizer, or at least at  $\bar{x}$ .

The second-order study in [Lemaréchal and Sagastizábal \[1997a\]](#) gave a negative answer to this issue, if the whole space is considered. Specifically, for any closed convex function  $f$  satisfying a second order growth condition the Moreau-Yosida regularization has a Hessian at every  $x \in \mathbb{R}^n$  if and only if the function  $f$  has a Hessian everywhere. In other words, if a Newton-like scheme can be applied to  $F$ , it can as well be applied directly to the original function  $f$ !

This direction appeared unsuccessful but in fact it was not pointless, as revealed by the following example.

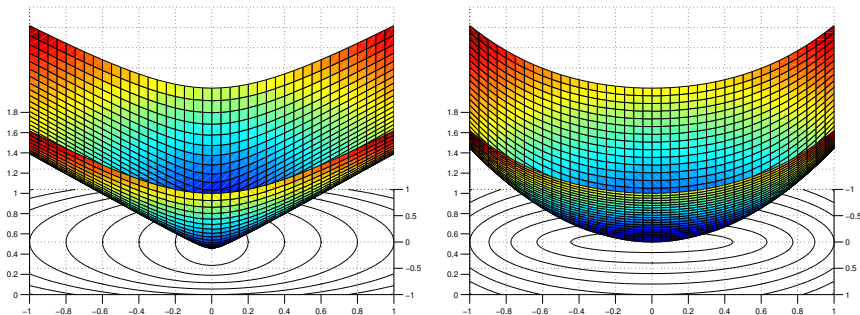
**3.2 A  $\mathcal{V}\mathcal{U}$  function.** Given a positive scalar  $a$ , consider the minimization in  $\mathbb{R}^2$  of

$$f(v, u) = f^1(v) + f^2(u) \quad \text{for } f^1(v) = |v| \text{ and } f^2(u) = \frac{a}{2}u^2.$$

The function  $f$  is differentiable everywhere except at the points with null first component:  $\partial f(v, u) = \text{sign}(v) \times \{au\}$ , with the convention that  $\text{sign}(v) = [-1, 1]$  if  $v = 0$ . Such is the case of the unique minimizer,  $\bar{x} = (0, 0)$ .

[Figure 2](#) shows two views of this bivariate function, on the left for fixed second component  $f(\cdot, u^{fix})$  and on the right when fixing the first component  $f(v^{fix}, \cdot)$ .

Figure 2: Two views of a simple  $\mathcal{V}\mathcal{U}$ -function



For this simple, sufficiently structured function, several calculations related to the Moreau-Yosida regularization can be easily performed explicitly. The separability of the

function is inherited by the proximal point,

$$p^f(v, u) = p^{f^1}(v) + p^{f^2}(u) \quad \text{with } p^{f^1}(v) = v - P_\mu(v) \text{ and } p^{f^2}(u) = \frac{\mu}{a + \mu}u,$$

where  $P_\mu(v)$  is the projection of  $v$  onto the interval  $[-\frac{1}{\mu}, \frac{1}{\mu}]$ .

The Hessian of the Moreau-Yosida regularization at a point  $x^\circ$  exists if and only if the proximal point operator has a Jacobian  $\nabla p(x^\circ)$  at  $x^\circ$ , and the equality  $\nabla^2 F(x^\circ) = H(I - \nabla p(x^\circ))$  holds, (Lemaréchal and Sagastizábal [ibid., Prop. 2.4]).

Back to our simple function, take  $x^\circ = (v^\circ, u^\circ)$  with first component satisfying  $|v^\circ| < \frac{1}{\mu}$ , so that  $x^\circ = (v^\circ, u^\circ)$  is close to  $\bar{x}$ . The corresponding Moreau-Yosida objects are

$$p^\circ := p^f(x^\circ) = \begin{pmatrix} 0 \\ \frac{\mu}{a+\mu}u^\circ \end{pmatrix}, \quad \nabla F(x^\circ) = \begin{pmatrix} \mu v^\circ \\ \frac{a\mu}{a+\mu}u^\circ \end{pmatrix}, \quad \text{and } \nabla^2 F(x^\circ) = \begin{pmatrix} \mu & 0 \\ 0 & \frac{a\mu}{a+\mu} \end{pmatrix}.$$

The Moreau-Yosida regularization has a Hessian at  $x^\circ$ , in particular at  $\bar{x}$ . The function  $f$  does not even have a gradient at  $\bar{x}$ , let alone a Hessian. Notwithstanding, the right graph in Figure 2 looks U-shaped, which indicates the function is sufficiently smooth to indeed have some sort of second-order object, reminiscent of a Hessian.

This intriguing observation raised the following question:

### Is it possible to find a region where a nonsmooth function behaves smoothly?

The answer to this point was explored by several authors in the 2000s, adopting different perspectives. There was a Convex Analysis viewpoint, with the  $\mathcal{U}$ -Lagrangian Lemaréchal, Oustry, and Sagastizábal [2000]. Later on a theory for primal-dual gradient structured functions Mifflin and Sagastizábal [2003] was developed on the basis of implicit function theorems and algebra. The elegant geometrical approach of partly smooth functions (not necessarily convex) A. S. Lewis [2002] relies on Variational Analysis.

Already in Lemaréchal and Sagastizábal [1997a] there was a hint of the answer, pointing at the need of suitably *decomposing the space*  $\mathbb{R}^n$  according to the structure of nonsmoothness of  $f$  at  $\bar{x}$ . More precisely, Lemaréchal and Sagastizábal [ibid., Sec. 3] showed that when it comes to second order the regularizing effect of the Moreau-Yosida operator is driven by the prox-Jacobian, whose image is entirely contained in the normal cone

$$\mathcal{U}(p^\circ) := N_{\partial f(p^\circ)}(\nabla F(x^\circ)).$$

Such a cone is in fact a subspace because the Moreau-Yosida gradient is in the relative interior of the subdifferential. Lemaréchal and Sagastizábal [ibid., Sec. 3]. Consider the linear subspace spanned by  $\partial f(p^\circ)$ , i.e.,  $\mathcal{U}(p^\circ) \perp \mathcal{U}(p^\circ)$ . Any closed convex function

looks “kinky” along  $p^\circ + \mathcal{V}(p^\circ)$  and smooth in  $p^\circ + \mathcal{U}(p^\circ)$ . For the example function such relations and the facts that

$$\partial f(p^\circ) = [-1, 1] \times \left\{ \frac{a\mu}{a+\mu} u^\circ \right\} \quad \text{and} \quad \mathcal{U}(p^\circ) = \{0\} \times \mathbb{R},$$

explain the respective “kinkiness” and smoothness of the left and right graphs in [Figure 2](#) (keep in mind that  $\bar{x} = p^f(\bar{x})$ , so  $p^\circ = \bar{x}$  in this case).

**Bibliographical note.** Proximal bundle methods designed to seek faster convergence by means of the Moreau-Yosida regularization were studied in [Lemaréchal and Sagastizábal \[1994\]](#), [Mifflin \[1996\]](#), [Mifflin, Sun, and Qi \[1998\]](#), [Chen and Fukushima \[1999\]](#), [Rauf and Fukushima \[2000\]](#).

## 4 Walking the path of superlinear rate

Once the  $\mathcal{V}$  and  $\mathcal{U}$  subspaces were brought to light, two important issues remained. First, there was the question of the algorithmic potential of the approach. In Nonlinear Programming a Newton direction is the result of minimizing a *second-order model* of the smooth function. For a nonsmooth function  $f$  we describe below how to make a second-order expansion along a trajectory related to the  $\mathcal{U}$ -subspace.

The second issue, crucial for applicability regards turning the conceptual  $\mathcal{V}\mathcal{U}$ -algorithm into an actual implementable method. Properties of the proximal point operator and the relations (2-2) revealed once more providential in this respect.

**4.1 Fast tracks.** In the  $\mathcal{V}\mathcal{U}$ -space decomposition every  $x \in \mathbb{R}^n$  has a  $\mathcal{V}$  and a  $\mathcal{U}$  component, say  $x_v$  and  $x_u$ . For our simple example the subspaces are

$$(4-1) \quad \mathcal{V} = \mathbb{R} \times \{0\} \quad \text{and} \quad \mathcal{U} = \{0\} \times \mathbb{R}$$

(this was the reason for denoting by  $v$  and  $u$  the two vector coordinates.)

Many nonsmooth convex functions admit a second-order expansion when considering special trajectories, parameterized by  $u \in \mathcal{U}$  sufficiently small. Such trajectories, called *fast track* in [Mifflin and Sagastizábal \[2002\]](#) and denoted by  $\chi(u)$  have as  $\mathcal{V}$ -component a very special function of the  $\mathcal{U}$ -component,  $v = v(u)$ :

$$\chi(u) := \bar{p} + (v(u), u).$$

The  $\mathcal{V}$ -component is special in the sense given by the essential relation from [Lemaréchal, Oustry, and Sagastizábal \[2000, Cor.3.5\]](#), stating that  $v(u)$  goes to zero faster than  $u$ .

To illustrate the fast track, the example in [Section 3.2](#) is too simple:  $v(u) \equiv 0$ . Consider a slightly more involved function,

$$f(v, u) = \max(f^1(v), f^2(u)) = \max(|v|, \frac{1}{2}au^2),$$

and recall that  $a > 0$ . The function  $f$  fails to be differentiable on the locus of the equation  $|v| = \frac{1}{2}au^2$ . In particular, its minimizer  $\bar{p} = (0, 0)$  has the subdifferential  $\partial f(\bar{p}) = [-1, 1] \times \{0\}$  so the  $\mathcal{V}\mathcal{U}$ -subspaces remain those in [\(4-1\)](#).

For a given interior subgradient with  $\mathcal{V}$ -component  $\gamma$  the fast track is defined as follows:

$$v(u) = v(u; \gamma) \in \arg \min \{f(\bar{p} + (v, u)) - \langle \gamma, v \rangle_{\mathcal{V}} : v \in \mathcal{V}\}.$$

Working out the calculations with  $|\gamma| < 1$  for the example function gives that

$$v(u) = \frac{1}{2} \text{sign}(\gamma)u^2 \quad \text{and, hence,} \quad \chi(u) = \left(\frac{1}{2} \text{sign}(\gamma)u^2, u\right).$$

The fast track exists and its  $\mathcal{V}$ -component is smooth for a fairly general class of functions, with sufficiently structured nonsmoothness, for details see [Mifflin and Sagastizábal \[2002\]](#). Furthermore the function can be expanded up to second order along the fast track. In our example, for any  $u \in \mathbb{R}$

(4-2)

$$f(\chi(u)) = \frac{1 - |\gamma|}{2}au^2, \nabla_{\mathcal{U}} f(\chi(u)) = (1 - |\gamma|)au, \text{ and } \nabla_{\mathcal{U}}^2 f(\chi(u)) = (1 - |\gamma|)a.$$

The second-order object, called the  $\mathcal{U}$ -Hessian and denoted  $H_{\mathcal{U}} f$ , can be used in a  $\mathcal{U}$ -Newton scheme performing the following steps

1. Having  $u$ , compute  $v(u)$ .
2. Compute  $\gamma(u) = \nabla_{\mathcal{U}} f(\chi(u))$ , an element tangent to the fast track.
3. Update  $u$  by solving the system  $H_{\mathcal{U}} f \Delta_{\mathcal{U}} = -\gamma(u)$ .

This conceptual scheme converges superlinearly to  $\bar{p}$  because the  $\mathcal{U}$ -update is superlinear on the  $\mathcal{U}$ -component (it is a Newton move on the smooth function  $f \circ \chi$ ). Since  $v(u) = o(\|u\|)$  by the important Corollary 3.5 in [Lemaréchal, Oustry, and Sagastizábal \[2000\]](#), the speed of convergence of the overall process is directed by the speed of the  $\mathcal{U}$ -component.

**4.2 Putting the conceptual scheme in practice.** So far superlinear  $\mathcal{V}\mathcal{U}$  steps are stated on a conceptual level, since for their computation the subspaces and the  $\mathcal{U}$ -Hessian must



be known. The question of how to compute elements in the primal-dual track  $(\chi(u), \gamma(u))$  was resolved by the following fundamental result, [Mifflin and Sagastizábal \[2002, Thm.5.1\]](#), where  $\text{ri } S$  stands for the relative interior of a convex set  $S$ .

**Theorem 4.1** (Proximal points are on the fast track). *Let  $f : \mathbb{R}^n \rightarrow \mathbb{R}$  be a convex function with minimizer  $\bar{x} \in \mathbb{R}^n$ . Suppose that  $0 \in \text{ri } \partial f(\bar{x})$  and let  $\chi(u)$  be a fast track. Given a positive parameter  $\mu$ , for all  $x$  close enough to  $\bar{x}$  there exists  $u(x)$  such that  $p^f(x) = \chi(u(x))$ .*

Together with (2-2) this result opens the door towards implementability. The primal-dual fast track is approximated using the proximal point of the cutting-plane model:

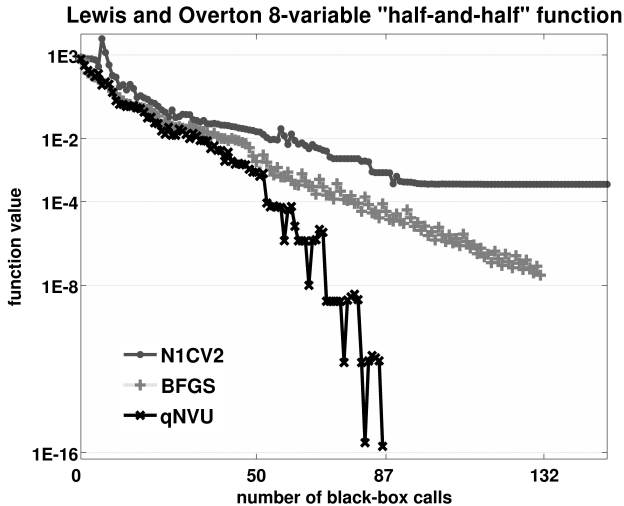
$$\chi(u) \approx p^{\mathfrak{m}^k}(\hat{x}^k) \quad \text{and} \quad \gamma(u) \approx \hat{g}^k \in \partial_{\varepsilon^k} f(\hat{x}^k).$$

The  $\mathcal{V}\mathcal{U}$ -bundle method [Mifflin and Sagastizábal \[2005\]](#) is the first globally convergent algorithm with Q-superlinear rate designed according to these premises. To approximate the  $\mathcal{U}$ -Hessian, a quasi-Newton matrix is updated in the  $\mathcal{U}$ -subspace by means of the well-known Broyden-Fletcher-Goldfarb-Shanno (BFGS) formula. Approximations for the  $\mathcal{V}\mathcal{U}$ -subspaces result from the solution to a new QP problem, yielding the dual component of the fast track. Indeed, an important distinctive feature of the method is that approximating the primal-dual track requires solving *two* QP problems per iteration. This is key to compute the  $\mathcal{V}\mathcal{U}$  objects using the most updated information.

The following excerpt from [Mifflin and Sagastizábal \[2012\]](#), slightly modified to fit the current setting, together with [Figure 3](#), illustrates the superlinear behaviour of the  $\mathcal{V}\mathcal{U}$ -bundle method on a test function that is a multidimensional version of the  $\mathcal{V}\mathcal{U}$ -function in [Section 3.2](#).

*“The half-and-half function  $f(x) = \sqrt{x^T B x} + x^T A x$  was created by A. Lewis and M. Overton to analyze BFGS behavior when minimizing a nonsmooth function. The 8-variable example in the figure has a matrix  $A$  with all elements zero, except for ones on the diagonal at odd numbered locations ( $B(i, i) = 1$  for  $i = 1, 3, 5, 7$ ). The matrix  $A$  is diagonal with elements  $A(i, i) = 1/i^2$  for  $i = 1, \dots, 8$ . The minimizer of this partly smooth convex function is at  $\bar{x} = 0$ , where the  $\mathcal{V}$  and  $\mathcal{U}$  subspaces both have dimension 4; hence, the name half-and-half.*

*Each graph in the figure shows function values from all points generated by its corresponding algorithm starting from the point having all components equal to 20.08. The top curve indicates sublinear rate of convergence. It was obtained with a proximal bundle method, implemented in the code N1CV2 by C. Lemaréchal and C. Sagastizábal [Lemaréchal and Sagastizábal \[1997b\]](#). The middle curve exhibits a linear convergence rate and corresponds to the BFGS implementation by M. Overton, who adapted the method for non-smooth functions via a suitable line search developed with A. Lewis in [A. S. Lewis and](#)*

Figure 3: Superlinear speed of a fully implementable  $\mathcal{U}\mathcal{U}$ -method

[Overton \[2013\]](#). They argue that in NSO the linear convergence of “vanilla BFGS” as exhibited by this example is surprisingly typical. The intriguing assertion, mentioned in [A. S. Lewis \[2014\]](#), has been proved only for a two variable example with the use of exact line searches, i.e. by exploiting nonsmoothness.

It pays indeed to exploit nonsmoothness, as shown by the curve at the bottom of the figure, which exhibits a superlinear rate and results from the quasi-Newton  $\mathcal{U}\mathcal{U}$ -bundle algorithm [Mifflin and Sagastizábal \[2005\]](#).

**Bibliographical note.** The Eigenvalue Optimization works [Oustry \[1999\]](#), [Oustry \[2000\]](#), with two QP problems per iteration, laid the groundwork for the  $\mathcal{U}\mathcal{U}$ -bundle method in [Mifflin and Sagastizábal \[2005\]](#). An important difference is that while the latter uses a black-box oracle as in [Figure 1](#), the former works exploit *rich* oracles, delivering more than one subgradient at once (in the eigenvalue context, this amounts to computing most of the eigenvectors at each iteration).

The smooth activity manifold in [A. S. Lewis \[2002\]](#) is the primal track  $\chi(u)$ , while the composition  $f \circ \chi$  is the  $\mathcal{U}$ -Lagrangian [Lemaréchal, Oustry, and Sagastizábal \[2000\]](#). Interesting geometrical relations with the sequential quadratic programming method and the predictor-corrector type algorithms in Nonlinear Programming were analyzed in [Miller and Malick \[2005\]](#) and [Daniilidis, Hare, and Malick \[2006\]](#).

Computational issues on how to numerically identify the primal track and the  $\mathcal{V}\mathcal{U}$ -objects were respectively addressed in [Daniilidis, Sagastizábal, and M. Solodov \[2009\]](#) and [Hare \[2014\]](#).

Finally, the case of (nonconvex) prox-regular functions was considered, among other authors, by [Hare and A. Lewis \[2007\]](#) and [Mifflin and Sagastizábal \[2004\]](#); see also [Huang, Pang, Lu, and Xia \[2017\]](#).

## 5 NSO models: Going above and beyond

Superlinear speed is undeniably a desirable feature since it results in the algorithm making less iterations and perhaps even more importantly, in achieving higher precision. This is not the only concern in NSO, however. When it comes to running times, a bottleneck refers to the time spent inside of the oracle, computing the function value and one subgradient for a given iterate. In many real-life applications the overall CPU time is typically divided into 15%-25% for the bundle calculations. The rest, that is more than three quarters of the total time, is consumed by the oracle.

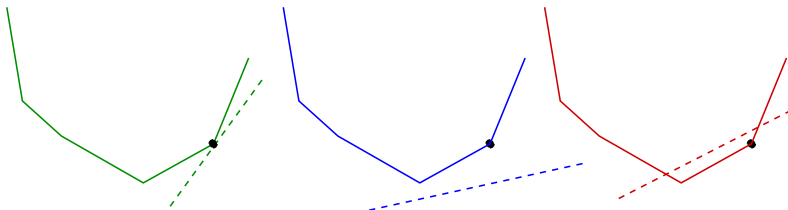
Central to this issue is the distinction between serious and null steps. Roughly speaking, if one is to cut short the oracle computing times, better do it for those iterates that result in a null step. After all, the subsequence that “matters the most” from the convergence viewpoint is the one made up of serious steps. With this philosophy, the oracle precision becomes *variable*, bringing a new paradigm into the field, in which there is an interaction between the NSO algorithm and the oracle.

**5.1 Dealing with inexactness.** When the oracle delivers inexact information

$$f_{x^k} \approx f(x^k) \quad \text{and} \quad g_{x^k} \approx g^k \in \partial f(x^k)$$

there are two different situations, represented by the two rightmost graphs in [Figure 4](#).

Figure 4: Exact, lower inexact, and upper inexact linearizations



In the figure, the full lines display the function  $f$ , with a black circle corresponding to  $f(x^k)$ . Dotted lines are the linearizations built with an exact oracle on the left, and two inexact oracles on the center and right. In the middle graph the linearization does not cut off a portion of the graph of  $f$  so the situation is more favourable. Models based on such linearizations may not miss a region where  $f$  attains its minimum.

The formal identification of the favourable situation is given by requiring that the output of the oracle satisfies, for a nonnegative error  $\eta^k$ , that

$$f_{x^k} \in [f(x^k) - \eta^k, f(x^k)] \quad \text{and} \quad g_{x^k} \in \partial_{\eta^k} f(x^k).$$

The resulting inexact linearization  $\ell^k$  remains everywhere below the function, hence justifying the name of “lower” oracle.

In the less favourable rightmost situation, the “upper” oracle output is

$$f_{x^k} = f(x^k) - \eta^k \quad \text{and} \quad g_{x^k} \in \partial_{\eta^k + \eta_g^k} f(x^k),$$

without any specification on the sign of the errors  $\eta^k$  and  $\eta_g^k$  (nevertheless  $\eta^k + \eta_g^k \geq 0$ , by construction). In this case, the approximate functional value can be *above* the exact one, as in the figure.

Lower oracles appear in situations when the function  $f$  is the result of some maximization process, like a dual function in Lagrangian relaxation, or a value function in Benders decomposition, or the recourse function in a two-stage stochastic program. In such circumstances  $f(x) = \max\{F(x, y) : y \in Y\}$  for certain smooth function  $F$ . Typically, the feasible set is approximated by  $Y^k \subset Y$ , so

$$f_{x^k} := \max\{F(x, y) : y \in Y^k\} = F(x^k, y^k) \quad \text{and} \quad g_{x^k} := \nabla_x F(x^k, y^k),$$

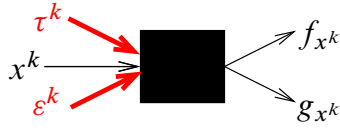
where  $y^k$  stands for a maximizer.

Lower approximations can also be obtained by stopping before optimality the solution process of maximizing  $F(x^k, \cdot)$  over  $Y$ , for instance giving to the oracle a maximum computing time. This is a good example of the important class of *on-demand accuracy* oracles considered in [de Oliveira and Sagastizábal \[2014\]](#). When compared with the initial black boxes the originality of such oracles lies in the fact that their output can be delivered with more or less precision, depending on the nature of the evaluation point. Specifically, recall that to be declared the next serious step the new functional value needs to be below the target, that in the inexact setting has the expression

$$\tau^k := f_{\hat{x}^k} - m\delta^k.$$

[Figure 5](#) represents schematically the new setting, in addition to the evaluation point, the oracle receives a target  $\tau^k$  and an error bound  $\varepsilon^k \geq 0$ .

Figure 5: On-demand accuracy oracle



An on-demand accuracy oracle has a maximum computing time to produce its output and operates as follows. If the oracle reaches a functional value that is below the target, then the computation error must be smaller than the given bound:

$$f_{x^k} \leq \tau^k \implies \eta^k + \eta_g^k \leq \varepsilon^k.$$

If the oracle reaches the maximum time and the approximate functional value is not below the target, then the returned value may have any precision (the value of  $\varepsilon^k$  is not used). This ingenious mechanism was initiated with the works by K. Kiwiel [K. C. Kiwiel \[2008\]](#), [K. C. Kiwiel and Lemaréchal \[2009\]](#) and was later on systematically studied for level bundle variants in [de Oliveira and Sagastizábal \[2014\]](#). By managing the error bound so that  $\varepsilon^k \rightarrow 0$ , any infinite sequence of serious steps eventually has exact functional values and therefore converges to an *exact* minimizer.

Examples of upper oracles abound in Derivative-Free NSO [Hare and Planiden \[2016\]](#), Stochastic Programming [van Ackooij, de Oliveira, and Song \[2018\]](#), and Probabilistic Optimization [van Ackooij, Berge, de Oliveira, and Sagastizábal \[2017\]](#). In the latter the probability distribution is continuous, and its discretization yields approximations that can be above or below the exact functional value. On-demand accuracy oracles are still a possibility, for example for Gaussian distributions, since in this case the approximation can be computed with any desired accuracy.

If the error bound sent to the oracle is too large the resulting linearization may be so bad that, to prevent the algorithm from breaking down, certain *noise attenuation* step needs to be put in place. When there is noise attenuation, the parameter  $\mu$  is increased and a third subsequence needs to be considered in the convergence analysis. We do not go into more details of this rather technical issue. We just mention that the situation can be avoided if the oracle is of lower type. A comprehensive convergence analysis theory, covering many different oracle situations can be found in [de Oliveira, Sagastizábal, and Lemaréchal \[2014\]](#).

## 6 Final words

As announced, the presented material is by no means exhaustive. Rather, the writing reflects insights gained mostly thanks to the generosity of colleagues and mentors I was lucky to work with. I am very much indebted to Claude Lemaréchal and Robert Mifflin, for passing on their passion for “kinks”, space decompositions, and black boxes. The works by Krzysztof Kiwiel have been enlightening more than once. And my PhD students have taught me how to learn when teaching.

This survey concludes with a tribute to the late Jonathan M. Borwein, an enthusiastic pioneer of blending Variational Analysis with Computational Mathematics so that researchers get to “visualise mathematics for greater understanding” (sic).

The readers wishing to catch a glimpse of Jon’s very rich and long lasting legacy, can visit the page

<https://carma.newcastle.edu.au/jon/>

In the essay [Borwein \[2017\]](#), one of his last publications, Jon gives advice to young mathematicians (author, referee, or editor, he says). We include below some of Jon’s final recommendations in the essay, to encourage its reading, that we found inspirational in many levels:

- (a) Not all questions deserve to be answered.
- (b) Aim to have two qualitatively different examples. Ask if they are natural or contrived.
- (c) Better an interesting new proof of a substantial known result than a modest and routine generalisation of an uninteresting result.
- (d) Don’t imagine many people are reading your paper linearly. Most readers - if one is lucky enough to have any - are leafing through looking for the punchlines. So avoid too many running hypotheses or at least make a full statement of each major result.
- (e) Remember that your readers or audience may well be using English as a second language. This does not mean you should dumb down your language but - as with the advice to restate your main hypotheses - the key points should be made in simple declarative English.
- (f) Most research mathematicians are not scholars let alone trained mathematical historians. Avoid the temptation to say that an idea was invented or introduced by someone - whether it is Hilbert or your supervisor. Say rather that you first learned about the topic from a paper you cite.

- (g) Make sure your citation list is up to date. In the current digital environment there is no excuse for failing to do a significant literature search.
- (h) When you submit your well-motivated and carefully written paper (including a reasonable literature discussion and great examples) to a journal remember that you alone, and not the referee, are responsible for correctness of your arguments.
- (i) Above all be honest.

## References

- W. van Ackooij, V. Berge, W. de Oliveira, and C. Sagastizábal (2017). “[Probabilistic optimization via approximate p-efficient points and bundle methods](#)”. *Computers & Operations Research* 77, pp. 177–193 (cit. on p. 3832).
- W. van Ackooij, W. de Oliveira, and Y. Song (2018). “[Adaptive Partition-Based Level Decomposition Methods for Solving Two-Stage Stochastic Programs with Fixed Recourse](#)”. *INFORMS Journal on Computing* 30.1, pp. 57–70 (cit. on p. 3832).
- P. Apkarian, D. Noll, and L. Ravanbod (2016). “Nonsmooth Bundle Trust-region Algorithm with Applications to Robust Stability”. *Set-Valued and Variational Analysis*. 24.1, pp. 115–148 (cit. on p. 3822).
- A. Auslender (1987). “Numerical methods for nondifferentiable convex optimization”. *Mathematical Programming Studies* 30, pp. 102–126 (cit. on p. 3821).
- J.F. Bonnans, J.Ch. Gilbert, C. Lemaréchal, and C. Sagastizábal (2006). *Numerical Optimization. Theoretical and Practical Aspects*. Universitext. Berlin: Springer-Verlag, 2nd. edition, xiv+423 pp. (Cit. on p. 3817).
- J. M. Borwein (2017). “Generalisations, Examples, and Counter-examples in Analysis and Optimisation”. *Set-Valued and Variational Analysis*. 24.3. Special issue Advances in Monotone Operators Theory and Optimization in honour of Michel Théra at 70. Edited by F. J. Aragón Artacho, R. Henrion, M.A. López-Cerdá, C. Sagastizábal, J. M. Borwein, pp. 467–479 (cit. on p. 3833).
- X. Chen and M. Fukushima (1999). “Proximal quasi-Newton methods for nondifferentiable convex optimization”. *Math. Program.* 85.2, Ser. A, pp. 313–334 (cit. on p. 3826).
- E. Cheney and A. Goldstein (1959). “Newton’s Method for Convex Programming and Tchebycheff approximations”. *Numerische Mathematik* 1, pp. 253–268 (cit. on p. 3820).
- R. Correa and C. Lemaréchal (1993). “Convergence of some algorithms for convex minimization”. *Mathematical Programming* 62.2, pp. 261–275 (cit. on p. 3821).
- A. Daniilidis, W. Hare, and J. Malick (2006). “[Geometrical interpretation of the predictor-corrector type algorithms in structured optimization problems](#)”. *Optimization* 55.5-6, pp. 481–503 (cit. on p. 3829).

- A. Daniilidis, C. Sagastizábal, and M. Solodov (2009). “Identifying Structure of Nonsmooth Convex Functions by the Bundle Technique”. *SIAM Journal on Optimization* 20.2, pp. 820–840 (cit. on p. 3830).
- C. I. Fábián (2000). “Bundle-type methods for inexact data”. In: *Proceedings of the XXIV Hungarian Operations Research Conference (Veszprém, 1999)*. Vol. 8. 1, pp. 35–55. MR: 1778487 (cit. on p. 3822).
- A. Frangioni (2002). “Generalized Bundle Methods”. *SIAM Journal on Optimization* 13.1, pp. 117–156 (cit. on p. 3822).
- M. Fukushima (1984). “A descent algorithm for nonsmooth convex optimization”. *Mathematical Programming* 30, pp. 163–175 (cit. on p. 3821).
- M. Haara, K. Miettinen, and M. M. Makela (2004). “New limited memory bundle method for large-scale nonsmooth optimization”. *Optimization Methods and Software* 6, pp. 673–692 (cit. on p. 3822).
- W. Hare (2014). “Numerical Analysis of  $\mathcal{UU}$ -Decomposition,  $\mathcal{U}$ -Gradient, and  $\mathcal{U}$ -Hessian Approximations”. *SIAM Journal on Optimization* 24.4, pp. 1890–1913 (cit. on p. 3830).
- W. Hare and A. Lewis (2007). “Identifying Active Manifolds”. *Algorithmic Operations Research* 2.2 (cit. on p. 3830).
- W. Hare and C. Planiden (2016). “Computing Proximal Points of Convex Functions with Inexact Subgradients”. *Set-Valued and Variational Analysis*. On Line First (cit. on p. 3832).
- C. Helmberg and K. Kiwiel (2002). “A spectral bundle method with bounds”. *Mathematical Programming* 93.1, pp. 173–194 (cit. on p. 3822).
- M. Hintermüller (2001). “A Proximal Bundle Method Based on Approximate Subgradients”. *Computational Optimization and Applications* 20.3, pp. 245–266 (cit. on p. 3818).
- J.-B. Hiriart-Urruty and C. Lemaréchal (1993). *Convex Analysis and Minimization Algorithms*. Grund. der math. Wiss 305-306. (two volumes). Springer-Verlag (cit. on p. 3817).
- M. Huang, L.P. Pang, Y. Lu, and Z.Q. Xia (2017). “A Fast Space-decomposition Scheme for Nonconvex Eigenvalue Optimization”. *Set-Valued and Variational Analysis*. 25.1, pp. 43–67 (cit. on p. 3830).
- J. E. Kelley (1960). “The Cutting Plane Method for Solving Convex Programs”. *J. Soc. Indust. Appl. Math.* 8, pp. 703–712 (cit. on p. 3820).
- K. C. Kiwiel (1990). “Proximity Control in Bundle Methods for Convex Nondifferentiable Minimization”. *Mathematical Programming* 46, pp. 105–122 (cit. on p. 3820).
- (2006). “A Proximal Bundle Method with Approximate Subgradient Linearizations”. *SIAM Journal on Optimization* 16.4, pp. 1007–1023 (cit. on p. 3818).
- (2008). “A Method of Centers with Approximate Subgradient Linearizations for Nonsmooth Convex Optimization”. *SIAM Journal on Optimization* 18.4, pp. 1467–1489 (cit. on p. 3832).



- K. C. Kiwiel and Claude Lemaréchal (2009). “[An inexact bundle variant suited to column generation](#)”. *Math. Program.* 118.1, pp. 177–206 (cit. on p. [3832](#)).
- K.C. Kiwiel (1985). *Methods of descent for nondifferentiable optimization*. Berlin: Springer-Verlag, pp. vi+362 (cit. on p. [3822](#)).
- C. Lemaréchal and R. Mifflin, eds. (1978). *Nonsmooth optimization*. Vol. 3. IASA proceedings series. Pergamon Press, Oxford, x+188 pp. (Cit. on p. [3818](#)).
- C. Lemaréchal (1975). “An extension of Davidon methods to nondifferentiable problems”. *Mathematical Programming Studies* 3, pp. 95–109 (cit. on p. [3822](#)).
- (1980). *Extensions diverses des méthodes de gradient et applications*. Thèse d’Etat, Université de Paris IX (cit. on p. [3818](#)).
- C. Lemaréchal, A. Nemirovskii, and Yu. Nesterov (1995). “New variants of bundle methods”. *Mathematical Programming* 69, pp. 111–148 (cit. on p. [3822](#)).
- C. Lemaréchal, F. Oustry, and C. Sagastizábal (2000). “The  $\mathcal{U}$ -Lagrangian of a convex function”. *Trans. Amer. Math. Soc.* 352.2, pp. 711–729 (cit. on pp. [3825](#)–[3827](#), [3829](#)).
- C. Lemaréchal and C. Sagastizábal (1994). “An approach to variable metric bundle methods”. In: *Systems Modelling and Optimization*. Ed. by J. Henry and J-P. Yvon. Lecture Notes in Control and Information Sciences 197. Springer-Verlag, pp. 144–162 (cit. on pp. [3823](#), [3826](#)).
- (1997a). “[Practical aspects of the Moreau-Yosida regularization: theoretical preliminaries](#)”. *SIAM Journal on Optimization* 7.2, pp. 367–385 (cit. on pp. [3824](#), [3825](#)).
- (1997b). “[Variable metric bundle methods: from conceptual to implementable forms](#)”. *Mathematical Programming* 76, pp. 393–410 (cit. on pp. [3822](#), [3828](#)).
- A. S. Lewis (2002). “[Active Sets, Nonsmoothness, and Sensitivity](#)”. *SIAM Journal on Optimization* 13.3, pp. 702–725 (cit. on pp. [3825](#), [3829](#)).
- (2014). “Nonsmooth optimization: conditioning, convergence and semi-algebraic models”. In: *Proceedings of the International Congress of Mathematicians, Seoul*. Vol. IV, Invited Lectures. Kyung Moon SA, pp. 871–896 (cit. on pp. [3817](#), [3829](#)).
- A. S. Lewis and M. L. Overton (2013). “Nonsmooth optimization via quasi-Newton methods”. *Mathematical Programming* 141 (1–2), pp. 135–163 (cit. on p. [3828](#)).
- L. Lukšan and J. Vlček (1999). “Globally convergent variable metric method for convex nonsmooth unconstrained minimization”. *J. Optim. Theory Appl.* 102.3, pp. 593–613 (cit. on p. [3822](#)).
- R. Mifflin (1977). “An algorithm for constrained optimization with semismooth functions”. *Mathematics of Operations Research* 2, pp. 191–207 (cit. on p. [3822](#)).
- (1996). “A quasi-second-order proximal bundle algorithm”. *Mathematical Programming* 73.1, pp. 51–72 (cit. on p. [3826](#)).
- R. Mifflin and C. Sagastizábal (2002). “[Proximal Points are on the Fast Track](#)”. *Journal of Convex Analysis* 9.2, pp. 563–579 (cit. on pp. [3826](#)–[3828](#)).

- (2003). “Primal-Dual Gradient Structured Functions: second-order results; links to epiderivatives and partly smooth functions”. *SIAM Journal on Optimization* 13.4, pp. 1174–1194 (cit. on p. 3825).
- (2004). “ $\mathcal{U}\mathcal{U}$ -Smoothness and Proximal Point Results for Some Nonconvex Functions”. *Optimization Methods and Software* 19.5, pp. 463–478 (cit. on p. 3830).
- (2005). “A  $\mathcal{U}\mathcal{U}$ -algorithm for convex minimization”. *Mathematical Programming* 104.2–3, pp. 583–608 (cit. on pp. 3828, 3829).
- (2012). “A Science Fiction Story in Nonsmooth Optimization Originating at IIASA”. In: *Optimization Stories*. Vol. Extra for ISMP 2012, ed. by M. Grötschel, 460 pp. Documenta Mathematica (cit. on p. 3828).
- R. Mifflin, D.F. Sun, and L.Q. Qi (1998). “Quasi-Newton bundle-type methods for non-differentiable convex optimization”. *SIAM Journal on Optimization* 8.2, pp. 583–603 (cit. on p. 3826).
- S. A. Miller and J. Malick (2005). “Connections between  $\mathcal{U}$ -Lagrangian, Riemannian Newton, and SQP Methods”. *Mathematical Programming* 104, pp. 609–633 (cit. on p. 3829).
- J.J. Moreau (1965). “Proximité et dualité dans un espace Hilbertien”. *Bulletin de la Société Mathématique de France* 93, pp. 273–299 (cit. on pp. 3820, 3822).
- W. de Oliveira (2017). “Target radius methods for nonsmooth convex optimization”. *Operations Research Letters* 45.6, pp. 659–664 (cit. on p. 3822).
- W. de Oliveira and C. Sagastizábal (2014). “Level Bundle Methods for Oracles with On-Demand Accuracy”. *Optimization Methods and Software* 29.6. Charles Broyden Prize for best paper published by the journal in 2014. (cit. on pp. 3831, 3832).
- W. de Oliveira, C. Sagastizábal, and C. Lemaréchal (2014). “Convex proximal bundle methods in depth: a unified analysis for inexact oracles”. English. *Mathematical Programming* 148.1-2, pp. 241–277 (cit. on p. 3832).
- W. de Oliveira and M. Solodov (2016). “A doubly stabilized bundle method for nonsmooth convex optimization”. *Mathematical Programming* 156.1, pp. 125–159 (cit. on p. 3822).
- A. Ouerou (2013). “The proximal Chebychev center cutting plane algorithm for convex additive functions”. *Math. Program.* 140.1, pp. 163–187 (cit. on p. 3822).
- F. Oustry (1999). “The  $\mathcal{U}$ -Lagrangian of the maximum eigenvalue function”. *SIAM Journal on Optimization* 9, pp. 526–549 (cit. on p. 3829).
- (2000). “A second-order bundle method to minimize the maximum eigenvalue function”. *Mathematical Programming* 89.1, Ser. A, pp. 1–33 (cit. on p. 3829).
- A.I. Rauf and M. Fukushima (2000). “Globally convergent BFGS method for nonsmooth convex optimization”. *J. Optim. Theory Appl.* 104.3, pp. 539–558 (cit. on p. 3826).

- S. M. Robinson (1999). “Linear convergence of epsilon-subgradient descent methods for a class of convex functions”. *Mathematical Programming* 86.1, pp. 41–58 (cit. on p. 3822).
- H. Schramm and J. Zowe (1992). “A version of the bundle idea for minimizing a nonsmooth function: conceptual idea, convergence analysis, numerical results”. *SIAM Journal on Optimization* 2.1, pp. 121–152 (cit. on p. 3822).
- N. Shor (1970). “Utilization of the Operation of Space Dilatation in the minimization of convex function”. *Cybernetics* 6, pp. 7–15 (cit. on p. 3818).
- M. V. Solodov (2003). “On Approximations with Finite Precision in Bundle Methods for Nonsmooth Optimization”. *Journal of Optimization Theory and Applications* 119.1, pp. 151–165 (cit. on p. 3818).
- K. Yosida (1964). *Functional Analysis*. Springer Verlag (cit. on p. 3822).

Received 2017-12-01.

CLAUDIA SAGASTIZÁBAL: IMECC - UNICAMP, RUA SERGIO BUARQUE DE HOLANDA, 651,  
13083-859, CAMPINAS, SP, BRAZIL  
[sagastiz@unicamp.br](mailto:sagastiz@unicamp.br)

# SPECTRAHEDRAL LIFTS OF CONVEX SETS

REKHA R. THOMAS

## Abstract

Efficient representations of convex sets are of crucial importance for many algorithms that work with them. It is well-known that sometimes, a complicated convex set can be expressed as the projection of a much simpler set in higher dimensions called a *lift* of the original set. This is a brief survey of recent developments in the topic of lifts of convex sets. Our focus will be on lifts that arise from affine slices of real positive semidefinite cones known as *psd* or *spectrahedral lifts*. The main result is that projection representations of a convex set are controlled by factorizations, through closed convex cones, of an operator that comes from the convex set. This leads to several research directions and results that lie at the intersection of convex geometry, combinatorics, real algebraic geometry, optimization, computer science and more.

## 1 Introduction

Efficient representations of convex sets are of fundamental importance in many areas of mathematics. An old idea from optimization for creating a compact representation of a convex set is to express it as the projection of a higher-dimensional set that might potentially be simpler, see for example [Conforti, Cornuéjols, and Zambelli \[2010\]](#), [Ben-Tal and Nemirovski \[2001\]](#). In many cases, this technique offers surprisingly compact representations of the original convex set. We present the basic questions that arise in the context of projection representations, provide some answers, pose more questions, and examine the current limitations and challenges.

As a motivating example, consider a full-dimensional convex polytope  $P \subset \mathbb{R}^n$ . Recall that  $P$  can be expressed either as the convex hull of a finite collection of points in  $\mathbb{R}^n$

---

The author was partially supported by the U.S. National Science Foundation grant DMS-1719538. This paper was written while the author was in residence at the Mathematical Sciences Research Institute in Berkeley, California, during the Fall 2017 semester, and based on work supported by the National Science Foundation under Grant No. 1440140.

MSC2010: primary 52A02; secondary 90C02.

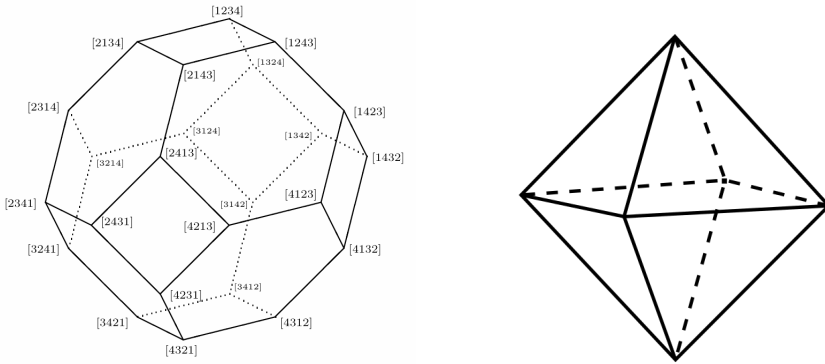


Figure 1: The permutahedron  $\Pi_4$  and the crosspolytope  $C_3$ .

or as the intersection of a finite set of linear halfspaces. The minimal set of points needed in the convex hull representation are the *vertices* of  $P$ , and the irredundant inequalities needed are in bijection with the *facets* (codimension-one faces) of  $P$ . Therefore, if the number of facets of  $P$  is exponential in  $n$ , then the linear inequality representation of  $P$  is of size exponential in  $n$ . The complexity of optimizing a linear function over  $P$  depends on the size of its inequality representation and hence it is worthwhile to ask if efficient inequality representations can be obtained through some indirect means such as projections. We illustrate the idea on two examples.

**Example 1.1.** The  $n$ -dimensional *crosspolytope*  $C_n$  is the convex hull of the standard unit vectors  $e_i \in \mathbb{R}^n$  and their negatives Ziegler [1995, Example 0.4]. For example,  $C_2$  is a square and  $C_3$  is an octahedron, see Figure 1. Written in terms of inequalities,

$$C_n = \{x \in \mathbb{R}^n : \pm x_1 \pm x_2 \pm \cdots \pm x_n \leq 1\}$$

and all  $2^n$  inequalities listed are needed as they define facets of  $C_n$ . However,  $C_n$  is also the projection onto  $x$ -coordinates of the polytope

$$Q_n = \left\{ (x, y) \in \mathbb{R}^{2n} : \sum_{i=1}^n y_i = 1, -y_i \leq x_i \leq y_i \quad \forall i = 1, \dots, n \right\}$$

which involves only  $2n$  inequalities and one equation. □

**Example 1.2.** The *permutahedron*  $\Pi_n$  is the  $(n-1)$ -dimensional polytope that is the convex hull of all vectors obtained by permuting the coordinates of the  $n$ -dimensional vector  $(1, 2, 3, \dots, n)$ . It has  $2^n - 2$  facets, each indexed by a proper subset of  $[n] :=$

$\{1, 2, \dots, n\}$  Ziegler [ibid., Example 0.10]. In 2015, Goemans used sorting networks to show that  $\Pi_n$  is the linear image of a polytope  $Q_n$  that has  $\Theta(n \log n)$  variables and facets, and also argued that one cannot do better.  $\square$

The key takeaway from the above examples is that one can sometimes find efficient linear representations of polytopes if extra variables are allowed; a complicated polytope  $P \subset \mathbb{R}^n$  might be the linear projection of a polytope  $Q \subset \mathbb{R}^{n+k}$  with many fewer facets. To be considered efficient, both  $k$  and the number of facets of  $Q$  must be polynomial functions of  $n$ . Such a polytope  $Q$  is called a *lift* or *extended formulation* of  $P$ . Since optimizing a linear function over  $P$  is equivalent to optimizing the same function over a lift of it, these projection representations offer the possibility of efficient algorithms for linear programming over  $P$ .

Polytopes are special cases of closed convex sets and one can study lifts in this more general context. All convex sets are slices of closed convex cones by affine planes and hence we will look at lifts of convex sets that have this form. Formally, given a closed convex cone  $K \subset \mathbb{R}^m$ , an affine plane  $L \subset \mathbb{R}^m$ , and a convex set  $C \subset \mathbb{R}^n$ , we say that  $K \cap L$  is a  $K$ -lift of  $C$  if  $C = \pi(K \cap L)$  for some linear map  $\pi : \mathbb{R}^m \rightarrow \mathbb{R}^n$ . Recall that every polytope is an affine slice of a nonnegative orthant  $\mathbb{R}_+^k$  and hence polyhedral lifts of polytopes, as we saw in Examples 1.1 and 1.2, are special cases of cone lifts. A polytope can also have non-polyhedral lifts.

The main source of non-polyhedral lifts in this paper will come from the *positive semidefinite cone*  $\mathcal{S}_+^k$  of  $k \times k$  real symmetric positive semidefinite (psd) matrices. If a matrix  $X$  is psd, we write  $X \succeq 0$ . An affine slice of  $\mathcal{S}_+^k$  is called a *spectrahedron* of size  $k$ . If a spectrahedron (of size  $k$ ) is a lift of a convex set  $C$ , we say that  $C$  admits a *spectrahedral* or *psd lift* (of size  $k$ ). It is also common to say that  $C$  is *sdp representable* or a *projected spectrahedron* or a *spectrahedral shadow*. Note that a spectrahedron in  $\mathcal{S}_+^k$  can also be written in the form

$$\left\{ x \in \mathbb{R}^t : A_0 + \sum_{i=1}^t A_i x_i \succeq 0 \right\}$$

where  $A_0, A_1, \dots, A_t$  are real symmetric matrices of size  $k$ .

**Example 1.3.** The square  $P \subset \mathbb{R}^2$  with vertices  $(\pm 1, \pm 1)$  can be expressed as the projection of a spectrahedron as follows:

$$P = \left\{ (x, y) \in \mathbb{R}^2 : \exists z \in \mathbb{R} \text{ s.t. } \begin{pmatrix} 1 & x & y \\ x & 1 & z \\ y & z & 1 \end{pmatrix} \succeq 0 \right\}.$$

The spectrahedral lift in this example is known as the *elliptope* and is shown in Figure 2. It consists of all  $X \in \mathcal{S}_+^3$  such that  $X_{ii} = 1$  for  $i = 1, 2, 3$ .  $\square$

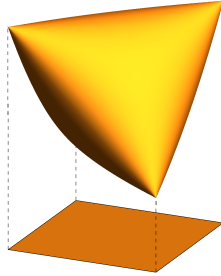


Figure 2: A spectrahedral lift of a square.

**Example 1.4.** Given a graph  $G = ([n], E)$  with vertex set  $[n]$  and edge set  $E$ , a collection  $S \subseteq [n]$  is called a *stable set* if for each  $i, j \in S$ , the pair  $\{i, j\} \notin E$ . Each stable set  $S$  is uniquely identified by its incidence vector  $\chi^S \in \{0, 1\}^n$  defined as  $(\chi^S)_i = 1$  if  $i \in S$  and 0 otherwise. The *stable set polytope* of  $G$  is

$$\text{STAB}(G) := \text{conv}\{\chi^S : S \text{ stable set in } G\}$$

where  $\text{conv}$  denotes convex hull. For  $x = \chi^S$ , consider the rank one matrix in  $\mathcal{S}_+^{n+1}$

$$\begin{pmatrix} 1 \\ x \end{pmatrix} (1 \ x^\top) = \begin{pmatrix} 1 & x^\top \\ x & xx^\top \end{pmatrix} = \begin{pmatrix} 1 & x^\top \\ x & U \end{pmatrix}.$$

Since  $\chi^S \in \{0, 1\}^n$ ,  $U_{ii} = x_i$  for all  $i \in [n]$ , and since  $S$  is stable,  $U_{ij} = 0$  for all  $\{i, j\} \in E$ . Therefore, the convex set

$$\text{TH}(G) := \left\{ x \in \mathbb{R}^n : \begin{array}{l} \exists U \in \mathcal{S}_+^n \text{ s.t. } \begin{pmatrix} 1 & x^\top \\ x & U \end{pmatrix} \succeq 0, \\ U_{ii} = x_i \ \forall i \in [n], \\ U_{ij} = 0 \ \forall \{i, j\} \in E \end{array} \right\}$$

known as the *theta body* of  $G$ , contains all the vertices of  $\text{STAB}(G)$ , and hence by convexity, all of  $\text{STAB}(G)$ . In general, this containment is strict. The theta body  $\text{TH}(G)$  is the projection onto  $x$ -coordinates of the set of all matrices in  $\mathcal{S}_+^{n+1}$  whose entries satisfy a set of linear constraints. The latter is a spectrahedron.

Theta bodies of graphs were defined in [Lovász \[1979\]](#). He proved that  $\text{STAB}(G) = \text{TH}(G)$  if and only if  $G$  is a perfect graph. Even for perfect graphs,  $\text{STAB}(G)$  can have exponentially many facets, but by Lovász's result, it admits a spectrahedral lift of size  $n + 1$ .  $\square$

We close the introduction with a psd lift of a non-polytopal convex set. Since polyhedra can only project to polyhedra, any lift of a non-polyhedral convex set is necessarily non-polyhedral.

**Example 1.5.** Let  $X$  be the  $n \times n$  symbolic matrix with entries  $x_1, \dots, x_{n^2}$  written consecutively along its  $n$  rows. and let  $I_n$  denote the  $n \times n$  identity matrix. Consider the spectrahedron of size  $2n$  defined by the conditions

$$\begin{pmatrix} Y & X \\ X^\top & I_n \end{pmatrix} \succeq 0, \quad \text{trace}(Y) = 1.$$

The psd condition is equivalent to  $Y - XX^\top \succeq 0$  via Schur complement. Taking the trace on both sides we get  $1 = \text{trace}(Y) \geq \text{trace}(XX^\top) = \sum_{i=1}^{n^2} x_i^2$ . Thus, the projection of the above spectrahedron onto  $x = (x_1, \dots, x_{n^2})$  is contained in the unit ball

$$B_{n^2} := \{(x_1, \dots, x_{n^2}) : \sum x_i^2 \leq 1\}.$$

On the other hand, for any  $x$  on the boundary of  $B_{n^2}$ , the matrix

$$\begin{pmatrix} XX^\top & X \\ X^\top & I_n \end{pmatrix}$$

lies in the above spectrahedron and projects onto  $x$ . We conclude that  $B_{n^2}$  has a spectral lift of size  $O(n)$ .  $\square$

In many of the above cases, projections offer a more compact representation of the convex set in question compared to the natural representation the set came with. Two fundamental questions we can ask now are the following.

*Question 1.6.* Given a convex set  $C \subset \mathbb{R}^n$  and a closed convex cone  $K \subset \mathbb{R}^m$ , does  $C$  admit a  $K$ -lift?

*Question 1.7.* If  $K$  comes from a family of cones  $\{K_t \subset \mathbb{R}^t\}$  such as the set of all positive orthants or the set of all psd cones, what is the smallest  $t$  for which  $C$  admits a  $K_t$ -lift? The smallest such  $t$  is a measure of complexity of  $C$ .

We will address both these questions and discuss several further related directions and results. In [Section 2](#) we prove that the existence of a  $K$ -lift for a convex set  $C$  is controlled by the existence of a  $K$ -factorization of an operator associated to  $C$ . This result specializes nicely to polytopes as we will see in [Section 3](#). These factorization theorems generalize a celebrated result of [Yannakakis \[1991\]](#) about polyhedral lifts of polytopes. The rest of



the sections are focussed on spectrahedral lifts of convex sets. In [Section 4](#) we define the notion of positive semidefinite rank (psd rank) of a convex set and explain the known bounds on this invariant. We also mention recent results about psd ranks of certain families of convex sets. The psd rank of an  $n$ -dimensional polytope is known to be at least  $n + 1$ . In [Section 5](#), we explore the class of polytopes that have this minimum possible psd rank. We conclude in [Section 6](#) with the basic connections between sum of squares polynomials and spectrahedral lifts. We also describe the recent breakthrough by Scheiderer that provides the first examples of convex semialgebraic sets that do not admit spectrahedral lifts.

## 2 The Factorization Theorem for Convex Sets

A convex set is called a *convex body* if it is compact and contains the origin in its interior. For simplicity, we will always assume that all our convex sets are convex bodies. Recall that the *polar* of a convex set  $C \subset \mathbb{R}^n$  is the set

$$C^\circ = \{y \in \mathbb{R}^n : \langle x, y \rangle \leq 1, \forall x \in C\}.$$

Let  $\text{ext}(C)$  denote the set of *extreme points* of  $C$ , namely, all points  $p \in C$  such that if  $p = (p_1 + p_2)/2$ , with  $p_1, p_2 \in C$ , then  $p = p_1 = p_2$ . Both  $C$  and  $C^\circ$  are convex hulls of their respective extreme points. Consider the operator  $S : \mathbb{R}^n \times \mathbb{R}^n \rightarrow \mathbb{R}$  defined by  $S(x, y) = 1 - \langle x, y \rangle$ . The *slack operator*  $S_C$ , of a convex set  $C \subset \mathbb{R}^n$ , is the restriction of the operator  $S$  to  $\text{ext}(C) \times \text{ext}(C^\circ)$ . Note that the range of  $S_C$  is contained in  $\mathbb{R}_+$ , the set of nonnegative real numbers.

**Definition 2.1.** Let  $K \subset \mathbb{R}^m$  be a full-dimensional closed convex cone and  $C \subset \mathbb{R}^n$  a full-dimensional convex body. A *K-lift* of  $C$  is a set  $Q = K \cap L$ , where  $L \subset \mathbb{R}^m$  is an affine subspace, and  $\pi : \mathbb{R}^m \rightarrow \mathbb{R}^n$  is a linear map such that  $C = \pi(Q)$ . If  $L$  intersects the interior of  $K$  we say that  $Q$  is a *proper K-lift* of  $C$ .

We will see that the existence of a  $K$ -lift of  $C$  is intimately related to properties of the slack operator  $S_C$ . Recall that the *dual* of a closed convex cone  $K \subset \mathbb{R}^m$  is

$$K^* = \{y \in \mathbb{R}^m : \langle x, y \rangle \geq 0, \forall x \in K\}.$$

A cone  $K$  is *self-dual* if  $K^* = K$ . The cones  $\mathbb{R}_+^n$  and  $\mathcal{S}_+^k$  are self-dual.

**Definition 2.2.** Let  $C$  and  $K$  be as in [Definition 2.1](#). We say that the slack operator  $S_C$  is *K-factorizable* if there exist maps (not necessarily linear)

$$A : \text{ext}(C) \rightarrow K \quad \text{and} \quad B : \text{ext}(C^\circ) \rightarrow K^*$$

such that  $S_C(x, y) = \langle A(x), B(y) \rangle$  for all  $(x, y) \in \text{ext}(C) \times \text{ext}(C^\circ)$ .

We can now characterize the existence of a  $K$ -lift of  $C$  in terms of the operator  $S_C$ , answering [Question 1.6](#). The proof relies on the theory of *convex cone programming* which is the problem of optimizing a linear function over an affine slice of a closed convex cone, see [Ben-Tal and Nemirovski \[2001\]](#), or [Blekherman, Parrilo, and Thomas \[2013, §2.1.4\]](#) for a quick introduction.

**Theorem 2.3.** *Gouveia, Parrilo, and Thomas [2013, Theorem 1] If  $C$  has a proper  $K$ -lift then  $S_C$  is  $K$ -factorizable. Conversely, if  $S_C$  is  $K$ -factorizable then  $C$  has a  $K$ -lift.*

*Proof.* Suppose  $C$  has a proper  $K$ -lift. Then there exists an affine space  $L = w_0 + L_0$  in  $\mathbb{R}^m$  ( $L_0$  is a linear subspace) and a linear map  $\pi : \mathbb{R}^m \rightarrow \mathbb{R}^n$  such that  $C = \pi(K \cap L)$  and  $w_0 \in \text{int}(K)$ . Equivalently,

$$C = \{x \in \mathbb{R}^n : x = \pi(w), \quad w \in K \cap (w_0 + L_0)\}.$$

We need to construct the maps  $A : \text{ext}(C) \rightarrow K$  and  $B : \text{ext}(C^\circ) \rightarrow K^*$  that factorize the slack operator  $S_C$ , from the  $K$ -lift of  $C$ . For  $x_i \in \text{ext}(C)$ , define  $A(x_i) := w_i$ , where  $w_i$  is any point in the non-empty convex set  $\pi^{-1}(x_i) \cap K \cap L$ .

Let  $c$  be an extreme point of  $C^\circ$ . Then  $\max\{\langle c, x \rangle : x \in C\} = 1$  since  $\langle c, x \rangle \leq 1$  for all  $x \in C$ , and if the maximum was smaller than one, then  $c$  would not be an extreme point of  $C^\circ$ . Let  $M$  be a full row rank matrix such that  $\ker M = L_0$ . Then the following hold:

$$1 = \max_{x \in C} \langle c, x \rangle = \max_{w \in K \cap (w_0 + L_0)} \langle c, \pi(w) \rangle = \max_{\substack{w \in K \\ Mw = Mw_0}} \langle \pi^*(c), w \rangle$$

Since  $w_0$  lies in the interior of  $K$ , by Slater's condition we have strong duality for the above cone program, and we get

$$1 = \min \langle Mw_0, y \rangle : M^T y - \pi^*(c) \in K^*$$

with the minimum being attained. Further, setting  $z = M^T y$  we have that

$$1 = \min \langle w_0, z \rangle : z - \pi^*(c) \in K^*, z \in L_0^\perp$$

with the minimum being attained. Now define  $B : \text{ext}(C^\circ) \rightarrow K^*$  as the map that sends  $y_i \in \text{ext}(C^\circ)$  to  $B(y_i) := z - \pi^*(y_i)$ , where  $z$  is any point in the nonempty convex set  $L_0^\perp \cap (K^* + \pi^*(y_i))$  that satisfies  $\langle w_0, z \rangle = 1$ . Note that for such a  $z$ ,  $\langle w_i, z \rangle = 1$  for all  $w_i \in L$ . Then  $B(y_i) \in K^*$ , and for an  $x_i \in \text{ext}(C)$ ,

$$\begin{aligned} \langle x_i, y_i \rangle &= \langle \pi(w_i), y_i \rangle = \langle w_i, \pi^*(y_i) \rangle = \langle w_i, z - B(y_i) \rangle \\ &= 1 - \langle w_i, B(y_i) \rangle = 1 - \langle A(x_i), B(y_i) \rangle. \end{aligned}$$

Therefore,  $S_C(x_i, y_i) = 1 - \langle x_i, y_i \rangle = \langle A(x_i), B(y_i) \rangle$  for all  $x_i \in \text{ext}(C)$  and  $y_i \in \text{ext}(C^\circ)$ .

Suppose now  $S_C$  is  $K$ -factorizable, i.e., there exist maps  $A : \text{ext}(C) \rightarrow K$  and  $B : \text{ext}(C^\circ) \rightarrow K^*$  such that  $S_C(x, y) = \langle A(x), B(y) \rangle$  for all  $(x, y) \in \text{ext}(C) \times \text{ext}(C^\circ)$ . Consider the affine space

$$L = \{(x, z) \in \mathbb{R}^n \times \mathbb{R}^m : 1 - \langle x, y \rangle = \langle z, B(y) \rangle, \forall y \in \text{ext}(C^\circ)\},$$

and let  $L_K$  be its coordinate projection into  $\mathbb{R}^m$ . Note that  $0 \notin L_K$  since otherwise, there exists  $x \in \mathbb{R}^n$  such that  $1 - \langle x, y \rangle = 0$  for all  $y \in \text{ext}(C^\circ)$  which implies that  $C^\circ$  lies in the affine hyperplane  $\langle x, y \rangle = 1$ . This is a contradiction since  $C^\circ$  contains the origin. Also,  $K \cap L_K \neq \emptyset$  since for each  $x \in \text{ext}(C)$ ,  $A(x) \in K \cap L_K$  by assumption.

Let  $x$  be some point in  $\mathbb{R}^n$  such that there exists some  $z \in K$  for which  $(x, z)$  is in  $L$ . Then, for all extreme points  $y$  of  $C^\circ$  we will have that  $1 - \langle x, y \rangle$  is nonnegative. This implies, using convexity, that  $1 - \langle x, y \rangle$  is nonnegative for all  $y$  in  $C^\circ$ , hence  $x \in (C^\circ)^\circ = C$ .

We now argue that this implies that for each  $z \in K \cap L_K$  there exists a unique  $x_z \in \mathbb{R}^n$  such that  $(x_z, z) \in L$ . That there is one, comes immediately from the definition of  $L_K$ . Suppose now that there is another such point  $x'_z$ . Then  $(tx_z + (1-t)x'_z, z) \in L$  for all reals  $t$  which would imply that the line through  $x_z$  and  $x'_z$  would be contained in  $C$ , contradicting our assumption that  $C$  is compact.

The map that sends  $z$  to  $x_z$  is therefore well-defined in  $K \cap L_K$ , and can be easily checked to be affine. Since the origin is not in  $L_K$ , we can extend it to a linear map  $\pi : \mathbb{R}^m \rightarrow \mathbb{R}^n$ . To finish the proof it is enough to show  $C = \pi(K \cap L_K)$ . We have already seen that  $\pi(K \cap L_K) \subseteq C$  so we just have to show the reverse inclusion. For all extreme points  $x$  of  $C$ ,  $A(x)$  belongs to  $K \cap L_K$ , and therefore,  $x = \pi(A(x)) \in \pi(K \cap L_K)$ . Since  $C = \text{conv}(\text{ext}(C))$  and  $\pi(K \cap L_K)$  is convex,  $C \subseteq \pi(K \cap L_K)$ .  $\square$

The asymmetry in the two directions of [Theorem 2.3](#) disappears for many nice cones including  $\mathbb{R}_+^k$  and  $\mathbb{S}_+^k$ . For more on this, see [Gouveia, Parrilo, and Thomas \[2013, Corollary 1\]](#). In these nice cases,  $C$  has a  $K$ -lift if and only if  $S_C$  has a  $K$ -factorization. [Theorem 2.3](#) generalizes the original factorization theorem of Yannakakis for polyhedral lifts of polytopes [Yannakakis \[1991, Theorem 3, §4\]](#) to arbitrary cone lifts of convex sets.

Recall that in the psd cone  $\mathbb{S}_+^k$ , the inner product  $\langle A, B \rangle = \text{trace}(AB)$ .

**Example 2.4.** The unit disk  $C \subset \mathbb{R}^2$  is a spectrahedron in  $\mathbb{S}_+^2$  as follows

$$C = \left\{ (x, y) \in \mathbb{R}^2 : \begin{pmatrix} 1+x & y \\ y & 1-x \end{pmatrix} \geq 0 \right\},$$

and hence trivially has a  $\mathbb{S}_+^2$ -lift. This means that the slack operator  $S_C$  must have a  $\mathbb{S}_+^2$ -factorization. Since  $C^\circ = C$ ,  $\text{ext}(C) = \text{ext}(C^\circ) = \partial C$ , and so we have to find maps

$A, B : \text{ext}(C) \rightarrow \mathcal{S}_+^2$  such that for all  $(x_1, y_1), (x_2, y_2) \in \text{ext}(C)$ ,

$$\langle A(x_1, y_1), B(x_2, y_2) \rangle = 1 - x_1 x_2 - y_1 y_2.$$

This is accomplished by the maps

$$A(x_1, y_1) = \begin{pmatrix} 1 + x_1 & y_1 \\ y_1 & 1 - x_1 \end{pmatrix}$$

and

$$B(x_2, y_2) = \frac{1}{2} \begin{pmatrix} 1 - x_2 & -y_2 \\ -y_2 & 1 + x_2 \end{pmatrix}$$

which factorize  $S_C$  and are positive semidefinite in their domains.  $\square$

**Example 2.5.** Consider the spectrahedral lift of the unit ball  $B_{n^2}$  from [Example 1.5](#). Again, we have that  $\text{ext}(B_{n^2}) = \text{ext}(B_{n^2}^\circ) = \partial B_{n^2}$ . The maps

$$A(x) = \begin{pmatrix} XX^\top & X \\ X^\top & I_n \end{pmatrix}, \quad B(y) = \frac{1}{2} \begin{pmatrix} I_n & -Y \\ -Y^\top & YY^\top \end{pmatrix}$$

where  $X$  is defined as in [Example 1.5](#) and  $Y$  is defined the same way, offer a  $\mathcal{S}_+^{2n}$ -factorization of the slack operator of  $B_{n^2}$ .  $\square$

The existence of cone lifts of convex bodies is preserved under many geometric operations [Gouveia, Parrilo, and Thomas \[2013, Propositions 1 and 2\]](#). For instance, if  $C$  has a  $K$ -lift, then so does any compact image of  $C$  under a projective transformation. An elegant feature of this theory is that the existence of lifts is invariant under polarity/duality;  $C$  has a  $K$ -lift if and only if  $C^\circ$  has a  $K^*$ -lift. In particular, if  $C$  has a spectrahedral lift of size  $k$ , then so does  $C^\circ$ .

### 3 The Factorization Theorem for Polytopes

When the convex body  $C$  is a polytope, [Theorem 2.3](#) becomes rather simple. This specialization also appeared in [Fiorini, Massar, Pokutta, Tiwary, and de Wolf \[2012\]](#).

**Definition 3.1.** Let  $P$  be a full-dimensional polytope in  $\mathbb{R}^n$  with vertex set  $V_P = \{p_1, \dots, p_v\}$  and an irredundant inequality representation

$$P = \{x \in \mathbb{R}^n : h_1(x) \geq 0, \dots, h_f(x) \geq 0\}.$$

Since  $P$  is a convex body, we may assume that the constant in each  $h_j(x)$  is 1. The *slack matrix* of  $P$  is the nonnegative  $v \times f$  matrix whose  $(i, j)$ -entry is  $h_j(p_i)$ , the *slack* of vertex  $p_i$  in the facet inequality  $h_j(x) \geq 0$ .

When  $P$  is a polytope,  $\text{ext}(P)$  is just  $V_P$ , and  $\text{ext}(P^\circ)$  is in bijection with  $F_P$ , the set of facets of  $P$ . The facet  $F_j$  is defined by  $h_j(x) \geq 0$  and  $f := |F_P|$ . Then the slack operator  $S_P$  is the map from  $V_P \times F_P$  to  $\mathbb{R}_+$  that sends the vertex facet pair  $(p_i, F_j)$  to  $h_j(p_i)$ . Hence, we may identify the slack operator of  $P$  with the slack matrix of  $P$  and use  $S_P$  to also denote this matrix. Since the facet inequalities of  $P$  are only unique up to multiplication by positive scalars, the matrix  $S_P$  is also only unique up to multiplication of its columns by positive scalars. Regardless, we will call  $S_P$ , derived from the given presentation of  $P$ , the slack matrix of  $P$ .

**Definition 3.2.** Let  $M = (M_{ij}) \in \mathbb{R}_+^{p \times q}$  be a nonnegative matrix and  $K$  a closed convex cone. Then a  $K$ -factorization of  $M$  is a pair of ordered sets  $\{a^1, \dots, a^p\} \subset K$  and  $\{b^1, \dots, b^q\} \subset K^*$  such that  $\langle a^i, b^j \rangle = M_{ij}$ .

Note that  $M \in \mathbb{R}_+^{p \times q}$  has a  $\mathbb{R}_+^k$ -factorization if and only if there exist a  $p \times k$  nonnegative matrix  $A$  and a  $k \times q$  nonnegative matrix  $B$  such that  $M = AB$ , called a *nonnegative factorization* of  $M$ . [Definition 3.2](#) generalizes nonnegative factorizations of nonnegative matrices to cone factorizations.

**Theorem 3.3.** *If a full-dimensional polytope  $P$  has a proper  $K$ -lift then every slack matrix of  $P$  admits a  $K$ -factorization. Conversely, if some slack matrix of  $P$  has a  $K$ -factorization then  $P$  has a  $K$ -lift.*

[Theorem 3.3](#) is a direct translation of [Theorem 2.3](#) using the identification between the slack operator of  $P$  and the slack matrix of  $P$ . The original theorem of [Yannakakis \[1991, Theorem 3, §4\]](#) proved this result in the case where  $K = \mathbb{R}_+^k$ .

**Example 3.4.** Consider the regular hexagon with inequality description

$$H = \left\{ (x_1, x_2) \in \mathbb{R}^2 : \begin{pmatrix} 1 & \sqrt{3}/3 \\ 0 & 2\sqrt{3}/3 \\ -1 & \sqrt{3}/3 \\ -1 & -\sqrt{3}/3 \\ 0 & -2\sqrt{3}/3 \\ 1 & -\sqrt{3}/3 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \leq \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \right\}.$$

We will denote the coefficient matrix by  $F$  and the right hand side vector by  $d$ . It is easy to check that  $H$  cannot be the projection of an affine slice of  $\mathbb{R}_+^k$  for  $k < 5$ . Therefore, we ask whether it can be the linear image of an affine slice of  $\mathbb{R}_+^5$ . Using [Theorem 3.3](#) this is

equivalent to asking if the slack matrix of the hexagon,

$$S_H := \begin{pmatrix} 0 & 0 & 1 & 2 & 2 & 1 \\ 1 & 0 & 0 & 1 & 2 & 2 \\ 2 & 1 & 0 & 0 & 1 & 2 \\ 2 & 2 & 1 & 0 & 0 & 1 \\ 1 & 2 & 2 & 1 & 0 & 0 \\ 0 & 1 & 2 & 2 & 1 & 0 \end{pmatrix},$$

has a  $\mathbb{R}_+^5$ -factorization. Check that

$$S_H = \begin{pmatrix} 1 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 1 & 2 \\ 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 2 & 1 & 0 \end{pmatrix} \begin{pmatrix} 0 & 0 & 0 & 1 & 2 & 1 \\ 1 & 2 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 0 & 0 & 1 \end{pmatrix},$$

where we call the first matrix  $A$  and the second matrix  $B$ . We may take the rows of  $A$  as elements of  $\mathbb{R}_+^5$ , and the columns of  $B$  as elements of  $\mathbb{R}_+^5 = (\mathbb{R}_+^5)^*$ , and they provide us a  $\mathbb{R}_+^5$ -factorization of the slack matrix  $S_H$ , proving that this hexagon has a  $\mathbb{R}_+^5$ -lift while the trivial polyhedral lift would have been to  $\mathbb{R}_+^6$ .

We can construct the lift using the proof of the [Theorem 2.3](#). Note that

$$H = \{(x_1, x_2) \in \mathbb{R}^2 : \exists y \in \mathbb{R}_+^5 \text{ s.t. } Fx + B^T y = d\}.$$

Hence, the exact slice of  $\mathbb{R}_+^5$  that is mapped to the hexagon is simply

$$\{y \in \mathbb{R}_+^5 : \exists x \in \mathbb{R}^2 \text{ s.t. } B^T y = d - Fx\}.$$

By eliminating the  $x$  variables in the system we get

$$\{y \in \mathbb{R}_+^5 : y_1 + y_2 + y_3 + y_5 = 2, y_3 + y_4 + y_5 = 1\},$$

and so we have a three dimensional slice of  $\mathbb{R}_+^5$  projecting down to  $H$ . This projection is visualized in [Figure 3](#).

The hexagon is a good example to see that the existence of lifts depends on more than the combinatorics of the polytope. If instead of a regular hexagon we take the hexagon

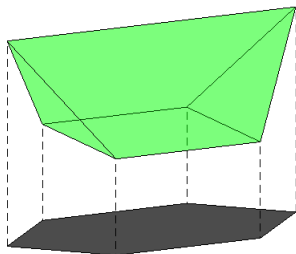


Figure 3: A  $\mathbb{R}_+^5$ -lift of the regular hexagon.

with vertices  $(0, -1)$ ,  $(1, -1)$ ,  $(2, 0)$ ,  $(1, 3)$ ,  $(0, 2)$  and  $(-1, 0)$ , a valid slack matrix would be

$$S := \begin{pmatrix} 0 & 0 & 1 & 4 & 3 & 1 \\ 1 & 0 & 0 & 4 & 4 & 3 \\ 7 & 4 & 0 & 0 & 4 & 9 \\ 3 & 4 & 4 & 0 & 0 & 1 \\ 3 & 5 & 6 & 1 & 0 & 0 \\ 0 & 1 & 3 & 5 & 3 & 0 \end{pmatrix}.$$

One can check that if a  $6 \times 6$  matrix with the zero pattern of a slack matrix of a hexagon has a  $\mathbb{R}_+^5$ -factorization, then it has a factorization with either the same zero pattern as the matrices  $A$  and  $B$  obtained before, or the patterns given by applying a cyclic permutation to the rows of  $A$  and the columns of  $B$ . A simple algebraic computation then shows that the slack matrix  $S$  above has no such decomposition hence this irregular hexagon has no  $\mathbb{R}_+^5$ -lift.  $\square$

**Example 3.5.** In [Example 1.3](#) we saw a  $\mathbb{R}_+^3$ -lift of a square  $P$ . Up to scaling of columns by positive numbers, the slack matrix of  $P$  is

$$S_P = \begin{pmatrix} 0 & 0 & 1 & 1 \\ 0 & 1 & 1 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

where the rows are associated to the vertices  $(1, 1)$ ,  $(1, -1)$ ,  $(-1, -1)$ ,  $(-1, 1)$  in that order, and the columns to the facets defined by the inequalities

$$1 - x_1 \geq 0, \quad 1 - x_2 \geq 0, \quad 1 + x_1 \geq 0, \quad 1 + x_2 \geq 0.$$

The matrix  $S_P$  admits the following  $\mathcal{S}_+^3$ -factorization where the first four matrices are associated to the rows of  $S_P$  and the next four matrices are associated to the columns of  $S_P$ :

$$\begin{pmatrix} 1 & 1 & 1 \\ 1 & 1 & 1 \\ 1 & 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & 1 & -1 \\ 1 & 1 & -1 \\ -1 & -1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & -1 & -1 \\ -1 & 1 & 1 \\ -1 & 1 & 1 \end{pmatrix}, \begin{pmatrix} 1 & -1 & 1 \\ -1 & 1 & -1 \\ 1 & -1 & 1 \end{pmatrix} \\ \frac{1}{4} \begin{pmatrix} 1 & -1 & 0 \\ -1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \frac{1}{4} \begin{pmatrix} 1 & 0 & -1 \\ 0 & 0 & 0 \\ -1 & 0 & 1 \end{pmatrix}, \frac{1}{4} \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \frac{1}{4} \begin{pmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 1 & 0 & 1 \end{pmatrix}.$$

□

## 4 Positive Semidefinite Rank

From now on we focus on the special case of spectrahedral lifts of convex sets. Since the family of psd cones  $\{\mathcal{S}_+^k : k \in \mathbb{N}\}$  is closed in the sense that any face of a member  $\mathcal{S}_+^i$  in the family is isomorphic to  $\mathcal{S}_+^j$  for some  $j \leq i$ , we can look at the smallest index  $k$  for which a convex set  $C$  admits a  $\mathcal{S}_+^k$ -lift.

**Definition 4.1.** The *psd rank* of a convex set  $C \subset \mathbb{R}^n$ , denoted as  $\text{rank}_{\text{psd}}(C)$  is the smallest positive integer  $k$  such that  $C = \pi(\mathcal{S}_+^k \cap L)$  for some affine space  $L$  and linear map  $\pi$ . If  $C$  does not admit a psd lift, then define  $\text{rank}_{\text{psd}}(C) = \infty$ .

The following lemma is immediate from the previous sections and offers an explicit tool for establishing psd ranks.

**Lemma 4.2.** *The psd rank of a convex set  $C$  is the smallest  $k$  for which the slack operator  $S_C$  admits a  $\mathcal{S}_+^k$ -factorization. If  $P$  is a polytope, then  $\text{rank}_{\text{psd}}(P)$  is the smallest integer  $k$  for which the slack matrix  $S_P$  admits a  $\mathcal{S}_+^k$ -factorization.*

Following Definition 3.2, for any nonnegative matrix  $M \in \mathbb{R}_+^{p \times q}$ , one can define  $\text{rank}_{\text{psd}}(M)$  to be the smallest integer  $k$  such that  $M$  admits a  $\mathcal{S}_+^k$ -factorization. The relationship between  $\text{rank}_{\text{psd}}(M)$  and  $\text{rank}(M)$  is as follows:

$$(1) \quad \frac{1}{2} \left( \sqrt{1 + 8 \text{rank}(M)} - 1 \right) \leq \text{rank}_{\text{psd}}(M) \leq \min\{p, q\}.$$

For a proof, as well as a comprehensive comparison between psd rank and several other notions of rank of a nonnegative matrix, see [Fawzi, Gouveia, Parrilo, Robinson, and Thomas \[2015\]](#).

The goal of this section is to describe the known bounds on psd ranks of convex sets. As might be expected, the best results we have are for polytopes.



**4.1 Polytopes.** In the case of polytopes, there is a simple lower bound on psd rank. The proof relies on the following technique to increase the psd rank of a matrix by one.

**Lemma 4.3.** *Gouveia, Robinson, and Thomas [2013, Proposition 2.6]* Suppose  $M \in \mathbb{R}_+^{p \times q}$  and  $\text{rank}_{\text{psd}}(M) = k$ . If  $M$  is extended to  $M' = \begin{pmatrix} M & \mathbf{0} \\ w & \alpha \end{pmatrix}$  where  $w \in \mathbb{R}_+^q$ ,  $\alpha > 0$  and  $\mathbf{0}$  is a column of zeros, then  $\text{rank}_{\text{psd}}(M') = k + 1$ . Further, the factor associated to the last column of  $M'$  in any  $\mathcal{S}_+^{k+1}$ -factorization of  $M'$  has rank one.

**Theorem 4.4.** *Gouveia, Robinson, and Thomas [ibid., Proposition 3.2]* If  $P \subset \mathbb{R}^n$  is a full-dimensional polytope, then the psd rank of  $P$  is at least  $n + 1$ . If  $\text{rank}_{\text{psd}}(P) = n + 1$ , then every  $\mathcal{S}_+^{n+1}$ -factorization of the slack matrix of  $P$  only uses rank one matrices as factors.

*Proof.* The proof is by induction on  $n$ . If  $n = 1$ , then  $P$  is a line segment and we may assume that its vertices are  $p_1, p_2$  and facets are  $F_1, F_2$  with  $p_1 = F_2$  and  $p_2 = F_1$ . Hence its slack matrix is a  $2 \times 2$  diagonal matrix with positive diagonal entries. It is not hard to see that  $\text{rank}_{\text{psd}}(S_P) = 2$  and any  $\mathcal{S}_+^2$ -factorization of it uses only matrices of rank one.

Assume the first statement in the theorem holds up to dimension  $n - 1$  and consider a polytope  $P \subset \mathbb{R}^n$  of dimension  $n$ . Let  $F$  be a facet of  $P$  with vertices  $p_1, \dots, p_s$ , facets  $f_1, \dots, f_t$  and slack matrix  $S_F$ . Suppose  $f_i$  corresponds to facet  $F_i$  of  $P$  for  $i = 1, \dots, t$ . By induction hypothesis,  $\text{rank}_{\text{psd}}(F) = \text{rank}_{\text{psd}}(S_F) \geq n$ . Let  $p$  be a vertex of  $P$  not in  $F$  and assume that the top left  $(s + 1) \times (t + 1)$  submatrix of  $S_P$  is indexed by  $p_1, \dots, p_s, p$  in the rows and  $F_1, \dots, F_t, F$  in the columns. Then this submatrix of  $S_P$ , which we will call  $S'_F$ , has the form

$$S'_F = \begin{pmatrix} S_F & \mathbf{0} \\ * & \alpha \end{pmatrix}$$

with  $\alpha > 0$ . By Lemma 4.3, the psd rank of  $S'_F$  is at least  $n + 1$  since the psd rank of  $S_F$  is at least  $n$ . Hence,  $\text{rank}_{\text{psd}}(P) = \text{rank}_{\text{psd}}(S_P) \geq n + 1$ .

Suppose there is now a  $\mathcal{S}_+^{n+1}$ -factorization of  $S_P$  and therefore of  $S'_F$ . By Lemma 4.3 the factor corresponding to the facet  $F$  has rank one. Repeating the procedure for all facets  $F$  and all submatrices  $S'_F$  we get that all factors corresponding to the facets of  $P$  in this  $\mathcal{S}_+^{n+1}$ -factorization of  $S_P$  must have rank one. To prove that all factors indexed by the vertices of  $P$  also have rank one, we use the fact that the transpose of a slack matrix of  $P$  is (up to row scaling) a slack matrix of the polar polytope  $P^\circ$ , concluding the proof.  $\square$

For an  $n$ -dimensional polytope  $P \subset \mathbb{R}^n$ , it is well-known that  $\text{rank}(S_P) = n + 1$ , see for instance Gouveia, Robinson, and Thomas [ibid., Lemma 3.1]. Therefore, Theorem 4.4 implies that for a slack matrix  $S_P$  of a polytope  $P$  we have a simple relationship between

rank and psd rank, namely  $\text{rank}(S_P) \leq \text{rank}_{\text{psd}}(P)$ , as compared to (1). From (1) we also have that for a polytope  $P$  with  $v$  vertices and  $f$  facets,  $\text{rank}_{\text{psd}}(P) \leq \min\{v, f\}$ . In general, it is not possible to bound the psd rank of nonnegative matrices, even slack matrices, by a function in the rank of the matrix. For instance, all slack matrices of polygons have rank three. However, we will see as a consequence of the results in the next subsection that the psd rank of an  $n$ -gon grows with  $n$ .

In the next section we will see that the lower bound in Theorem 4.4 can be tight for several interesting classes of polytopes. Such polytopes include some 0/1-polytopes. However, Briët, Dadush and Pokutta showed that not all 0/1-polytopes can have small psd rank.

**Theorem 4.5.** *Briët, Dadush, and Pokutta [2015] For any  $n \in \mathbb{Z}_+$ , there exists  $U \subset \{0, 1\}^n$  such that*

$$\text{rank}_{\text{psd}}(\text{conv}(U)) = \Omega\left(\frac{2^{\frac{n}{4}}}{(n \log n)^{\frac{1}{4}}}\right).$$

Despite the above result, it is not easy to find explicit polytopes with high psd rank. The most striking results we have so far are the following by Lee, Raghavendra and Steurer, which provide super polynomial lower bounds on the psd rank of specific families of 0/1-polytopes.

**Theorem 4.6.** *Lee, Raghavendra, and Steurer [2015] The cut, TSP, and stable set polytopes of  $n$ -vertex graphs have psd rank at least  $2^{n^\delta}$ , for some constant  $\delta > 0$ .*

We saw the stable set polytope of an  $n$ -vertex graph before. The cut and TSP polytopes are other examples of polytopes that come from graph optimization problems. The TSP (traveling salesman problem) is the problem of finding a tour through all vertices of the  $n$ -vertex complete graph that minimizes a linear objective function. Each tour can be represented as a 0/1-vector in  $\{0, 1\}^{\binom{n}{2}}$  and the TSP polytope is the convex hull of all these tour vectors.

**4.2 General convex sets.** We now examine lower bounds on the psd rank of an arbitrary convex set  $C \subset \mathbb{R}^n$ . The following elegant lower bound was established by Fawzi and Safey El Din.

**Theorem 4.7.** *Fawzi and Safey El Din [2018] Suppose  $C \subset \mathbb{R}^n$  is a convex set and  $d$  is the minimum degree of a polynomial with real coefficients that vanishes on the boundary of  $C^\circ$ . Then  $\text{rank}_{\text{psd}}(C) \geq \sqrt{\log d}$ .*

The *algebraic degree* of a convex set  $C$  is the smallest degree of a polynomial with real coefficients that vanishes on the boundary of  $C$ . Suppose  $P$  is a polytope with  $v$  vertices

and the origin in its interior. Then  $P^\circ$  has  $v$  facets each corresponding to a linear polynomial  $l_i$  that vanishes on the facet. The polynomial  $p := \pi_{i=1}^v l_i$  vanishes on the boundary of  $P^\circ$  and has degree  $v$ . In fact, the algebraic degree of  $P^\circ$  is  $v$ . Hence by [Theorem 4.7](#),  $\text{rank}_{\text{psd}}(P) \geq \sqrt{\log v}$ . This result is analogous to an observation of [Goemans \[2015\]](#) that any polyhedral lift  $Q$  of  $P$  has at least  $\log v$  facets. The reason is that every vertex in  $P$  is the projection of a face of  $Q$  which in turn is the intersection of some set of facets of  $Q$ . Therefore,

$$v \leq \# \text{ faces of } Q \leq 2^{\# \text{ facets of } Q}.$$

Even for polytopes there are likely further factors from combinatorics and topology that can provide stronger lower bounds on psd rank.

The lower bound in [Theorem 4.7](#) is very explicit and simple, but it does not involve  $n$ . We now exhibit a simple lower bound that does.

**Proposition 4.8.** *Let  $C \subset \mathbb{R}^n$  be an  $n$ -dimensional convex body. Then  $\text{rank}_{\text{psd}}(C) = \Omega(\sqrt{n})$ .*

*Proof.* Suppose  $\text{rank}_{\text{psd}}(C) = k$ . Then there exists maps  $A : \text{ext}(C) \rightarrow \mathcal{S}_+^k$  and  $B : \text{ext}(C^\circ) \rightarrow \mathcal{S}_+^k$  such that for all  $(x, y) \in \text{ext}(C) \times \text{ext}(C^\circ)$ ,

$$(2) \quad S_C((x, y)) = 1 - \langle x, y \rangle = (1, x^\top) \cdot \begin{pmatrix} 1 \\ -y \end{pmatrix} = \text{trace}(A(x)B(y)).$$

Define  $\text{rank}(S_C)$  to be the minimum  $l$  such that  $S_C((x, y)) = a_x^\top b_y$  for  $a_x, b_y \in \mathbb{R}^l$ . Equality of the first and third expressions in (2) implies that  $\text{rank}(S_C) \leq n + 1$ . Now consider  $n + 1$  affinely independent extreme points  $x_1, \dots, x_{n+1}$  of  $C$  and  $n + 1$  affinely independent extreme points  $y_1, \dots, y_{n+1}$  of  $C^\circ$ . Then the values of  $S_C$  restricted to  $(x, y)$  as  $x$  and  $y$  vary in these chosen sets are the entries of the matrix

$$\begin{pmatrix} 1 & x_1^\top \\ \vdots & \\ 1 & x_{n+1}^\top \end{pmatrix} \begin{pmatrix} 1 & \cdots & 1 \\ -y_1 & \cdots & -y_{n+1} \end{pmatrix}$$

which has rank  $n + 1$ . Therefore,  $\text{rank}(S_C) = n + 1$ . Equality of the first and last expressions in (2) implies that the first inequality in (1) holds with  $M$  replaced by  $S_C$  via the same proof, see [Gouveia, Parrilo, and Thomas \[2013, Proposition 4\]](#). In other words,  $\frac{1}{2} \left( \sqrt{1 + 8(n + 1)} - 1 \right) \leq \text{rank}_{\text{psd}}(S_C) = \text{rank}_{\text{psd}}(C)$ , and we get the result.  $\square$

**Example 4.9.** The spectrahedral lift of  $B_{n^2}$  in [Example 1.5](#) is optimal, and  $\text{rank}_{\text{psd}}(B_{n^2}) = \Theta(n)$ .  $\square$

The lower bounds in [Theorem 4.7](#) and [Proposition 4.8](#) depend solely on the algebraic degree of  $C^\circ$  and  $n$  respectively. A question of interest is how the bound might jointly depend on both these parameters?

While the lower bounds in [Theorems 4.4, 4.7](#) and [Proposition 4.8](#) can be tight, we do not have much understanding of the psd ranks of specific polytopes or convex sets except in a few cases. For example, [Theorem 4.7](#) implies that the psd rank of polygons must grow to infinity as the number of vertices grows to infinity. However, we do not know if the psd rank of polygons is monotone in the number of vertices.

## 5 Psd-Minimal Polytopes

Recall from [Theorem 4.4](#) that the psd rank of an  $n$ -dimensional polytope is at least  $n + 1$ . In this section we study those polytopes whose psd rank is exactly this lower bound. Such polytopes are said to be *psd-minimal*. The key to understanding psd-minimality is another notion of rank of a nonnegative matrix.

**Definition 5.1.** A *Hadamard square root* of a nonnegative real matrix  $M$ , denoted as  $\sqrt{M}$ , is any matrix whose  $(i, j)$ -entry is a square root (positive or negative) of the  $(i, j)$ -entry of  $M$ .

Let  $\text{rank}_{\sqrt{}}(M) := \min\{\text{rank}(\sqrt{M})\}$  be the minimum rank of a Hadamard square root of a nonnegative matrix  $M$ . We recall the basic connection between the psd rank of a nonnegative matrix  $M$  and  $\text{rank}_{\sqrt{}}(M)$  shown in [Gouveia, Robinson, and Thomas \[2013, Proposition 2.2\]](#).

**Proposition 5.2.** *If  $M$  is a nonnegative matrix, then  $\text{rank}_{\text{psd}}(M) \leq \text{rank}_{\sqrt{}}(M)$ . In particular, the psd rank of a 0/1 matrix is at most the rank of the matrix.*

*Proof.* Let  $\sqrt{M}$  be a Hadamard square root of  $M \in \mathbb{R}_+^{p \times q}$  of rank  $r$ . Then there exist vectors  $a_1, \dots, a_p, b_1, \dots, b_q \in \mathbb{R}^r$  such that  $(\sqrt{M})_{ij} = \langle a_i, b_j \rangle$ . Therefore,  $M_{ij} = \langle a_i, b_j \rangle^2 = \langle a_i a_i^T, b_j b_j^T \rangle$  where the second inner product is the trace inner product for symmetric matrices defined earlier. Hence,  $\text{rank}_{\text{psd}}(M) \leq r$ .  $\square$

Even though  $\text{rank}_{\sqrt{}}(M)$  is only an upper bound on  $\text{rank}_{\text{psd}}(M)$ , we cannot find  $\mathcal{S}_+^k$ -factorizations of  $M$  with only rank one factors if  $k < \text{rank}_{\sqrt{}}(M)$ .

**Lemma 5.3.** [Gouveia, Robinson, and Thomas \[ibid., Lemma 2.4\]](#) *The smallest  $k$  for which a nonnegative real matrix  $M$  admits a  $\mathcal{S}_+^k$ -factorization in which all factors are matrices of rank one is  $k = \text{rank}_{\sqrt{}}(M)$ .*

*Proof.* If  $k = \text{rank}_{\sqrt{}}(M)$ , then there is a Hadamard square root of  $M \in \mathbb{R}_+^{p \times q}$  of rank  $k$  and the proof of [Proposition 5.2](#) gives a  $\mathcal{S}_+^k$ -factorization of  $M$  in which all factors

have rank one. On the other hand, if there exist  $a_1 a_1^T, \dots, a_p a_p^T, b_1 b_1^T, \dots, b_q b_q^T \in \mathcal{S}_+^k$  such that  $M_{ij} = \langle a_i a_i^T, b_j b_j^T \rangle = \langle a_i, b_j \rangle^2$ , then the matrix with  $(i, j)$ -entry  $\langle a_i, b_j \rangle$  is a Hadamard square root of  $M$  of rank at most  $k$ .  $\square$

This brings us to a characterization of psd-minimal polytopes.

**Theorem 5.4.** *If  $P \subset \mathbb{R}^n$  is a full-dimensional polytope, then  $\text{rank}_{\text{psd}}(P) = n + 1$  if and only if  $\text{rank}_{\sqrt{}}(S_P) = n + 1$ .*

*Proof.* By Proposition 5.2,  $\text{rank}_{\text{psd}}(P) \leq \text{rank}_{\sqrt{}}(S_P)$ . Therefore, if  $\text{rank}_{\sqrt{}}(S_P) = n + 1$ , then by Theorem 4.4, the psd rank of  $P$  is exactly  $n + 1$ .

Conversely, suppose  $\text{rank}_{\text{psd}}(P) = n + 1$ . Then there exists a  $\mathcal{S}_+^{n+1}$ -factorization of  $S_P$  which, by Theorem 4.4, has all factors of rank one. Thus, by Lemma 5.3, we have  $\text{rank}_{\sqrt{}}(S_P) \leq n + 1$ . Since  $\text{rank}_{\sqrt{}}$  is bounded below by  $\text{rank}_{\text{psd}}$ , we must have  $\text{rank}_{\sqrt{}}(S_P) = n + 1$ .  $\square$

Our next goal is to find psd-minimal polytopes. Recall that two polytopes  $P$  and  $Q$  are *combinatorially equivalent* if they have the same vertex-facet incidence structure. In this section we describe a simple algebraic obstruction to psd-minimality based on the combinatorics of a given polytope, therefore providing an obstruction for all polytopes in the given combinatorial class. Our main tool is a symbolic version of the slack matrix of a polytope.

**Definition 5.5.** The *symbolic slack matrix* of a  $d$ -polytope  $P$  is the matrix,  $S_P(x)$ , obtained by replacing all positive entries in the slack matrix  $S_P$  of  $P$  with distinct variables  $x_1, \dots, x_t$ .

Note that two  $d$ -polytopes  $P$  and  $Q$  are in the same combinatorial class if and only if  $S_P(x) = S_Q(x)$  up to permutations of rows and columns, and names of variables. Call a polynomial  $f \in \mathbb{R}[x_1, \dots, x_t]$  a *monomial* if it is of the form  $f = \pm x^a$  where  $x^a = x_1^{a_1} \cdots x_t^{a_t}$  and  $a = (a_1, \dots, a_t) \in \mathbb{N}^t$ . We refer to a sum of two distinct monomials as a *binomial* and to the sum of three distinct monomials as a *trinomial*. This differs from the usual terminology that allows nontrivial coefficients.

**Lemma 5.6** (Trinomial Obstruction Lemma). *Suppose the symbolic slack matrix  $S_P(x)$  of an  $n$ -polytope  $P$  has a  $(n + 2)$ -minor that is a trinomial. Then no polytope in the combinatorial class of  $P$  can be psd-minimal.*

*Proof.* Suppose  $Q$  is psd-minimal and combinatorially equivalent to  $P$ . Hence, we can assume that  $S_P(x)$  equals  $S_Q(x)$ . By Theorem 5.4 there is some  $u = (u_1, \dots, u_t) \in \mathbb{R}^t$ , with no coordinate equal to zero, such that  $S_Q = S_P(u_1^2, \dots, u_t^2)$  and  $\text{rank}(S_P(u)) = n + 1$ . Since  $S_Q$  is the slack matrix of an  $n$ -polytope, we have

$$\text{rank}(S_P(u_1^2, \dots, u_t^2)) = n + 1 = \text{rank}(S_P(u_1, \dots, u_t)).$$

Now suppose  $D(x)$  is a trinomial  $(n+2)$ -minor of  $S_P(x)$ . Up to sign,  $D(x)$  has the form  $x^a + x^b + x^c$  or  $x^a - x^b + x^c$  for some  $a, b, c \in \mathbb{N}^t$ . In either case, it is not possible for  $D(u_1^2, \dots, u_t^2) = D(u_1, \dots, u_t) = 0$ .  $\square$

### 5.1 Psd-minimal polytopes of dimension up to four.

**Proposition 5.7.** *Gouveia, Robinson, and Thomas [2013, Theorem 4.7] The psd-minimal polygons are precisely all triangles and quadrilaterals.*

*Proof.* Let  $P$  be an  $n$ -gon where  $n > 4$ . Then  $S_P(x)$  has a submatrix of the form

$$\begin{bmatrix} 0 & x_1 & x_2 & x_3 \\ 0 & 0 & x_4 & x_5 \\ x_6 & 0 & 0 & x_7 \\ x_8 & x_9 & 0 & 0 \end{bmatrix},$$

whose determinant is  $x_1x_4x_7x_8 - x_2x_5x_6x_9 + x_3x_4x_6x_9$  up to sign. By Lemma 5.6, no  $n$ -gon with  $n > 4$  can be psd-minimal.

Since all triangles are projectively equivalent, by verifying the psd-minimality of one, they are all seen to be psd-minimal. Similarly, for quadrilaterals.  $\square$

Lemma 5.6 can also be used to classify up to combinatorial equivalence all 3-polytopes that are psd-minimal. Using Proposition 5.7, together with the fact that faces of psd-minimal polytopes are also psd-minimal, and the invariance of psd rank under polarity, we get that any 3-polytope  $P$  with a vertex of degree larger than four, or a facet that is an  $n$ -gon where  $n > 4$ , cannot be psd-minimal.

**Lemma 5.8.** *If  $P$  is a 3-polytope with a vertex of degree four and a quadrilateral facet incident to this vertex, then  $S_P(x)$  contains a trinomial 5-minor.*

*Proof.* Let  $v$  be the vertex of degree four incident to facets  $F_1, F_2, F_3, F_4$  such that  $[v_1, v] = F_1 \cap F_2$ ,  $[v_2, v] = F_2 \cap F_3$ ,  $[v_3, v] = F_3 \cap F_4$  and  $F_4 \cap F_1$  are edges of  $P$ , where  $v_1, v_2$  and  $v_3$  are vertices of  $P$ .

Suppose  $F_4$  is quadrilateral. Then  $F_4$  has a vertex  $v_4$  that is different from, and non-adjacent to,  $v$ . Therefore,  $v_4$  does not lie on  $F_1, F_2$  or  $F_3$ . Consider the  $5 \times 5$  submatrix of  $S_P(x)$  with rows indexed by  $v, v_1, v_2, v_3, v_4$  and columns by  $F_1, F_2, F_3, F_4, F$  where  $F$  is a facet not containing  $v$ . This matrix has the form

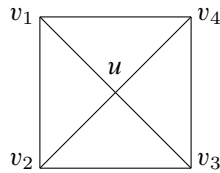
$$\begin{bmatrix} 0 & 0 & 0 & 0 & x_1 \\ 0 & 0 & x_2 & x_3 & * \\ x_4 & 0 & 0 & x_5 & * \\ x_6 & x_7 & 0 & 0 & * \\ x_8 & x_9 & x_{10} & 0 & * \end{bmatrix},$$

and its determinant is a trinomial. □

**Proposition 5.9.** *The psd-minimal 3-polytopes are combinatorially equivalent to simplices, quadrilateral pyramids, bisimplices, octahedra or their duals.*

*Proof.* Suppose  $P$  is a psd-minimal 3-polytope. If  $P$  contains only vertices of degree three and triangular facets, then  $P$  is a simplex.

For all remaining cases,  $P$  must have a vertex of degree four or a quadrilateral facet. Since psd rank is preserved under polarity, we may assume that  $P$  has a vertex  $u$  of degree four. By Lemma 5.8, the neighborhood of  $u$  looks as follows.



Suppose  $P$  has five vertices. If all edges of  $P$  are in the picture, i.e. the picture is a Schlegel diagram of  $P$ , then  $P$  is a quadrilateral pyramid. Otherwise  $P$  has one more edge, and this edge is  $[v_1, v_3]$  or  $[v_2, v_4]$ , yielding a bisimplex in either case.

If  $P$  has more than five vertices, then we may assume that  $P$  has a vertex  $v$  that is a neighbor of  $v_1$  different from  $u, v_2, v_4$ . Then  $v_1$  is a degree four vertex and thus, by Lemma 5.8, all facets of  $P$  containing  $v_1$  are triangles. This implies that  $v$  is a neighbor of  $v_2$  and  $v_4$ . Applying the same logic to either  $v_2$  or  $v_4$ , we get that  $v$  is also a neighbor of  $v_3$ . Since all these vertices now have degree four, there could be no further vertices in  $P$ , and so  $P$  is an octahedron. Hence  $P$  is combinatorially equal to, or dual to, one of the polytopes seen so far. □

Call an octahedron in  $\mathbb{R}^3$ , *biplanar*, if there are two distinct planes each containing four vertices of the octahedron. The complete classification of psd-minimal 3-polytopes is as follows.

**Theorem 5.10.** *Gouveia, Robinson, and Thomas [2013, Theorem 4.11] The psd-minimal 3-polytopes are precisely simplices, quadrilateral pyramids, bisimplices, biplanar octahedra and their polars.*

In dimension four, the classification of psd-minimal polytopes becomes quite complicated. The full list consists of 31 combinatorial classes of polytopes including the 11 known projectively unique polytopes in  $\mathbb{R}^4$ . These 11 are *combinatorially psd-minimal*, meaning that all polytopes in each of their combinatorial classes are psd-minimal. For

the remaining 20 classes, there are non-trivial conditions on psd-minimality. We refer the reader to [Gouveia, Pashkovich, Robinson, and Thomas \[2017\]](#) for the result in  $\mathbb{R}^4$ .

Beyond  $\mathbb{R}^4$ , a classification of all psd-minimal polytopes looks to be cumbersome. On the other hand, there are families of polytopes of increasing dimension that are all psd-minimal. A polytope  $P \subset \mathbb{R}^n$  is *2-level* if for every facet of  $P$ , all vertices of  $P$  are either on this facet or on a single other parallel translate of the affine span of this facet. Examples of 2-level polytopes include simplices, regular hypercubes, regular cross-polytopes, and hypersimplices. All 2-level polytopes are psd-minimal, but not conversely. For example, the regular bisimplex in  $\mathbb{R}^3$  is psd-minimal but not 2-level. Recall from [Example 1.4](#) that the stable set polytopes of perfect graphs are psd-minimal. In fact, they are also 2-level and it was shown in [Gouveia, Parrilo, and Thomas \[2010, Corollary 4.11\]](#) that all down-closed 0/1-polytopes that are 2-level are in fact stable set polytopes of perfect graphs. On the other hand, [Gouveia, Parrilo, and Thomas \[2013, Theorem 9\]](#) shows that  $\text{STAB}(G)$  is not psd-minimal if  $G$  is not perfect.

## 6 Spectrahedral lifts and sum of squares polynomials

We now look at a systematic technique that creates a sequence of nested outer approximations of the convex hull of an algebraic set. These approximations come from projections of spectrahedra and are called *theta bodies*. In many cases, the theta body at the  $k$ th step will equal the closure of the convex hull of the algebraic set and hence the spectrahedron that it was a projection of, is a lift of this convex set. We examine how this type of lift fits into our general picture.

Let  $I = \langle p_1, \dots, p_s \rangle \subset \mathbb{R}[x] := \mathbb{R}[x_1, \dots, x_n]$  be a polynomial ideal and let  $\mathcal{V}_{\mathbb{R}}(I) \subset \mathbb{R}^n$  be the real points in its variety. Then the closure of the convex hull of  $\mathcal{V}_{\mathbb{R}}(I)$ ,  $C := \overline{\text{conv}(\mathcal{V}_{\mathbb{R}}(I))}$ , is a closed convex semialgebraic set. Since we are only interested in the convex hull of  $\mathcal{V}_{\mathbb{R}}(I)$ , and the convex hull is defined by its extreme points, we may assume without loss of generality that  $I$  is the largest ideal that vanishes on the extreme points of  $C$ .

Recall that  $C$  is the intersection of all half spaces containing  $\mathcal{V}_{\mathbb{R}}(I)$ . Each half space is expressed as  $l(x) \geq 0$  for some linear polynomial  $l \in \mathbb{R}[x]$  that is nonnegative on  $\mathcal{V}_{\mathbb{R}}(I)$ . A linear polynomial  $l$  is nonnegative on  $\mathcal{V}_{\mathbb{R}}(I)$  if there exists polynomials  $h_i \in \mathbb{R}[x]$  such that  $l - \sum h_i^2 \in I$ . In this case we say that  $l$  is a *sum of squares (sos)* mod  $I$ , and if the degree of each  $h_i$  is at most  $k$ , then we say that  $l$  is *k-sos mod I*. Define the  $k$ th theta body of  $I$  to be the set

$$\text{TH}_k(I) := \{x \in \mathbb{R}^n : l(x) \geq 0 \ \forall \ l \text{ linear and } k\text{-sos mod } I\}.$$



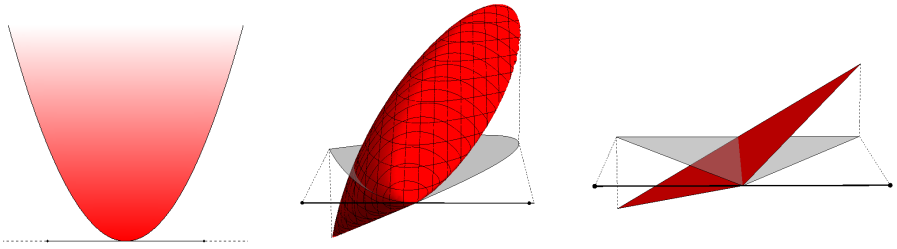


Figure 4: The theta bodies of  $I = \langle (x+1)x(x-1)^2 \rangle$  and their spectrahedral lifts. The first theta body is the entire real line, the second is slightly larger than  $[-1, 1]$  and the third is exactly  $[-1, 1]$ .

Note that all theta bodies are closed convex semialgebraic sets and they form a series of nested outer approximations of  $C$  since

$$\text{TH}_i(I) \supseteq \text{TH}_{i+1}(I) \supseteq C \quad \text{for all } i \geq 1.$$

We say that  $I$  is  $\text{TH}_k$ -exact if  $\text{TH}_k(I) = C$ . The terminology is inspired by Lovász's theta body  $\text{TH}(G)$  from [Example 1.4](#) which is precisely  $\text{TH}_1(I_G)$  of the ideal

$$I_G = \langle x_i^2 - x_i, \forall i = 1, \dots, n \rangle + \langle x_i x_j, \forall \{i, j\} \in E(G) \rangle.$$

In our terminology,  $I_G$  is  $\text{TH}_1$ -exact when  $G$  is a perfect graph.

Theta bodies of a general polynomial ideal  $I \subset \mathbb{R}[x]$  were defined in [Gouveia, Parrilo, and Thomas \[2010\]](#), and it was shown there that  $I$  is  $\text{TH}_k$ -exact if and only if  $C$  admits a specific type of spectrahedral lift. This lift has size equal to the number of monomials in  $\mathbb{R}[x]$  of degree at most  $k$ . Let  $[x]_k$  denote the vector of all monomials of degree at most  $k$  in  $\mathbb{R}[x]$ . When  $\text{TH}_k(I) = C$ , [Theorem 2.3](#) promises two maps  $A$  and  $B$  that factorize the slack operator of  $C$ . These operators are very special.

**Theorem 6.1.** *Gouveia, Parrilo, and Thomas [2013, Theorem 11] The slack operator of  $C = \text{conv}(\mathbb{U}_{\mathbb{R}}(I))$  has a factorization in which  $A(x) = [x]_k [x]_k^\top$  if and only if  $C = \text{TH}_k(I)$ . Further, the map  $B$  sends each linear functional  $l(x)$  corresponding to an extreme point of the polar of  $C$  to a psd matrix  $Q_l$  such that  $l(x) - x^\top Q_l x \in I$  certifying that  $l(x)$  is nonnegative on  $\mathbb{U}_{\mathbb{R}}(I)$ .*

In fact, each theta body is the projection of a spectrahedron. [Figure 4](#) shows the theta bodies and their spectrahedral lifts of the ideal  $I = \langle (x+1)x(x-1)^2 \rangle$ . In this case,  $C = [-1, 1] \subset \mathbb{R}$ .

While theta bodies offer a systematic method to sometimes construct a spectrahedral lift of  $C$ , they may not offer the most efficient lift of this set. So an immediate question is whether there might be radically different types of spectrahedral lifts for  $C$ . Since the projection of a spectrahedron is necessarily convex and semialgebraic, a set  $C$  can have a spectrahedral lift only if it is convex and semialgebraic. So a second question is whether every convex semialgebraic set has a spectrahedral lift. This question gained prominence from Nemirovski [2007], and Helton and Nie showed that indeed a compact convex semialgebraic set has a spectrahedral lift if its boundary is sufficiently smooth and has positive curvature. They then conjectured that every convex semialgebraic set has a spectrahedral lift, see Helton and Nie [2009] and Helton and Nie [2010]. This conjecture was very recently disproved by Scheiderer who exhibited many explicit counter-examples Scheiderer [2018b]. All these sets therefore have infinite psd rank.

Recall that a morphism  $\phi : X \rightarrow Y$  between two affine real varieties creates a ring homomorphism  $\phi^* : \mathbb{R}[Y] \rightarrow \mathbb{R}[X]$  between their coordinate rings. By a real variety we mean a variety defined by polynomials with real coefficients. Let  $X_{\mathbb{R}}$  denote the  $\mathbb{R}$ -points of  $X$ .

**Theorem 6.2.** *Scheiderer [ibid., Theorem 3.14] Let  $S \subset \mathbb{R}^n$  be a semialgebraic set and let  $C$  be the closure of its convex hull. Then  $C$  has a spectrahedral lift if and only if there is a morphism  $\phi : X \rightarrow \mathbb{A}^n$  of affine real varieties and a finite-dimensional  $\mathbb{R}$ -linear subspace  $U$  in the coordinate ring  $\mathbb{R}[X]$  such that*

1.  $S \subset \phi(X_{\mathbb{R}})$ ,
2. *for every linear polynomial  $l \in \mathbb{R}[x]$  that is nonnegative on  $S$ , the element  $\phi^*(l)$  of  $\mathbb{R}[X]$  is a sum of squares of elements in  $U$ .*

This theorem offers a set of necessary and sufficient conditions for the existence of a spectrahedral lift of the convex hull of a semialgebraic set by working through an intermediate variety  $X$ . The setting is more general than that in Theorem 6.1 where we only considered convex hulls of algebraic sets. Regardless, the spirit of condition (2) is that the theta body method (or more generally, Lasserre's method Lasserre [2000/01]) is essentially universal with the subspace  $U \subseteq \mathbb{R}[X]$  playing the role of degree bounds on the sos nonnegativity certificates that were required for  $\text{TH}_k$ -exactness. Theorem 6.2 provides counterexamples to the Helton-Nie conjecture.

**Theorem 6.3.** *Scheiderer [2018b, Theorem 4.23] Let  $S \subset \mathbb{R}^n$  be any semialgebraic set with  $\dim(S) \geq 2$ . Then for some positive integer  $k$ , there exists a polynomial map  $\phi : S \rightarrow \mathbb{R}^k$  such that the closed convex hull of  $\phi(S) \subset \mathbb{R}^k$  is not the linear image of a spectrahedron.*

These results show, among other examples, that there are high enough Veronese embeddings of semialgebraic sets that cannot be the projections of spectrahedra.

**Corollary 6.4.** *Scheiderer [2018b, Corollary 4.24] Let  $n, d$  be positive integers with  $n \geq 3, d \geq 4$  or  $n = 2$  and  $d \geq 6$ . Let  $m_1, \dots, m_N$  be the non-constant monomials in  $\mathbb{R}[x]$  of degree at most  $d$ . Then for any semialgebraic set  $S \subseteq \mathbb{R}^n$  with non-empty interior, the closed convex hull of*

$$m(S) := \{(m_1(s), \dots, m_N(s)) : s \in S\} \subset \mathbb{R}^N$$

*is not the linear image of a spectrahedron.*

In contrast, Scheiderer had previously shown that all convex semialgebraic sets in  $\mathbb{R}^2$  have spectrahedral lifts [Scheiderer \[2018a\]](#), thus proving the Helton-Nie conjecture in the plane. The current smallest counterexamples to the Helton-Nie conjecture are in  $\mathbb{R}^{11}$ . Is it possible that there is a counterexample in  $\mathbb{R}^3$ ?

## 7 Notes

There are many further results on spectrahedral lifts of convex sets beyond those mentioned here. An important topic that has been left out is that of symmetric spectrahedral lifts which are lifts that respect the symmetries of the convex set. Due to the symmetry requirement, such lifts are necessarily of size at least as large as the psd rank of the convex set. On the other hand, the symmetry restriction provides more tools to study such lifts and there are many beautiful results in this area, see [Fawzi, Saunderson, and Parrilo \[2017\]](#), [Fawzi, Saunderson, and Parrilo \[2015\]](#), [Fawzi, Saunderson, and Parrilo \[2016\]](#).

Many specific examples of spectrahedral lifts of convex sets exist, and several of them have significance in applications. An easy general source is the book [Blekherman, Parrilo, and Thomas \[2013\]](#). In particular, Chapter 6 is dedicated to sdp representability of convex sets. This book includes a number of further topics in the area of *Convex Algebraic Geometry*.

**Acknowledgments.** I am indebted to all my collaborators on the projects that contributed to this paper. I thank Pablo Parrilo for the construction in [Example 1.5](#) and for several useful conversations. I also thank Hamza Fawzi and João Gouveia for comments on this paper, and Hamza for pointing out the bound in [Proposition 4.8](#). Claus Scheiderer and Daniel Plaumann were very helpful with the content of [Section 6](#).

## References

- A. Ben-Tal and A. Nemirovski (2001). *Lectures on Modern Convex Optimization*. MPS SIAM Series on Optimization. Analysis, algorithms, and engineering applications. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA; Mathematical Programming Society (MPS), Philadelphia, PA (cit. on pp. [3837](#), [3843](#)).
- G Blekherman, P. A. Parrilo, and R. R. Thomas, eds. (2013). *Semidefinite Optimization and Convex Algebraic Geometry*. Vol. 13. MOS-SIAM Ser. Optim. SIAM, Philadelphia, PA (cit. on pp. [3843](#), [3860](#)).
- J. Briët, D. Dadush, and S. Pokutta (2015). “On the existence of 0/1 polytopes with high semidefinite extension complexity”. *Math. Program.* 153.1, Ser. B, pp. 179–199 (cit. on p. [3851](#)).
- M. Conforti, G. Cornuéjols, and G. Zambelli (2010). “Extended formulations in combinatorial optimization”. *4OR* 8.1, pp. 1–48 (cit. on p. [3837](#)).
- H. Fawzi, J. Gouveia, P. A. Parrilo, R. Z. Robinson, and R. R. Thomas (2015). “Positive semidefinite rank”. *Math. Program.* 153.1, Ser. B, pp. 133–177 (cit. on p. [3849](#)).
- H. Fawzi and M. Safey El Din (2018). “[A lower bound on the positive semidefinite rank of convex bodies](#)”. To appear in SIAM J. Applied Algebra and Geometry (cit. on p. [3851](#)).
- H. Fawzi, J. Saunderson, and P. A. Parrilo (2015). “Equivariant semidefinite lifts and sum-of-squares hierarchies”. *SIAM J. Optim.* 25.4, pp. 2212–2243 (cit. on p. [3860](#)).
- (2016). “Sparse sums of squares on finite abelian groups and improved semidefinite lifts”. *Math. Program.* 160.1-2, Ser. A, pp. 149–191 (cit. on p. [3860](#)).
- (2017). “Equivariant semidefinite lifts of regular polygons”. *Math. Oper. Res.* 42.2, pp. 472–494 (cit. on p. [3860](#)).
- S. Fiorini, S. Massar, S. Pokutta, H. R. Tiwary, and R. de Wolf (2012). “Linear vs. semidefinite extended formulations: exponential separation and strong lower bounds”. In: *STOC’12—Proceedings of the 2012 ACM Symposium on Theory of Computing*. ACM, New York, pp. 95–106 (cit. on p. [3845](#)).
- M. X. Goemans (2015). “Smallest compact formulation for the permutahedron”. *Math. Program.* 153.1, Ser. B, pp. 5–11 (cit. on pp. [3839](#), [3852](#)).
- J. Gouveia, P. A. Parrilo, and R. R. Thomas (2010). “Theta bodies for polynomial ideals”. *SIAM J. Optim.* 20.4, pp. 2097–2118 (cit. on pp. [3857](#), [3858](#)).
- (2013). “Lifts of convex sets and cone factorizations”. *Math. Oper. Res.* 38.2, pp. 248–264 (cit. on pp. [3843](#)–[3845](#), [3852](#), [3857](#), [3858](#)).
- J. Gouveia, K. Pashkovich, R. Z. Robinson, and R. R. Thomas (2017). “Four-dimensional polytopes of minimum positive semidefinite rank”. *J. Combin. Theory Ser. A* 145, pp. 184–226 (cit. on p. [3857](#)).

- J. Gouveia, R. Z. Robinson, and R. R. Thomas (2013). “Polytopes of minimum positive semidefinite rank”. *Discrete Comput. Geom.* 50.3, pp. 679–699 (cit. on pp. [3850](#), [3853](#), [3855](#), [3856](#)).
- J. W. Helton and J. Nie (2009). “Sufficient and necessary conditions for semidefinite representability of convex hulls and sets”. *SIAM J. Optim.* 20.2, pp. 759–791 (cit. on p. [3859](#)).
- (2010). “Semidefinite representation of convex sets”. *Math. Program.* 122.1, Ser. A, pp. 21–64 (cit. on p. [3859](#)).
- J. B. Lasserre (2000/01). “Global optimization with polynomials and the problem of moments”. *SIAM J. Optim.* 11.3, pp. 796–817 (cit. on p. [3859](#)).
- J. R. Lee, P. Raghavendra, and D. Steurer (2015). “Lower bounds on the size of semidefinite programming relaxations”. In: *STOC’15—Proceedings of the 2015 ACM Symposium on Theory of Computing*. ACM, New York, pp. 567–576 (cit. on p. [3851](#)).
- L. Lovász (1979). “On the Shannon capacity of a graph”. *IEEE Trans. Inform. Theory* 25.1, pp. 1–7 (cit. on p. [3840](#)).
- A. Nemirovski (2007). “Advances in convex optimization: conic programming”. In: *International Congress of Mathematicians. Vol. I*. Eur. Math. Soc., Zürich, pp. 413–444 (cit. on p. [3859](#)).
- C. Scheiderer (2018a). “Semidefinite representation for convex hulls of real algebraic curves”. *SIAM J. Appl. Algebra Geom.* 2.1, pp. 1–25 (cit. on p. [3860](#)).
- (2018b). “Spectrahedral shadows”. *SIAM J. Appl. Algebra Geom.* 2.1, pp. 26–44 (cit. on pp. [3859](#), [3860](#)).
- M. Yannakakis (1991). “Expressing combinatorial optimization problems by linear programs”. *J. Comput. System Sci.* 43.3, pp. 441–466 (cit. on pp. [3841](#), [3844](#), [3846](#)).
- G. M. Ziegler (1995). *Lectures on Polytopes*. Vol. 152. Graduate Texts in Mathematics. New York: Springer-Verlag (cit. on pp. [3838](#), [3839](#)).

Received 2017-12-02.

REKHA R. THOMAS  
DEPARTMENT OF MATHEMATICS  
UNIVERSITY OF WASHINGTON  
BOX 354350  
SEATTLE, WA 98195 USA  
[rrthomas@uw.edu](mailto:rrthomas@uw.edu)

## OPTIMAL SHAPE AND LOCATION OF SENSORS OR ACTUATORS IN PDE MODELS

EMMANUEL TRÉLAT

### Abstract

We report on a series of works done in collaboration with Y. Privat and E. Zuazua, concerning the problem of optimizing the shape and location of sensors and actuators for systems whose evolution is driven by a linear partial differential equation. This problem is frequently encountered in applications where one wants to optimally design sensors in order to maximize the quality of the reconstruction of solutions by using only partial observations, or to optimally design actuators in order to control a given process with minimal efforts. For example, we model and solve the following informal question: what is the optimal shape and location of a thermometer?

Note that we want to optimize not only the placement but also the shape of the observation or control subdomain over the class of all possible measurable subsets of the domain having a prescribed Lebesgue measure. By probabilistic considerations we model this optimal design problem as the one of maximizing a spectral functional interpreted as a randomized observability constant, which models optimal observability for random initial data.

Solving this problem strongly depends on the operator in the PDE model and requires fine knowledge on the asymptotic properties of eigenfunctions of that operator. For parabolic equations like heat, Stokes or anomalous diffusion equations, we prove the existence and uniqueness of a best domain, proved to be regular enough, and whose algorithmic construction depends in general on a finite number of modes. In contrast, for wave or Schrödinger equations, relaxation may occur and our analysis reveals intimate relations with quantum chaos, more precisely with quantum ergodicity properties of the Laplacian eigenfunctions.

---

*MSC2010:* primary 93B07; secondary 49K20, 49Q10, 35P20, 58J51.

*Keywords:* observability, controllability, partial differential equations, shape optimization, spectral inequalities, quantum ergodicity.

## 1 Introduction and modeling

Our objective is to address the problem of optimizing the shape and location of sensors and actuators for processes modeled by a linear partial differential equation. Such questions are frequently encountered in engineering applications in which one aims at placing optimally, for instance, some given sensors on a system in order to achieve then the best possible reconstruction from observed signals. Here we also want to optimize the shape of sensors, without prescribing any a priori restriction on their regularity. Such problems have been little treated from the mathematical point of view. Our aim is to provide a relevant and rigorous mathematical model and setting in which the question can be addressed. Since controllability and observability are dual notions, we essentially focus on observability. The equations that we will investigate are mainly the wave equation

$$(1) \quad \partial_{tt} y = \Delta y$$

or the Schrödinger equation

$$(2) \quad \partial_t y = i \Delta y$$

or general parabolic equations

$$(3) \quad \partial_t y = Ay$$

like heat-like, Stokes and anomalous diffusion equations for instance, settled on some open bounded connected subset  $\Omega$  of a Riemannian manifold, with various possible boundary conditions that can be Dirichlet, Neumann, mixed or Robin.

**1.1 Spectral optimal design formulation.** The first question arising is the one of formulating the problem in a relevant way. There are indeed several possible approaches to model the optimal observation problem; in particular we have to make precise the meaning of optimality here.

**Informal considerations.** To begin with, let us focus on the wave equation (1) with Dirichlet conditions on  $\partial\Omega$  (considerations for (2) and (3) are similar). The domain  $\Omega$  may represent a cavity in which some signals are propagating, in which we want to design and place some sensors that will then perform some measurements over a certain horizon of time, in view of a reconstruction inverse problem aiming at getting full information on the wave signals from the knowledge of these partial measurements.

We want to settle a relevant and appropriate mathematical formulation of the question of knowing what is the best possible shape and location of sensors, achieving the “best possible” observation in some sense.

A first obvious but important remark is that, in the absence of any constraint, certainly the best strategy consists of observing the solutions over the whole domain  $\Omega$ , that is, place sensors everywhere. This is however clearly not reasonable and in practice the subdomain covered by sensors is limited, due for instance to cost considerations. From the mathematical point of view, this constraint is taken into account by considering as the set of unknowns, the set of all possible measurable subsets  $\omega$  of  $\Omega$  that are of Lebesgue measure  $|\omega| = L|\Omega|$ , where  $L \in (0, 1)$  is some fixed real number.

Given such a subset  $\omega$  representing the sensors (and that we will try to optimize), we observe the restriction  $y|_{\omega}$  of solutions of (1) over a certain time interval  $[0, T]$  for some fixed  $T > 0$ , while wanting that these observations be enough to be indeed able to reconstruct the whole solutions in the most efficient way. This injectivity property is usually called *observability*.

**Observability inequality.** We recall that the wave equation (1) is observable on  $\omega$  in time  $T$  if there exists  $C > 0$  such that

$$(4) \quad C \|(y(0, \cdot), \partial_t y(0, \cdot))\|_{L^2(\Omega) \times H^{-1}(\Omega)}^2 \leq \int_0^T \int_{\omega} |y(t, x)|^2 dx dt,$$

for all solutions  $y$  of (1). This is called an *observability inequality*.

It is well known that, for  $\omega$  open and  $\partial\Omega$  smooth, observability holds if the pair  $(\omega, T)$  satisfies the *Geometric Control Condition* (GCC) in  $\Omega$  (see [Bardos, Lebeau, and Rauch \[1992\]](#)), according to which every geodesic ray that propagates in  $\Omega$  at unit speed and reflects on its boundary according to the laws of geometric optics (like in a billiard) should intersect  $\omega$  within time  $T$  (note that this result has been extended to the case of time-varying domains  $\omega(t)$  in [Le Rousseau, Lebeau, Terpolilli, and Trélat \[2017\]](#)). On [Figure 1](#), on the

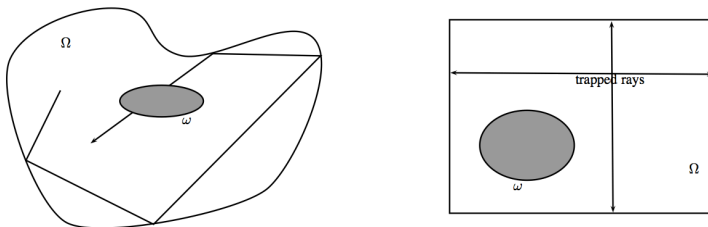


Figure 1: Illustration of the Geometric Control Condition

right, GCC is not satisfied because of the existence of trapped rays: there are solutions of



the wave equation that can never be observed on  $\omega$ . Note that GCC is necessary if there is no geodesic ray grazing  $\omega$  (see [Humbert, Privat, and Trélat \[2016\]](#)).

The *observability constant*  $C_T(\chi_\omega)$ , defined as

$$C_T(\chi_\omega) = \inf \left\{ \int_0^T \int_\Omega \chi_\omega(x) |y(t, x)|^2 dx dt \mid y \text{ is solution of (1),} \right. \\ \left. \|(y(0, \cdot), \partial_t y(0, \cdot))\|_{L^2(\Omega) \times H^{-1}(\Omega)} = 1 \right\},$$

is the largest nonnegative constant  $C_T(\chi_\omega)$  such that (4) holds true. Here, the notation  $\chi_\omega$  stands for the characteristic function of  $\omega$ . We have observability if  $C_T(\chi_\omega) > 0$ . The observability constant is defined in a similar way for the Schrödinger equation (2) and for the general parabolic equation (3) (see [Privat, Trélat, and Zuazua \[2015b, 2016a\]](#)). The constant  $C_T(\chi_\omega)$  measures the well-posedness of the inverse problem of reconstructing the whole solutions of (1) from partial measurements on  $[0, T] \times \omega$ .

At first sight it seems therefore relevant to model the problem of maximizing observability as the optimal design problem

$$(5) \quad \sup_{\chi_\omega \in \mathcal{U}_L} C_T(\chi_\omega)$$

where

$$\mathcal{U}_L = \{\chi_\omega \mid \omega \subset \Omega \text{ measurable, } |\omega| = L|\Omega|\}.$$

We stress that, in this problem, we want to optimize not only the placement but also the shape of  $\omega$  over all possible measurable subsets of  $\Omega$  having a prescribed measure. We do not put any restriction on the a priori regularity of  $\omega$ : the search is over subsets that do not have a prescribed shape, that are not necessarily BV, etc. This lack of compactness shall naturally raise important mathematical difficulties.

Anyway, modeling optimal observability as the problem (5) of maximizing the *deterministic* observability constant leads to a mathematical problem that is difficult to handle from the theoretical point of view, and more importantly, that is not fully relevant in view of practical issues. Let us explain these two difficulties and let us then explain how to adopt a slightly different model.

The first difficulty is due to the emergence of *crossed terms* in the spectral expansion of solutions. More precisely, let us fix in what follows a Hilbert basis  $(\phi_j)_{j \in \mathbb{N}^*}$  of  $L^2(\Omega)$  consisting of eigenfunctions of the Dirichlet-Laplacian operator  $-\Delta$  on  $\Omega$ , associated with the positive eigenvalues  $(\lambda_j)_{j \in \mathbb{N}^*}$  with  $\lambda_1 \leq \dots \leq \lambda_j \rightarrow +\infty$ . Since any solution  $y$  of (1) can be expanded as

$$y(t, x) = \sum_{j=1}^{+\infty} \left( a_j e^{i\sqrt{\lambda_j}t} + b_j e^{-i\sqrt{\lambda_j}t} \right) \phi_j(x)$$

where the coefficients  $a_j$  and  $b_j$  account for initial data, it follows that

$$C_T(\chi_\omega) = \frac{1}{2} \inf_{\substack{(a_j), (b_j) \in \ell^2(\mathbb{C}) \\ \sum_{j=1}^{+\infty} (|a_j|^2 + |b_j|^2) = 1}} \int_0^T \int_\omega \left| \sum_{j=1}^{+\infty} (a_j e^{i\sqrt{\lambda_j}t} + b_j e^{-i\sqrt{\lambda_j}t}) \phi_j(x) \right|^2 dx dt,$$

and then maximizing this functional over  $\mathcal{U}_L$  appears to be very difficult from the theoretical point of view, due to the crossed terms  $\int_\omega \phi_j \phi_k dx$  measuring the interaction over  $\omega$  between distinct eigenfunctions. The difficulty is similar to the one appearing in the well known problem of determining what are the best constants in Ingham's inequalities (see Jaffard and Micu [2001], Jaffard, Tucsnak, and Zuazua [1997], and Privat, Trélat, and Zuazua [2013b]).

The second difficulty with the model (5) is its lack of practical relevance. Indeed, the observability constant  $C_T(\chi_\omega)$  is *deterministic* and provides an account for the *worst possible case*: in this sense, it is a *pessimistic* constant. In practice, we perform the optimal design of sensors a priori, once for all, in view then of realizing a large number of measures (i.e., for many initial conditions). While performing many measurements, it may be expected that the worst case does not occur so often, and one would like that the observation be optimal for *most* of experiments. This leads us to consider rather an *averaged version* of the observability inequality over random initial data.

**Randomized observability constant.** We define what we call the *randomized observability constant* by

$$C_{T,\text{rand}}(\chi_\omega) = \frac{1}{2} \inf \left\{ \mathbb{E} \int_0^T \int_\omega \left| \sum_{j=1}^{+\infty} (\beta_{1,j}^v a_j e^{i\sqrt{\lambda_j}t} + \beta_{2,j}^v b_j e^{-i\sqrt{\lambda_j}t}) \phi_j(x) \right|^2 dx dt, \right. \\ \left. (a_j), (b_j) \in \ell^2(\mathbb{C}), \sum_{j=1}^{+\infty} (|a_j|^2 + |b_j|^2) = 1 \right\}$$

where  $(\beta_{1,j}^v)_{j \in \mathbb{N}^*}$  and  $(\beta_{2,j}^v)_{j \in \mathbb{N}^*}$  are two sequences of i.i.d. random laws (for instance, Bernoulli) on a probability space  $(\mathcal{X}, \mathcal{G}, \mathbb{P})$ , and  $\mathbb{E}$  is the expectation over the  $\mathcal{X}$  with respect to the probability measure  $\mathbb{P}$ . Definitions are similar for other equations (Schrödinger, heat, Stokes, etc). The constant  $C_{T,\text{rand}}(\chi_\omega)$  corresponds to the largest nonnegative constant of an averaged version of the observability inequality over random initial data. Indeed, with respect to the previous expression, the Fourier coefficients of the initial data have been randomized. In turn, by independence and taking the expectation, all crossed terms disappear and we obtain the following explicit expression of  $C_{T,\text{rand}}(\chi_\omega)$ , for any of

the equations (1), (2) and (3). For the latter, we assume that  $(\phi_j)_{j \in \mathbb{N}^*}$  is a Hilbert basis of  $L^2(\Omega, \mathbb{C})$  consisting of (complex-valued) eigenfunctions of the operator  $-A$ , associated with the (complex) eigenvalues  $(\lambda_j)_{j \in \mathbb{N}^*}$  such that  $\operatorname{Re}(\lambda_1) \leq \dots \leq \operatorname{Re}(\lambda_j) \leq \dots$ .

**Theorem 1** (Privat, Trélat, and Zuazua [2015b, 2016a]). *For every measurable subset  $\omega$  of  $\Omega$ , we have*

$$C_{T,\text{rand}}(\chi_\omega) = T \inf_{j \in \mathbb{N}^*} \gamma_j(T) \int_\omega |\phi_j(x)|^2 dx$$

where

$$\gamma_j(T) = \begin{cases} 1/2 & \text{for the wave equation (1),} \\ 1 & \text{for the Schrödinger equation (2),} \\ \frac{e^{2\operatorname{Re}(\lambda_j)T} - 1}{2\operatorname{Re}(\lambda_j)} & \text{for the parabolic equation (3).} \end{cases}$$

Note that we always have  $C_T(\chi_\omega) \leq C_{T,\text{rand}}(\chi_\omega)$  and that the inequality is strict for instance in each of the following cases:

- 1D Dirichlet waves on  $\Omega = (0, \pi)$ , whenever  $T$  is not an integer multiple of  $\pi$  (see Privat, Trélat, and Zuazua [2013b]);
- multi-D Dirichlet waves on  $\Omega$  stadium-shaped, when  $\omega$  contains an open neighborhood of the wings (in that case,  $C_T(\chi_\omega) = 0$ ; see Privat, Trélat, and Zuazua [2016a]).

**Formulation of the optimal observability problem.** Taking into account the fact that, in practice, it is expected that a large number of measurements is to be done, rather than (5), we finally choose to model the problem of best observability as the problem of maximizing the functional  $\chi_\omega \mapsto C_{T,\text{rand}}(\chi_\omega)$  over the set  $\mathcal{U}_L$ , that is:

$$(6) \quad \sup_{\chi_\omega \in \mathcal{U}_L} \inf_{j \in \mathbb{N}^*} \gamma_j(T) \int_\omega |\phi_j(x)|^2 dx$$

This is a spectral optimal design problem.

Note that the randomized observability constant  $C_{T,\text{rand}}(\chi_\omega)$  can also be interpreted as a time-asymptotic observability constant (see Privat, Trélat, and Zuazua [ibid.]).

**Remark 1.** Note that, in (5) or in (6) we take an infimum over all (randomized) initial data. In contrast, if we fix some given initial data, maximizing the functional  $\chi_\omega \mapsto \int_0^T \int_\omega |y(t, x)|^2 dx dt$  over  $\mathcal{U}_L$  is a problem that can be easily solved thanks to a decreasing rearrangement argument (see Privat, Trélat, and Zuazua [2015a]), showing that there always exists (at least) one optimal set  $\omega^*$ . The regularity of  $\omega^*$  depends on the initial data. We can show that it may be a Cantor set of positive measure, even for smooth

data. Of course, in practice designing optimal sensors depending on initial data would make no sense and this is why we consider in our model an infimum over all (or almost all) initial data.

**Remark 2.** As already underlined, in our search of the best possible subset  $\omega$ , we do not impose any restriction to  $\omega$  but its measurability. If we restrict the search to subsets having uniformly bounded (by some  $A > 0$ ) perimeter or total variation or satisfying the  $1/A$ -cone property, or if we restrict ourselves to subsets parametrized by some compact or finite-dimensional set, then quite straightforwardly there exists (at least) one optimal set  $\omega^*$ . But then the complexity of  $\omega^*$  may then increase with  $A$  (spillover phenomenon). We will observe this phenomenon when considering, further in the paper, a truncated version of (6) with a finite number of spectral modes.

Imposing no restriction on  $\omega$  is our choice here because we want to address the mathematical question of knowing if there is a "very best" subdomain over all possible measurable subsets  $\omega$  such that  $|\omega| = L|\Omega|$ .

**1.2 Related problems and existing results.** We first mention that the optimal observability problem on which we have focused up to now is related, by duality, to the problem of determining what is the best control domain for controlling to rest, for instance, the wave equation with internal control

$$\partial_{tt}y - \Delta y = \chi_\omega u.$$

We have addressed such best actuator problems in Privat, Trélat, and Zuazua [2013a, 2016b, 2017] with a similar randomization approach.

Another closely related problem is that of finding the best possible domain to stabilize the equation

$$\partial_{tt}y - \Delta y = -k\chi_\omega \partial_t y$$

thanks to a localized damping (see Privat and Trélat [2015] for results). Best means here that one may want to design  $\omega$  (over  $\mathcal{U}_L$ ) such that exponential decrease of solutions of the above locally damped wave equation is maximal. Historically, up to our knowledge, the first papers addressing this problem were Hébrard and Henrot [2003, 2005], in which the authors studied this problem in 1D and provided complete characterizations of the optimal set whenever it exists, for the problem of determining the best possible shape and position of the damping subdomain of a given measure.

Due to their relevance in engineering applications, optimal design problems for placing sensors or actuators for processes modeled by partial differential equations have been investigated in a large number of papers. Difficulties come from the facts that solutions live in infinite-dimensional spaces and that the class of admissible designs is not closed for the

standard and natural topology. Very few works take into consideration those aspects. In most of existing contributions, numerical tools are developed to solve a simplified version of the optimal design problem where either the PDE has been replaced with a discrete approximation, or the class of optimal designs is replaced with a compact finite dimensional set – see for example [Kumar and Seinfeld \[1978\]](#), [Morris \[2011\]](#), [Uciński and Patan \[2010\]](#), [van de Wal and de Jager \[2001\]](#), and [Wouwer, Point, Porteman, and Remy \[2000\]](#) where the aim is most often to optimize the number, the place and the type of sensors in order to improve the estimation of the state of the system. Sensors often have a prescribed shape (for instance, balls with a prescribed radius) and then the problem consists of placing optimally a finite number of points (the centers of the balls) and thus is finite-dimensional. Of course, the resulting optimization problem is already challenging. Here we want to optimize also the shape of the observation set without making any a priori restrictive assumption to the class of shapes (such as bounded variation) and the search is made over all possible measurable subsets.

From the mathematical point of view, the issue of studying a relaxed version of optimal design problems for shape and position of sensors or actuators has been investigated in a series of articles. In [Bellido and Donoso \[2007\]](#) the authors investigate the problem modeled in [Sigmund and Jensen \[2003\]](#) of finding the best possible distributions of two materials (with different elastic Young modulus and different density) in a rod in order to minimize the vibration energy in the structure. The authors of [Allaire, Aubry, and Jouve \[2001\]](#) also propose a convexification formulation of eigenfrequency optimization problems applied to optimal design. In [Fahroo and Ito \[1996\]](#) are discussed several possible criteria for optimizing the damping of abstract wave equations and derive optimality conditions for a certain criterion related to a Lyapunov equation. In [Münch and Periago \[2011\]](#), the authors study a homogenized version of the optimal location of controllers for the heat equation problem for fixed initial data, noticing that such problems are often ill-posed. In [Allaire, Münch, and Periago \[2010\]](#), the authors consider a similar problem and study the asymptotic behavior as the final time  $T$  goes to infinity of the solutions of the relaxed problem; they prove that optimal designs converge to an optimal relaxed design of the corresponding two-phase optimization problem for the stationary heat equation. We also mention [Fernández-Cara and Münch \[2012\]](#) where, still for fixed initial data, numerical investigations are used to provide evidence that the optimal location of null-controllers of the heat equation problem is an ill-posed problem.

## 2 Study of the optimal design problem

To address the optimal design problem (6), we distinguish between parabolic equations (3) (like heat, Stokes or anomalous diffusion equations) on the one part and the hyperbolic equations (1) and (2) on the other.

As a first remark, since the infimum in (6) involves all spectral modes  $j \in \mathbb{N}^*$ , solving the problem will require some knowledge on the asymptotic behavior of the squares  $|\phi_j|^2$  of the eigenfunctions as  $j \rightarrow +\infty$ . Note also that, because of the weights  $\gamma_j(T)$  in (6), there is a strong difference between the parabolic case where  $\gamma_j(T)$  is exponentially increasing as  $j \rightarrow +\infty$  and the hyperbolic (wave and Schrödinger) case where  $\gamma_j(T)$  remains constant.

**2.1 The parabolic case.** For parabolic equations (3), the situation is particularly nice and we have the following result, under several quite general assumptions on the operator  $A$ , which are satisfied for heat and Stokes equations and also for anomalous diffusion equations, i.e.,  $A = -(-\Delta)^\alpha$ , with  $\alpha > 1/2$ . Note that anomalous diffusion equations provide relevant models in many problems encountered in physics (plasma with slow or fast diffusion, aperiodic crystals, spins, etc), in biomathematics, in economy or in imaging sciences.

**Theorem 2** (Privat, Trélat, and Zuazua [2015b]). *Let  $T > 0$  be arbitrary. Assume that  $\partial\Omega$  is piecewise  $C^1$ . There exists a unique<sup>1</sup> optimal observation domain  $\omega^*$  solving (6). Moreover  $\omega^*$  is open and semi-analytic; in particular, it has a finite number of connected components. Additionally, we have  $C_T(\chi_{\omega^*}) < C_{T,\text{rand}}(\chi_{\omega^*})$ .*

Note that this existence and uniqueness result holds for every fixed orthonormal basis of eigenfunctions of the operator but the optimal set depends on the specific choice of the Hilbert basis.

This result (of which one can find an even more general version in Privat, Trélat, and Zuazua [ibid.]) gives a short and satisfactory positive answer to the question of knowing if there is a “very best” observation domain among all possible measurable subsets. Moreover, we are going to see further that there even exists a nice algorithmic procedure to compute the optimal set  $\omega^*$ , which happens to be fully characterized by a finite number of modes only.

The fact that the optimal set  $\omega^*$  is semi-analytic is a strong (and desirable) regularity property. In addition to the fact that  $\omega^*$  has a finite number of connected components, this implies also that  $\omega^*$  is Jordan measurable, that is,  $|\partial\omega^*| = 0$ . This is in contrast with

---

<sup>1</sup>Here, it is understood that the optimal set  $\omega^*$  is unique within the class of all measurable subsets of  $\Omega$  quotiented by the set of all measurable subsets of  $\Omega$  of zero measure.

the already mentioned fact that, for wave-like equations, when maximizing the energy for fixed data, the optimal set may be a Cantor set of positive measure, even for smooth initial data (see [Privat, Trélat, and Zuazua \[2015a\]](#)).

Let us explain shortly why [Theorem 2](#) applies to (3) with  $A = -(-\Delta)^\alpha$  (power of the Dirichlet-Laplacian) for every  $\alpha > 1/2$ . It is instrumental in the proof to use the fine lower estimates of [Apraiz, Escauriaza, Wang, and Zhang \[2014, Theorem 5\]](#), stating that

$$\int_{\omega} |\phi_j(x)|^2 \geq C e^{-C\sqrt{\mu_j}} \quad \forall j \in \mathbb{N}^*$$

(here, the  $\mu_j$ 's are the eigenvalues of  $-\Delta$ ) where the constant  $C > 0$  is uniform with respect to  $\chi_{\omega} \in \mathcal{U}_L$ . This uniform property is remarkable and particularly useful here in our context. The requirement  $\alpha > 1/2$  comes from a balance between the above lower estimate and the exponential weight  $\gamma_j(T) \sim e^{\mu_j^\alpha T}$ , yielding in that case a favorable coercivity property, itself implying compactness features that are crucial in the proof. Another instrumental tool in the proof is then a refined minimax theorem due to [Hartung \[1982\]](#).

In the critical case  $\alpha = 1/2$ , the conclusion of [Theorem 2](#) holds true as well provided that the time  $T$  is moreover large enough.

Furthermore, still considering  $A = -(-\Delta)^\alpha$ , it is proved in [Privat, Trélat, and Zuazua \[2016a\]](#) that:

- in the Euclidean square  $\Omega = (0, \pi)^2$ , when considering the usual Hilbert basis of eigenfunctions consisting of products of sine functions, for every  $\alpha > 0$  there exists a unique optimal set in  $\mathcal{U}_L$  (as in the theorem), which is moreover open and semi-analytic (whatever the value of  $\alpha > 0$  may be);
- in the Euclidean disk  $\Omega = \{x \in \mathbb{R}^2 \mid \|x\| < 1\}$ , when considering the usual Hilbert basis of eigenfunctions parametrized in terms of Bessel functions, for every  $\alpha > 0$  there exists a unique optimal set  $\omega^*$  (as in the theorem), which is moreover open, radial, with the following additional property:
  - if  $\alpha > 1/2$  then  $\omega^*$  consists of a finite number of concentric rings that are at a positive distance from the boundary (see [Figure 2](#));
  - if  $\alpha < 1/2$  (or if  $\alpha = 1/2$  and  $T$  is small enough) then  $\omega^*$  consists of an infinite number of concentric rings accumulating at the boundary.

This quite surprising result shows that the complexity of the optimal shape does not only depend on the operator but also depends on the geometry of the domain  $\Omega$ . The proof of these properties is difficult in the case  $\alpha < 1/2$ ; it requires involved estimates for Bessel functions combined with the use of quantum limits in the disk (like in the hyperbolic case in the next section) and analyticity considerations.

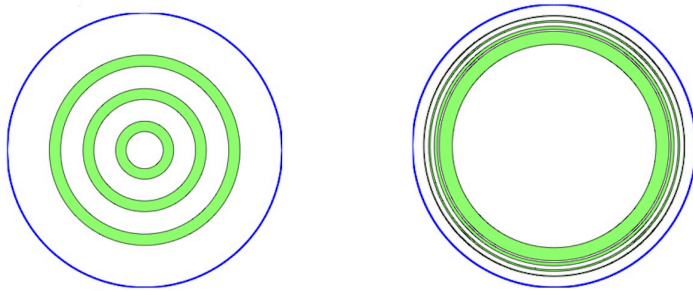


Figure 2: Optimal domain in the disk for  $L = 0.2$ ,  $T = 0.05$ . On the left:  $\alpha = 1$ . On the right:  $\alpha = 0.15$ .

**2.2 The hyperbolic case.** For the wave equation (1) and the Schrödinger equation (2), since all weights  $\gamma_j(T)$  in (6) are equal (and, in turn, the time  $T$  thus plays no role), in particular highfrequencies play an important role. Setting  $\mu_j = |\phi_j|^2 dx$ , we see that getting knowledge on the asymptotic behavior of  $\mu_j$  as  $j \rightarrow +\infty$  is now required. Noting that  $\mu_j$  is a probability measure for every  $j \in \mathbb{N}^*$ , the weak limits of the sequence  $(\mu_j)_{j \in \mathbb{N}^*}$  now enter into consideration.

**Theorem 3** (Privat, Trélat, and Zuazua [ibid.]). *Assume that the sequence of probability measures  $\mu_j = |\phi_j|^2(x) dx$  converges weakly to the uniform measure  $\frac{1}{|\Omega|} dx$  (assumption called Quantum Unique Ergodicity on the base) and that there exists  $p \in (2, +\infty]$  such that the sequence of eigenfunctions  $(\phi_j)_{j \in \mathbb{N}^*}$  is uniformly bounded in  $L^p(\Omega)$ . Then*

$$(7) \quad \sup_{\chi_\omega \in \mathcal{U}_L} \inf_{j \in \mathbb{N}^*} \int_\omega |\phi_j(x)|^2 dx = L \quad \forall L \in (0, 1).$$

To prove this result, we define  $J(\chi_\omega) = \inf_{j \in \mathbb{N}^*} \int_\omega |\phi_j(x)|^2 dx$  and we introduce a convexified version of the optimal design problem (5) (“relaxation” procedure in shape optimization), by considering the convex closure of the set  $\mathcal{U}_L$  for the  $L^\infty$  weak star topology, that is  $\overline{\mathcal{U}}_L = \{a \in L^\infty(\Omega, [0, 1]) \mid \int_\Omega a(x) dx = L|\Omega|\}$ . The convexified problem then consists of maximizing the functional  $a \mapsto J(a) = \inf_{j \in \mathbb{N}^*} \int_\Omega a(x) \phi_j(x)^2 dx$  over  $\overline{\mathcal{U}}_L$ . Clearly, a maximizer does exist, and it is easily seen by using Cesàro means of squares of eigenfunctions that the constant function  $a(\cdot) = L$  is a maximizer. But since the functional  $J$  is not lower semi-continuous it is not clear whether or not there may be a gap between the problem (5) and its convexified version. Theorem 3 above shows that, under appropriate spectral assumptions, there is no gap. The proof consists of a kind of homogenization procedure which consists of building a maximizing sequence of subsets for the problem of maximizing  $J$ , showing that it is always possible to increase the values



of  $J$  by considering subsets of measure  $L|\Omega|$  having an increasing number of connected components. The construction strongly uses the assumption that the sequence  $(\mu_j)_{j \in \mathbb{N}^*}$  has a unique weak limit (QUE on the base), which is very strong as we explain below.

**Link with quantum chaos.** Let us comment on the spectral assumptions done in the theorem.

They are satisfied in 1D: for instance in  $\Omega = (0, \pi)$ , the Dirichlet eigenfunctions  $\phi_j(x) = \sqrt{\frac{2}{\pi}} \sin(jx)$  are uniformly bounded in  $L^\infty(\Omega)$  and their squares weakly converge to  $1/\pi$ .

In multi-D, the assumptions are very strong and actually, except in the 1D case, we are not aware of domains  $\Omega$  for which the assumptions are satisfied. Firstly, in general the eigenfunctions are not uniformly bounded in  $L^\infty(\Omega)$  but, to the best of our knowledge, nothing seems to be known in general on the uniform  $L^p$ -boundedness property for some  $p > 2$ . Secondly the probability measures  $\mu_j = |\phi_j|^2 dx$  may have several weak limits. This question is related with deep open questions in mathematical physics and semi-classical analysis where one of the most fascinating open questions is to determine what can be these weak limits, called *quantum limits* or *semi-classical measures*. The famous Shnirelman theorem (see [Colin de Verdière \[1985\]](#), [Gérard and Leichtnam \[1993\]](#), [Šnirelman \[1974\]](#), and [Zelditch and Zworski \[1996\]](#)) states that, seeing the domain  $\Omega$  as a billiard, if the Riemannian geodesic flow is ergodic (for the canonical measure) then there exists a subsequence of  $(\mu_j)_{j \in \mathbb{N}^*}$  of density one converging vaguely to the uniform measure  $\frac{1}{|\Omega|} dx$  (Quantum Ergodicity, in short QE – still on the base, here). This result however lets open the possibility of having an exceptional subsequence of measures  $\mu_j$  converging vaguely to some other measure, for instance, the Dirac measure along a closed geodesic (*scars* in quantum physics, see [Faure, Nonnenmacher, and De Bièvre \[2003\]](#)). The QUE assumption mentioned above consists of assuming that the *whole* sequence  $(\mu_j)_{j \in \mathbb{N}^*}$  converges vaguely to the uniform measure. It is likely that QUE holds true on a negatively curved compact manifold (QUE conjecture, see [Sarnak \[2011\]](#) for a survey).

The idea is here that QUE ensures a delocalization property of the energy of high-frequency eigenfunctions. The quantity  $\int_\omega \phi_j^2(x) dx$  is interpreted as the probability of finding the quantum state of energy  $\lambda_j^2$  in  $\omega$ . The functional  $J(\chi_\omega)$  considered above can be viewed as a measure of eigenfunction concentration, which we seek to maximize over  $\mathcal{U}_L$ .

[Theorem 3](#) thus reveals intimate connections between domain optimization and asymptotic spectral properties or quantum ergodicity properties of  $\Omega$  (quantum chaos theory). It is interesting to notice that such a relationship was suggested in the early work [Chen, Fulling, Narcowich, and Sun \[1991\]](#) concerning the exponential decay properties of dissipative wave equations.

To end with these remarks on asymptotic properties on eigenfunctions, we note that the weak convergence of the measures  $\mu_j$  which is established in the several results mentioned above is however weaker than the convergence of the functions  $\phi_j^2$  for the weak topology of  $L^1(\Omega)$  that we need in our context. Indeed, the weak convergence of measures may fail to capture sets whose measure of the boundary is positive (such as Cantor of positive measure). This is why we also assume the  $L^p$  uniform boundedness property with  $p > 2$  because then, by the Portmanteau theorem and since  $\Omega$  is bounded, both notions of convergence coincide.

**The assumptions are not sharp.** The spectral assumptions made in [Theorem 3](#) are sufficient but are not necessary. It is indeed proved in [Privat, Trélat, and Zuazua \[2016a\]](#) that (7) is still satisfied if  $\Omega$  is a 2D square (with the usual eigenfunctions consisting of products of sine functions) or if  $\Omega$  is a 2D disk (with the usual eigenfunctions parametrized by Bessel functions), although, in the latter case, the eigenfunctions do not equidistribute as the eigenfrequencies increase, as illustrated by the well-known whispering galleries effect (see [Figure 3](#)): from the mathematical point of view, there exists a subsequence of  $(\mu_j)_{j \in \mathbb{N}^*}$  converging to the Dirac along the boundary of the disk.

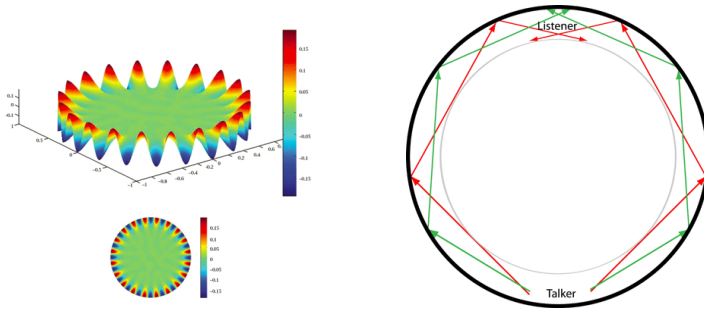


Figure 3: Whispering gallery phenomenon

**On the existence of an optimal set.** By [Theorems 1 and 3](#), the maximal possible value of  $C_{T,\text{rand}}(\chi_\omega)$  over the set  $\mathcal{U}_L$  is equal to  $TL/2$ . We now comment on the problem of existence of an optimal set: is the supremum reached in (7)? By compactness of the convexified set  $\overline{\mathcal{U}_L}$ , it is easy to see that the maximum of  $J$  over  $\overline{\mathcal{U}_L}$  is reached (in general in an infinite number of ways), but since  $\mathcal{U}_L$  is not compact for any appropriate topology, the question of the reachability of the supremum of  $J$  over  $\mathcal{U}_L$ , that is, the existence of an

optimal classical set, is a difficult question in general. In particular cases it can however be addressed using harmonic analysis (see [Privat, Trélat, and Zuazua \[2013b, 2016a\]](#)):

- In 1D, assume that  $\Omega = (0, \pi)$ , with the usual Hilbert basis of Dirichlet eigenfunctions made of sine functions. The supremum of  $J$  over  $\mathcal{U}_L$  (which is equal to  $L$ ) is reached if and only if  $L = 1/2$ . In that case, it is reached for all measurable subsets  $\omega \subset (0, \pi)$  of measure  $\pi/2$  such that  $\omega$  and its symmetric image  $\omega' = \pi - \omega$  are disjoint and complementary in  $(0, \pi)$ .
- In the 2D square  $\Omega = (0, \pi)^2$ , with the usual basis of Dirichlet eigenfunctions made of products of sine functions, the supremum of  $J$  over the more specific class of all possible subsets  $\omega = \omega_1 \times \omega_2$  of Lebesgue measure  $L\pi^2$ , where  $\omega_1$  and  $\omega_2$  are measurable subsets of  $(0, \pi)$ , is reached if and only if  $L \in \{1/4, 1/2, 3/4\}$ . In that case, it is reached for all such sets  $\omega$  satisfying

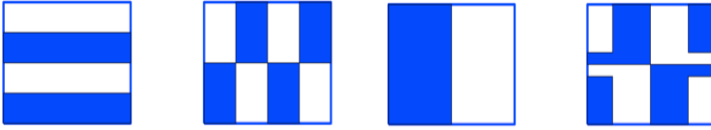


Figure 4:  $\Omega = (0, \pi)^2$ ,  $L = 1/2$ : examples of optimal sets. Note that the optimal sets on the left-side do not satisfy GCC and that  $C_T(\chi_\omega) = 0$  whereas  $C_{T,\text{rand}}(\chi_\omega) = TL/2$ .

$\frac{1}{4}(\chi_\omega(x, y) + \chi_\omega(\pi - x, y) + \chi_\omega(x, \pi - y) + \chi_\omega(\pi - x, \pi - y)) = L$  for almost all  $(x, y) \in [0, \pi]^2$  (see [Figure 4](#)).

- In the 2D disk  $\Omega = \{x \in \mathbb{R}^2 \mid \|x\| < 1\}$ , with the usual Hilbert basis of eigenfunctions defined in terms of Bessel functions, the supremum of  $J$  (which is equal to  $L$ ) over the class of all possible subsets  $\omega = \{(r, \theta) \in [0, 1] \times [0, 2\pi] \mid r \in \omega_r, \theta \in \omega_\theta\}$  such that  $|\omega| = L\pi$ , where  $\omega_r$  is any measurable subset of  $[0, 1]$  and  $\omega_\theta$  is any measurable subset of  $[0, 2\pi]$ , is reached if and only if  $L = 1/2$ . In that case, it is reached for all subsets  $\omega = \{(r, \theta) \in [0, 1] \times [0, 2\pi] \mid \theta \in \omega_\theta\}$  of measure  $\pi/2$ , where  $\omega_\theta$  is any measurable subset of  $[0, 2\pi]$  such that  $\omega_\theta$  and its symmetric image  $\omega'_\theta = 2\pi - \omega_\theta$  are disjoint and complementary in  $[0, 2\pi]$ .

In general, the question of the existence of an optimal set is completely open. In view of the partial results above and in view of the results of the next section, we conjecture that, for generic domains  $\Omega$  and for generic values of  $L \in (0, 1)$ , the supremum in (7) is not

reached and hence there does not exist any optimal set. We have no clue how to address this conjecture in general.

### 3 Spectral approximation of the optimal design problem

Motivated by the probable absence of optimal set for wave and Schrödinger equations as explained previously, and motivated by the objective of building the optimal set for parabolic equations, it is natural to consider the following finite-dimensional spectral approximation of the problem (6), namely:

$$(8) \quad \sup_{\chi_{\omega} \in \mathcal{U}_L} \min_{1 \leq j \leq N} \gamma_j(T) \int_{\omega} |\phi_j(x)|^2 dx$$

for any  $N \in \mathbb{N}^*$ . This is a spectral truncation where we keep only the  $N$  first modes. We have the following easy result.

**Theorem 4** (Privat, Trélat, and Zuazua [2015b, 2016a]). *Let  $T > 0$  be arbitrary. There exists a unique optimal observation domain  $\omega^N$  solving (8). Moreover  $\omega^N$  is open and semi-analytic and thus it has a finite number of connected components.*

Actually, since there is only a finite number of modes in (8), existence and uniqueness of an optimal set  $\omega^N$  is not difficult to prove (by a standard minimax argument), as well as a  $\Gamma$ -convergence property of  $J_N$  towards  $J$  for the weak star topology of  $L^\infty$ , where we have set  $J_N(\chi_{\omega}) = \min_{1 \leq j \leq N} \gamma_j(T) \int_{\omega} |\phi_j(x)|^2 dx$ . In particular, the sets  $\omega^N$  constitute a maximizing sequence for the (convexified) problem of maximizing  $J$  over  $\overline{\mathcal{U}}_L$ , and this, without geometric or ergodicity assumptions on  $\Omega$  (under the assumptions of Theorem 3, these sets constitute a maximizing sequence for the problem of maximizing  $J$  over  $\mathcal{U}_L$ ).

Let us now analyze how  $\omega^N$  behaves as  $N$  increases, by distinguishing between the parabolic case (3) and the hyperbolic case (1) and (2).

**3.1 The parabolic case.** For parabolic equations (3), under general assumptions on the operator  $A$ , which are satisfied for heat, Stokes equations and anomalous diffusion equations with  $\alpha > 1/2$ , remarkably, the sequence of optimal sets  $(\omega^N)_{N \in \mathbb{N}^*}$  is stationary.

**Theorem 5** (Privat, Trélat, and Zuazua [2015b]). *For every  $T > 0$  there exists  $N_0(T) \in \mathbb{N}^*$  such that*

$$\omega^{N_0(T)} = \omega^N = \omega^* \quad \forall N \geq N_0(T).$$

As a consequence, the optimal observation set  $\omega^*$  whose existence and uniqueness has been stated in Theorem 2 can actually be built from a finite-dimensional spectral approximation, by keeping only a finite number of modes. This stationarity property is illustrated

on [Figure 5](#) where we compute, as announced in the abstract, the “optimal thermometer in the square”. For this example, we have  $N_0(0.05) = 16$ , i.e., for  $T = 0.05$  the optimal

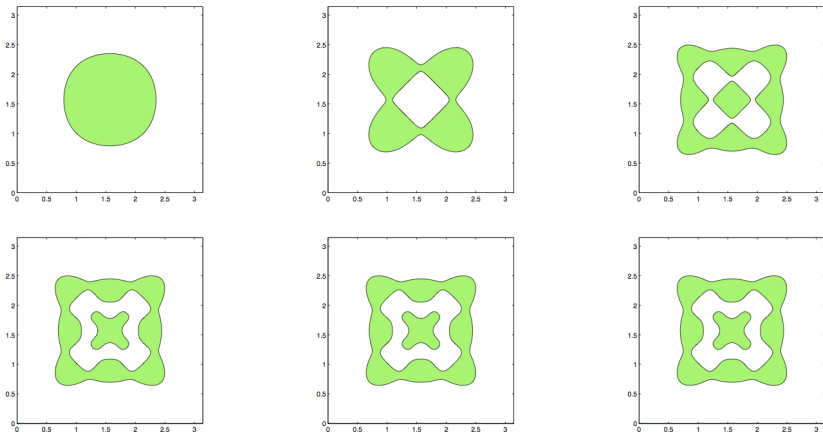


Figure 5: Dirichlet heat equation on  $\Omega = (0, \pi)^2$ ,  $L = 0.2$ ,  $T = 0.05$ . Row 1, from left to right: optimal domain  $\omega^N$  (in green) for  $N = 1, 4, 9$ . Row 2, from left to right: optimal domain  $\omega^N$  (in green) for  $N = 16, 25, 36$ .

domain is computed thanks to the 16 first eigenmodes.

It is also proved in [Privat, Trélat, and Zuazua \[2015b\]](#) that the function  $T \mapsto N_0(T) \in \mathbb{N}^*$  is nonincreasing and that if  $\operatorname{Re}(\lambda_1) < \operatorname{Re}(\lambda_2)$  then  $N_0(T) = 1$  as soon as  $T$  is large enough, which means that the optimal set  $\omega^*$  is entirely determined by the first eigenfunction if the observation time  $T$  is large.

**3.2 The hyperbolic case.** In contrast to the previous parabolic case, for wave and Schrödinger equations, the fact that all eigenmodes have the same weight ( $\gamma_j(T)$  remains constant) causes a strong instability of the optimal sets  $\omega^N$ , whose complexity increases drastically as  $N$  increases.

Moreover, the sets  $\omega^N$  have a finite number of connected components, expected to increase in function of  $N$ . The numerical simulations of [Figures 6 and 7](#) show the shapes of these sets. Their increasing complexity (number of connected components) which can be observed as  $N$  increases is in accordance with the conjecture of the nonexistence of an optimal set for (6).

Of course, however, up to some subsequence the sequence of maximizers  $\chi_{\omega^N}$  of  $J_N$  converges (in weak-star topology) to some maximizer  $a \in \overline{\mathcal{U}}_L$  of  $J$

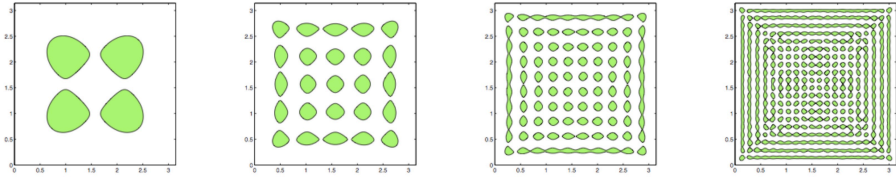


Figure 6:  $\Omega = (0, \pi)^2$ , Dirichlet boundary conditions,  $L = 0.2$ . From left to right:  $N = 4, 25, 100, 400$ . The optimal domain is in green.

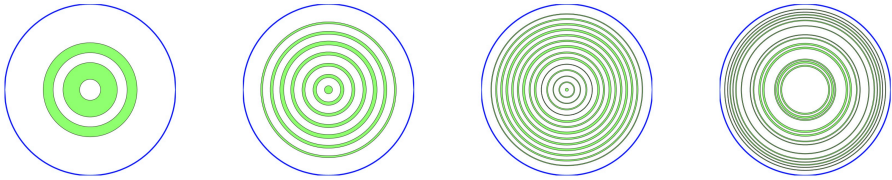


Figure 7:  $\Omega = \{x \in \mathbb{R}^2 \mid |x| \leq 1\}$ , Dirichlet boundary conditions,  $L = 0.2$ . From left to right:  $N = 4, 25, 100, 400$ . The optimal domain is in green.

In the 1D case  $\Omega = (0, \pi)$  with Dirichlet boundary conditions, it can be proved that, for  $L > 0$  sufficiently small, the optimal set  $\omega^N$  maximizing  $J_N$  is the union of  $N$  intervals concentrating around equidistant points and that  $\omega^N$  is actually the worst possible subset for the problem of maximizing  $J_{N+1}$ : in other words, the optimal domain for  $N$  modes is the worst possible one when considering the truncated problem with  $N + 1$  modes. This is the *spillover phenomenon*, noticed in Hébrard and Henrot [2005] and proved in Privat, Trélat, and Zuazua [2013b] (the proof is highly technical).

**Weighted observability inequalities.** This intrinsic instability is due to the fact that in (6) all modes have the same weight. This is so in the mathematical definition of the (deterministic or randomized) observability constant. One could argue that highfrequencies are difficult to observe and, trying to reflect the Heisenberg uncertainty principle of quantum physics, this leads to the intuition that lower frequencies should be in some sense more weighted than higher ones. One can then introduce a weighted version of the observability inequality (4), by considering, for instance the (equivalent) inequality

$$C_{T,\sigma}(\chi_\omega) (\|(y^0, y^1)\|_{L^2 \times H^{-1}}^2 + \sigma \|y^0\|_{H^{-1}}^2) \leq \int_0^T \int_\omega |y(t, x)|^2 dx dt$$

where  $\sigma \geq 0$  is some weight. We have  $C_{T,\sigma}(\chi_\omega) \leq C_T(\chi_\omega)$ , and considering as before an averaged version of this weighted observability inequality over random initial data leads to

$$C_{T,\sigma,\text{rand}}(\chi_\omega) = \frac{T}{2} \inf_{j \in \mathbb{N}^*} \frac{\lambda_j^2}{\sigma + \lambda_j^2} \int_\omega \phi_j(x)^2 dx$$

where now the weights are an increasing sequence of positive real numbers converging to 1. Actually, if  $\frac{\lambda_1^2}{\sigma + \lambda_1^2} < L < 1$  then highfrequencies do not play any role in the problem of maximizing  $C_{T,\sigma,\text{rand}}$  over  $\mathcal{U}_L$  and we have the following result for not too small values of  $L$

**Theorem 6 (Privat, Trélat, and Zuazua [2016a]).** *Assume that the whole sequence of probability measures  $\mu_j = \phi_j^2(x) dx$  converges vaguely to the uniform measure  $\frac{1}{|\Omega|} dx$  and that the sequence of eigenfunctions  $\phi_j$  is uniformly bounded in  $L^\infty(\Omega)$ . Then for every  $L \in \left(\frac{\lambda_1^2}{\sigma + \lambda_1^2}, 1\right)$  there exists  $N_0 \in \mathbb{N}^*$  such that*

$$\begin{aligned} \max_{\chi_\omega \in \mathcal{U}_L} \inf_{j \in \mathbb{N}^*} \frac{\lambda_j^2}{\sigma + \lambda_j^2} \int_\omega \phi_j(x)^2 dx &= \max_{\chi_\omega \in \mathcal{U}_L} \inf_{1 \leq j \leq N} \frac{\lambda_j^2}{\sigma + \lambda_j^2} \int_\omega \phi_j(x)^2 dx \\ &\leq \frac{\lambda_1^2}{\sigma + \lambda_1^2} < L \quad \forall N \geq N_0. \end{aligned}$$

*In particular, the problem of maximizing  $C_{T,\sigma,\text{rand}}$  over  $\mathcal{U}_L$  has a unique solution  $\chi_{\omega^{N_0}}$  and moreover the set  $\omega^{N_0}$  is open and semi-analytic.*

This result says that, when highfrequencies are weighted as above, there exists a unique optimal observation set if  $L$  is large enough, i.e., if one is allowed to cover a fraction of the whole domain  $\Omega$  that is large enough. This is similar to what we have obtained in the parabolic case. Moreover the optimal set can then be computed from a finite number of modes because the sequence of optimal sets  $\omega^N$  of the truncated problem is stationary. The threshold value  $\frac{\lambda_1^2}{\sigma + \lambda_1^2}$  becomes smaller when  $\sigma$  increases, in accordance with physical intuition. We do not know what may happen when  $L \leq \frac{\lambda_1^2}{\sigma + \lambda_1^2}$  but we suspect that the situation is the same as when  $\sigma = 0$  (spillover phenomenon and, probably, nonexistence of an optimal set); this conjecture is supported by some numerical simulations for the truncated problem (see Privat, Trélat, and Zuazua [ibid.]) which show that, when  $L$  is small, the optimal domains have an increasing complexity as  $N$  increases.

As before, we can notice that the assumptions of the above result, which are very strong, are not necessary and one can prove that the conclusion still holds true in a hypercube with Dirichlet boundary conditions when one considers the usual Hilbert basis made of products of sine functions.

## 4 Conclusion

We have modeled the problem of optimal shape and location of the observation or control domain having a prescribed measure, in terms of maximizing a spectral functional over all measurable subsets of fixed Lebesgue measure. This spectral functional can be interpreted as a randomized version of the observability constant over random initial data. For parabolic equations, we have existence and uniqueness of an optimal set, which can be determined from a finite number of modes. For wave and Schrödinger equations, the optimal observability problem is closely related to quantum chaos, in particular, asymptotic properties of eigenfunctions and we have seen that, generically, an optimal set should not exist, in accordance with the spillover phenomenon. In all cases, developing knowledge on concentration or delocalization properties of highfrequency eigenfunctions is crucial in order to address optimal observability issues.

We have seen that a way to avoid spillover and to recover existence and uniqueness of an optimal set for wave and Schrödinger equations is to consider weighted observability inequalities in which highfrequencies are penalized. Certainly, other approaches are possible, exploiting the physics of the problem.

**Optimal boundary observability.** In the paper we have focused on *internal* observation or control subdomains. Similar studies can be led for *boundary* subdomains. Optimal observability can be modeled by the optimal design problem

$$\sup_{|\omega|=L|\partial\Omega|} \inf_{j \in \mathbb{N}^*} \gamma_j(T) \int_{\omega} \frac{1}{\lambda_j} \left( \frac{\partial \phi_j}{\partial \nu} \right)^2 d\mathcal{H}^{n-1}$$

where now the Neumann traces of the Dirichlet-Laplacian eigenfunctions play a prominent role (in particular, their asymptotic properties). This problem, which interestingly can be interpreted as a spectral shape sensitivity problem, is studied in [Privat, Trélat, and Zuazua \[2018\]](#).

**On the deterministic observability constant.** We have let untouched the problem of maximizing the deterministic observability constant  $C_T(\chi_{\omega})$  over  $\mathcal{U}_L$ . Although we have explained that this problem is certainly less relevant in practice where a large number of measurements is performed, it is anyway very interesting from the mathematical point of view. The crossed terms (which we have ruled out by randomization) are then expected to have an important role. A first remark is that, extending the functional  $C_T$  to  $\overline{\mathcal{U}}_L$ , for 1D wave equations the constant density  $a \equiv L$  is not a maximizer of  $C_T$  if  $T \notin \pi\mathbb{N}^*$ . Knowing if there is a relaxation phenomenon or not is an open problem.



Another remark is the following. It is proved in [Humbert, Privat, and Trélat \[2016\]](#) that, for the wave equation (1), given any measurable subset  $\omega$  of  $\Omega$ , we have

$$\lim_{T \rightarrow +\infty} \frac{C_T(\chi_\omega)}{T} = \frac{1}{2} \min \left( \inf_{j \in \mathbb{N}^*} \int_{\omega} \phi_j(x)^2 dx, \lim_{T \rightarrow +\infty} \inf_{\gamma \in \Gamma} \frac{1}{T} \int_0^T \chi_{\bar{\omega}}(\gamma(t)) dt \right)$$

where  $\Gamma$  is the set of all geodesic rays on  $\Omega$ , provided that  $\omega$  has no grazing ray, i.e., provided that there exists no  $\gamma \in \Gamma$  such that  $\gamma(t) \in \partial\omega$  over a set of times of positive measure. This equality says that, in large time, the deterministic observability constant  $C_T(\chi_\omega)$  is the minimum of two quantities: the first one is exactly  $C_{T,\text{rand}}(\omega)$ , which is the functional we have focused on throughout the paper; the second one is of a geometric nature and provides an account for the average time spent by geodesic rays in the observation subset. Although this result is only valid asymptotically in time, it gives the intuition that geodesic rays play an important role. In order to address the problem of maximizing  $C_T(\chi_\omega)$  over  $\mathcal{U}_L$ , one should first try solve, for any  $T > 0$ ,

$$\sup_{\chi_\omega \in \mathcal{U}_L} \inf_{\gamma \in \Gamma} \frac{1}{T} \int_0^T \chi_\omega(\gamma(t)) dt.$$

This is an interesting optimal design problem.

**Discretization issues.** In the search of an optimal observation domain for a PDE model, certainly the most usual approach in engineering applications is to discretize the PDE (for instance by means of finite elements), thus obtaining a family of equations in finite dimension, indexed by some  $h > 0$  which can be thought as the size of the mesh. Given some fixed  $h$ , one then performs an optimal design procedure to find, if it exists, an optimal observation set  $\omega^h$ . The question is then natural to ask whether  $\omega^h$  converges, as  $h \rightarrow 0$ , to the (if it exists and is unique) optimal observation set  $\omega^*$  of the complete model. In other words, do the numerical optimal designs converge to the continuous optimal design as the mesh size tends to 0? Under which assumptions do the optimal designs commute with discretization schemes?

We have seen with the spectral truncation (which is a particular discretization method) that the answer is certainly negative for wave and Schrödinger equations but is positive for parabolic equations. The question is open for general discretization schemes and is of great interest in view of practical applications, all the more than discrete or semi-discrete models are often employed.

## References

- Grégoire Allaire, Sylvie Aubry, and François Jouve (2001). “Eigenfrequency optimization in optimal design”. *Comput. Methods Appl. Mech. Engrg.* 190.28, pp. 3565–3579. MR: [1819157](#) (cit. on p. [3868](#)).
- Grégoire Allaire, Arnaud Münch, and Francisco Periago (2010). “Long time behavior of a two-phase optimal design for the heat equation”. *SIAM J. Control Optim.* 48.8, pp. 5333–5356. MR: [2745777](#) (cit. on p. [3868](#)).
- J. Apraiz, L. Escauriaza, G. Wang, and C. Zhang (2014). “Observability inequalities and measurable sets”. *J. Eur. Math. Soc. (JEMS)* 16.11, pp. 2433–2475. MR: [3283402](#) (cit. on p. [3870](#)).
- Claude Bardos, Gilles Lebeau, and Jeffrey Rauch (1992). “Sharp sufficient conditions for the observation, control, and stabilization of waves from the boundary”. *SIAM J. Control Optim.* 30.5, pp. 1024–1065. MR: [1178650](#) (cit. on p. [3863](#)).
- J. C. Bellido and A. Donoso (2007). “An optimal design problem in wave propagation”. *J. Optim. Theory Appl.* 134.2, pp. 339–352. MR: [2332468](#) (cit. on p. [3868](#)).
- G. Chen, S. A. Fulling, F. J. Narcowich, and S. Sun (1991). “Exponential decay of energy of evolution equations with locally distributed damping”. *SIAM J. Appl. Math.* 51.1, pp. 266–301. MR: [1089141](#) (cit. on p. [3872](#)).
- Y. Colin de Verdière (1985). “Ergodicité et fonctions propres du laplacien”. *Comm. Math. Phys.* 102.3, pp. 497–502. MR: [818831](#) (cit. on p. [3872](#)).
- Yves Colin de Verdière, Luc Hillairet, and Emmanuel Trélat (2018). “Spectral asymptotics for sub-Riemannian Laplacians, I: Quantum ergodicity and quantum limits in the 3-dimensional contact case”. *Duke Math. J.* 167.1, pp. 109–174. MR: [3743700](#).
- Fariba Fahroo and Kazufumi Ito (1996). “Optimum damping design for an abstract wave equation”. *Kybernetika (Prague)* 32.6. New directions in control and automation, I (Li-massol, 1995), pp. 557–574. MR: [1438105](#) (cit. on p. [3868](#)).
- Frédéric Faure, Stéphane Nonnenmacher, and Stephan De Bièvre (2003). “Scarred eigenstates for quantum cat maps of minimal periods”. *Comm. Math. Phys.* 239.3, pp. 449–492. MR: [2000926](#) (cit. on p. [3872](#)).
- Enrique Fernández-Cara and Arnaud Münch (2012). “Numerical null controllability of semi-linear 1-D heat equations: fixed point, least squares and Newton methods”. *Math. Control Relat. Fields* 2.3, pp. 217–246. MR: [2991568](#) (cit. on p. [3868](#)).
- Patrick Gérard and Éric Leichtnam (1993). “Ergodic properties of eigenfunctions for the Dirichlet problem”. *Duke Math. J.* 71.2, pp. 559–607. MR: [1233448](#) (cit. on p. [3872](#)).
- Joachim Hartung (1982). “An extension of Sion’s minimax theorem with an application to a method for constrained games”. *Pacific J. Math.* 103.2, pp. 401–408. MR: [705239](#) (cit. on p. [3870](#)).

- Pascal Hébrard and Antoine Henrot (2003). “[Optimal shape and position of the actuators for the stabilization of a string](#)”. *Systems Control Lett.* 48.3-4. Optimization and control of distributed systems, pp. 199–209. MR: [2020637](#) (cit. on p. [3867](#)).
- (2005). “[A spillover phenomenon in the optimal location of actuators](#)”. *SIAM J. Control Optim.* 44.1, pp. 349–366. MR: [2177160](#) (cit. on pp. [3867](#), [3877](#)).
- Emmanuel Humbert, Yannick Privat, and Emmanuel Trélat (2016). “[Observability properties of the homogeneous wave equation on a closed manifold](#)”. arXiv: [1607.01535](#) (cit. on pp. [3864](#), [3880](#)).
- Stéphane Jaffard and Sorin Micu (2001). “Estimates of the constants in generalized Ingham’s inequality and applications to the control of the wave equation”. *Asymptot. Anal.* 28.3-4, pp. 181–214. MR: [1878794](#) (cit. on p. [3865](#)).
- Stéphane Jaffard, Marius Tucsnak, and Enrique Zuazua (1997). “[On a theorem of Ingham](#)”. *J. Fourier Anal. Appl.* 3.5. Dedicated to the memory of Richard J. Duffin, pp. 577–582. MR: [1491935](#) (cit. on p. [3865](#)).
- Sudarshan Kumar and J Seinfeld (1978). “Optimal location of measurements for distributed parameter estimation”. *IEEE Transactions on Automatic Control* 23.4, pp. 690–698 (cit. on p. [3868](#)).
- Jérôme Le Rousseau, Gilles Lebeau, Peppino Terpolilli, and Emmanuel Trélat (2017). “[Geometric control condition for the wave equation with a time-dependent observation domain](#)”. *Anal. PDE* 10.4, pp. 983–1015. MR: [3649373](#) (cit. on p. [3863](#)).
- Kirsten Morris (2011). “[Linear-quadratic optimal actuator location](#)”. *IEEE Trans. Automat. Control* 56.1, pp. 113–124. MR: [2777204](#) (cit. on p. [3868](#)).
- Arnaud Münch and Francisco Periago (2011). “[Optimal distribution of the internal null control for the one-dimensional heat equation](#)”. *J. Differential Equations* 250.1, pp. 95–111. MR: [2737836](#) (cit. on p. [3868](#)).
- Yannick Privat and Emmanuel Trélat (2015). “[Optimal design of sensors for a damped wave equation](#)”. *Discrete Contin. Dyn. Syst.* Dynamical systems, differential equations and applications. 10th AIMS Conference. Suppl. Pp. 936–944. MR: [3462528](#) (cit. on p. [3867](#)).
- Yannick Privat, Emmanuel Trélat, and Enrique Zuazua (2013a). “[Optimal location of controllers for the one-dimensional wave equation](#)”. *Ann. Inst. H. Poincaré Anal. Non Linéaire* 30.6, pp. 1097–1126. MR: [3132418](#) (cit. on p. [3867](#)).
- (2013b). “[Optimal observation of the one-dimensional wave equation](#)”. *J. Fourier Anal. Appl.* 19.3, pp. 514–544. MR: [3048589](#) (cit. on pp. [3865](#), [3866](#), [3874](#), [3877](#)).
  - (2015a). “[Complexity and regularity of maximal energy domains for the wave equation with fixed initial data](#)”. *Discrete Contin. Dyn. Syst.* 35.12, pp. 6133–6153. MR: [3393270](#) (cit. on pp. [3866](#), [3870](#)).

- (2015b). “Optimal shape and location of sensors for parabolic equations with random initial data”. *Arch. Ration. Mech. Anal.* 216.3, pp. 921–981. MR: [3325779](#) (cit. on pp. [3864](#), [3866](#), [3869](#), [3875](#), [3876](#)).
  - (2016a). “Optimal observability of the multi-dimensional wave and Schrödinger equations in quantum ergodic domains”. *J. Eur. Math. Soc. (JEMS)* 18.5, pp. 1043–1111. MR: [3500831](#) (cit. on pp. [3864](#), [3866](#), [3870](#), [3871](#), [3873–3875](#), [3878](#)).
  - (2016b). “Randomised observation, control and stabilization of waves [Based on the plenary lecture presented at the 86th Annual GAMM Conference, Lecce, Italy, March 24, 2015]”. *ZAMM Z. Angew. Math. Mech.* 96.5, pp. 538–549. MR: [3502963](#) (cit. on p. [3867](#)).
  - (2017). “Actuator design for parabolic distributed parameter systems with the moment method”. *SIAM J. Control Optim.* 55.2, pp. 1128–1152. MR: [3632257](#) (cit. on p. [3867](#)).
- Yannick Privat, Emmanuel Trélat, and Enrique Zuazua (2018). “Spectral shape optimization for Neumann traces of the Dirichlet-Laplacian eigenfunctions”. *Preprint Hal*, p. 40 (cit. on p. [3879](#)).
- Peter Sarnak (2011). “Recent progress on the quantum unique ergodicity conjecture”. *Bull. Amer. Math. Soc. (N.S.)* 48.2, pp. 211–228. MR: [2774090](#) (cit. on p. [3872](#)).
- Ole Sigmund and Jakob Søndergaard Jensen (2003). “Systematic design of phononic band-gap materials and structures by topology optimization”. *R. Soc. Lond. Philos. Trans. Ser. A Math. Phys. Eng. Sci.* 361.1806, pp. 1001–1019. MR: [1995446](#) (cit. on p. [3868](#)).
- A. I. Šnirelman (1974). “Ergodic properties of eigenfunctions”. *Uspehi Mat. Nauk* 29.6 (180), pp. 181–182. MR: [0402834](#) (cit. on p. [3872](#)).
- Dariusz Uciński and Maciej Patan (2010). “Sensor network design for the estimation of spatially distributed processes”. *Int. J. Appl. Math. Comput. Sci.* 20.3, pp. 459–481. MR: [2743850](#) (cit. on p. [3868](#)).
- Marc van de Wal and Bram de Jager (2001). “A review of methods for input/output selection”. *Automatica J. IFAC* 37.4, pp. 487–510. MR: [1832938](#) (cit. on p. [3868](#)).
- Alain Vande Wouwer, Nicolas Point, Stephanie Porteman, and Marcel Remy (2000). “An approach to the selection of optimal sensor locations in distributed parameter systems”. *Journal of process control* 10.4, pp. 291–300 (cit. on p. [3868](#)).
- Steven Zelditch and Maciej Zworski (1996). “Ergodicity of eigenfunctions for ergodic billiards”. *Comm. Math. Phys.* 175.3, pp. 673–682. MR: [1372814](#) (cit. on p. [3872](#)).

Received 2017-11-19.

SORBONNE UNIVERSITÉ  
UNIVERSITÉ PARIS-DIDEROT SPC  
CNRS, INRIA  
LABORATOIRE JACQUES-LOUIS LIONS, ÉQUIPE CAGE  
F-75005 PARIS  
FRANCE  
[emmanuel.trelat@upmc.fr](mailto:emmanuel.trelat@upmc.fr)  
[emmanuel.trelat@sorbonne-universite.fr](mailto:emmanuel.trelat@sorbonne-universite.fr)

# GRAPHICAL MODELS IN MACHINE LEARNING, NETWORKS AND UNCERTAINTY QUANTIFICATION

ANDREA L. BERTOZZI

## Abstract

This paper is a review article on semi-supervised and unsupervised graph models for classification using similarity graphs and for community detection in networks. The paper reviews graph-based variational models built on graph cut metrics. The equivalence between the graph mincut problem and total variation minimization on the graph for an assignment function allows one to cast graph-cut variational problems in the language of total variation minimization, thus creating a parallel between low dimensional data science problems in Euclidean space (e.g. image segmentation) and high dimensional clustering. The connection paves the way for new algorithms for data science that have a similar structure to well-known computational methods for nonlinear partial differential equations. This paper focuses on a class of methods build around diffuse interface models (e.g. the Ginzburg–Landau functional and the Allen–Cahn equation) and threshold dynamics, developed by the Author and collaborators. Semi-supervised learning with a small amount of training data can be carried out in this framework with diverse applications ranging from hyperspectral pixel classification to identifying activity in police body worn video. It can also be extended to the context of uncertainty quantification with Gaussian noise models. The problem of community detection in networks also has a graph-cut structure and algorithms are presented for the use of threshold dynamics for modularity optimization. With efficient methods, this allows for the use of network modularity for unsupervised machine learning problems with unknown number of classes.

## 1 Similarity Graphs and Spectral Clustering

Graphical models provide a mathematical structure for high dimensional data problems that yield important latent information in the data. They are also the basic building block

---

This work was supported by NSF grants DMS-1737770, DMS-1417674, NIJ grant 2014-R2-CX-0101, and ONR grant N00014-16-1-2119.

*MSC2010:* primary 65K10; secondary 35Q56, 49M20, 6209, 91C20, 91D30.

*Keywords:* diffuse interfaces, graphical models, graph Laplacian, machine learning, uncertainty quantification, social networks, community detection, data clustering, modularity, MBO scheme.

for network analysis. Graphical models yield useful information about connections between pieces of data from pairwise comparisons of that data, most notably via a *similarity graph* in which nodes represent pieces of data and edge weights are related to pairwise comparisons of the data. For machine learning methods, a major challenge is the inherent  $O(N^2)$  computational complexity of the weights (for  $N$  nodes) unless the graph is sparse. Another source of complexity is the number of classes. Furthermore, most machine learning methods, including those developed for complex graphical models, are based on linear algebra and linear models. Graph-based structures have the potential to provide a useful framework that is inherently nonlinear providing a broader framework for the data structures. For classification in machine learning there are basic methods like Support Vector Machine, which identifies a hyperplane separating different classes of data. This is a **supervised** algorithm involving a lot of training data with small amounts of unknown data. Kernel methods allow for unknown nonlinear mappings to be computed as part of the methodology. Still this restricts that data to have a certain form and for the mapping to be learned or computed.

In contrast, a similarity graph allows analysis of the data by performing operations on the graph itself, thus removing the original high-dimensionality of the problem. Linear structures have been studied, most notably the graph Laplacian matrix of the form  $L = D - W$  where  $W$  is the weight matrix of off-diagonal elements  $w_{ij}$  and the diagonal matrix  $D$  has each entry  $d_i$  equal to the sum of the weights connected to node  $i$ . Spectral clustering is an **unsupervised** method in which clusters are determined by a k-means method applied to a small set of eigenfunctions of the graph Laplacian matrix [von Luxburg \[2007\]](#). Spectral clustering can be paired with a random sampling method using the Nystrom extension, that allows for an approximately  $O(N)$  low-rank approximation of the graph Laplacian matrix. Spectral clustering in machine learning requires the graph to be constructed from data. Similarity graphs are well-known in machine learning and have each node corresponding to a feature vector  $V_i$  comprised of high-dimensional data to be classified, and the weights  $w_{ij}$  between nodes are computed as a pairwise comparison between the feature vectors. Some examples include:

1. The Gaussian function

$$(1) \quad w_{i,j} = \exp(-||V_i - V_j||^2/\tau)$$

Depending on the choice of metric, this similarity function includes the Yaroslavsky filter [Yaroslavsky \[1985\]](#) and the nonlocal means filter [Buades, Coll, and Morel \[2005\]](#).

2. Gaussian with cosine angle

$$(2) \quad w_{i,j} = \exp - \frac{(1 - \frac{\langle V_i, V_j \rangle}{|V_i||V_j|})^2}{2\sigma^2}$$

is a common similarity function used in hyperspectral imaging. In this case one is interested in alignment of feature vectors rather than their Euclidean distance.

3. Zelnik-Manor and Perona introduced local scaling weights for sparse matrix computations [Zelnik-Manor and Perona \[2004\]](#). Given a metric  $d(V_i, V_j)$  between each feature vector, they define a local parameter  $\sqrt{\tau(V_i)}$  for each  $V_i$ . The choice in [Zelnik-Manor and Perona \[ibid.\]](#) is  $\sqrt{\tau(V_i)} = d(V_i, V_M)$ , where  $V_M$  is the  $M$ th closest vector to  $V_i$ . The similarity matrix is then defined as

$$(3) \quad w_{i,j} = \exp\left(-\frac{d(V_i, V_j)^2}{\sqrt{\tau(V_i)\tau(V_j)}}\right).$$

This similarity matrix is better at segmentation when there are multiple scales that need to be segmented simultaneously.

There are two popular normalization procedures for the graph Laplacian, and the normalization has segmentation consequences [F. R. K. Chung \[1996\]](#) and [von Luxburg \[2007\]](#). The normalization that is often used for the nonlocal means graph for images is the symmetric Laplacian  $L_s$  defined as

$$(4) \quad L_s = D^{-1/2} L D^{-1/2} = I - D^{-1/2} W D^{-1/2}.$$

The symmetric Laplacian is named as such since it is a symmetric matrix. The random walk Laplacian is another important normalization given by

$$(5) \quad L_w = D^{-1} L = I - D^{-1} W.$$

The random walk Laplacian is closely related to discrete Markov processes.

A novel example of spectral clustering applied to social science data is presented in [van Gennip, Hunter, et al. \[2013\]](#). LAPD Field Interview (FI) cards provide a unique opportunity to investigate the relationships between individual use of space, social networks and group identities, specifically criminal street gang affiliation. FI cards are completed when a patrol officer comes into contact with a member of the public. They record spatio-temporal data about where and when the stop occurred, individual characteristics (e.g., name and home address) and demographic characteristics (e.g., age, sex, ethnic group). FI cards also record information about criminal activity and gang affiliation, if applicable. Critical here is information on gang membership. Known or suspected members of gangs have their gang affiliation recorded, gang moniker if known, and information on the duration of gang membership (e.g., member since 2004). FI cards also record instances where two or more gang members were stopped and interviewed together. Thus, each FI with two or more gang members represents a spatial sample of occasions when nodes in a social



network interacted. We developed a graphical model using both social network information from raw observations and spatial coordinates of these observations. Figure 1 shows results of spectral clustering using the composite graph with both information - the result finds latent groups of individuals that differ from the known gang affiliations as illustrated in the Pie chart. The work in Figure 1 used standard spectral clustering methods to identify

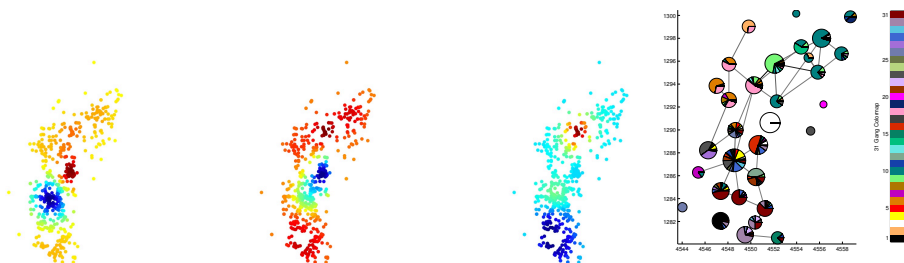


Figure 1: Spectral clustering applied to LAPD Hollenbeck Division Field Interview card data for 2009 [van Gennip, Hunter, et al. \[2013\]](#). Left eigenvalues associated with event data shown with geographic placement. (right) Pie charts showing clusters identified by spectral clustering compared with known ground truth of gang affiliation (shown in colors) for the 31 gangs in Hollenbeck. Copyright © 2013 Society for Industrial and Applied Mathematics. Reprinted with permission. All rights reserved.

latent groups in the geo-social set space. The results show that natural groupings Hollenbeck sometimes are comprised of mostly one gang but othertimes, especially in areas with different gangs in high spatial proximity, can have members of multiple gangs affiliated with the same observed group.

## 2 Data classification and the Ginzburg-Landau functional on graphs

The Author and Arjuna Flenner developed the first suite of binary classifiers for semi-supervised machine learning (minimal training data) using a fully nonlinear model for similarity graphs [Bertozzi and Flenner \[2012\]](#). We proposed the Ginzburg-Landau (GL) functional as a smooth relaxation of graph total variation (equivalent to graph cuts), as a regularizer for semi-supervised learning. For large datasets, we incorporate efficient linear algorithms into a nonlinear PDE-based method for non-convex optimization. Our work has been republished as a SIGEST paper [Bertozzi and Flenner \[2016\]](#). Over 50 new

papers and methods have arisen from this work including fast methods for nonlocal means image processing using the MBO scheme [Merkurjev, Kostic, and Bertozzi \[2013\]](#), multi-class learning methods [Garcia-Cardona, Merkurjev, Bertozzi, Flenner, and Percus \[2014\]](#) and [Iyer, Chanussot, and Bertozzi \[2017\]](#), parallel methods for exascale-ready platforms [Meng, Koniges, He, S. Williams, Kurth, Cook, Deslippe, and Bertozzi \[2016\]](#), hyperspectral video analysis [Hu, Sunu, and Bertozzi \[2015\]](#), [Merkurjev, Sunu, and Bertozzi \[2014\]](#), [Meng, Merkurjev, Koniges, and Bertozzi \[2017\]](#), and [W. Zhu, Chayes, Tiard, S. Sanchez, Dahlberg, Bertozzi, Osher, Zosso, and Kuang \[2017\]](#), modularity optimization for network analysis [Hu, Laurent, Porter, and Bertozzi \[2013\]](#) and [Boyd, Bai, X. C. Tai, and Bertozzi \[2017\]](#), measurement techniques in Zoology [Calatroni, van Gennip, Schönlieb, Rowland, and Flenner \[2017\]](#), generalizations to hypergraphs [Bosch, Klamt, and Stoll \[2016\]](#), Pagerank [Merkurjev, Bertozzi, and F. Chung \[2016\]](#) and Cheeger cut based methods [Merkurjev, Bertozzi, Yan, and Lerman \[2017\]](#). This paper reviews some of this literature and discusses future problem areas including crossover work between network modularity and machine learning and efforts in uncertainty quantification.

Given a phase field variable  $u$ , the Ginzburg-Landau energy, introduced for Euclidean space in the last century, involves a competition between the convex functional  $\int (\nabla u)^2 dx$  that induces smoothing, with a double well function  $\int W(u) dx$ , that separates its argument into phases. The Bertozzi-Flenner graph model replaces the first term with the graph Dirichlet energy,  $\sum_{ij} w_{ij} (u_i - u_j)^2$ , equivalent to the inner product of  $Lu$  with  $u$  where  $L$  is the graph Laplacian:

$$(6) \quad E_{GL}(f) = \frac{1}{\epsilon} \langle Lf, f \rangle + \epsilon \sum_i (W(f_i)).$$

For a variant of the GL functional in [Equation \(6\)](#) one can prove Gamma convergence of the vanishing  $\epsilon$  limit to the graph TV functional [van Gennip and Bertozzi \[2012\]](#), equivalent to the graph cut energy:

$$(7) \quad E_{TV}(u) = \sum_{ij} w_{ij} |u_i - u_j|,$$

for  $u$  defining a graph partition. More recent work extending these results is [Thorpe and Theil \[2017\]](#). An equivalent result has been known for Euclidean space for several decades [Kohn and Sternberg \[1989\]](#). Another variant involves a wavelet GL functional [Dobrosotskaya and Bertozzi \[2008\]](#) which has a Wulff shape energy as its sharp interface Gamma-limit [Dobrosotskaya and Bertozzi \[2010\]](#). The GL functional is useful, in lieu of L1 compressed sensing methods for minimizing total variation, because the computationally expensive graph information only arises in the Dirichlet energy, leveraging optimization algorithms that can exploit efficient approximations of the graph Laplacian.

The nonlinear structure can be reduced to simple calculations such as local thresholding, as shown in the MBO scheme below. The graph cut functional or equivalent TV functional can be incorporated into a semi-supervised or unsupervised learning problem. Without additional terms in the energy, the minimizer of the energy is trivial - simply pick  $u$  to be a constant, one of the minimizers of the well  $W$ . However nontrivial solutions can be found by modifying the energy to include a semi-supervised penalty term or additional balance terms in the case of unsupervised learning problems. For semi-supervised learning we consider an  $L^2$  penalty for known training data (defined to be set  $S$  and with values  $u_0$  along with a graph cut term to minimize the sum of the weights between unlike classes:

$$E_1(u) = |u|_{TV} + \sum_{i \in S} \frac{\lambda}{2} (u_0(i) - u(i))^2 \approx E_{GL}(u) + \sum_{i \in S} \frac{\lambda}{2} (u_0(i) - u(i))^2.$$

The second term is for semi-supervision and the first is for the graph cut. The parameter  $\lambda$  provides a soft constraint for semi-supervision. In many applications discussed below the supervision involves a small amount of training data, e.g. 10% or less, compared to the majority of the data for supervised learning such as SVM.

The semi-supervised learning problem described above can be minimized quickly on very large datasets using a pseudo-spectral method involving the eigenfunctions and eigenvalues of the graph Laplacian and convex splitting methods [Schönlieb and Bertozzi \[2011\]](#) from nonlinear PDE. The important eigenfunctions can be computed very quickly for large datasets using sub-sampling methods, e.g. the Nyström extension [Belongie, Fowlkes, F. Chung, and Malik \[2002\]](#), [Fowlkes, Belongie, F. Chung, and Malik \[2004\]](#), and [Fowlkes, Belongie, and Malik \[2001\]](#). What is remarkable is that the entire TV minimization problem can be solved without computing all the weights of the graph (which can be prohibitive in the case of e.g. nonlocal means used in image processing with textures) [Buades, Coll, and Morel \[2005\]](#), [Gilboa and Osher \[2007, 2008\]](#), and [Merkurjev, Sunu, and Bertozzi \[2014\]](#). While there are other fast algorithms out there for TV minimization (e.g. the split Bregman method [Goldstein and Osher \[2009\]](#)) none of them can easily be adapted to use the fast algorithms for eigenfunctions that rely on having a symmetric matrix. Indeed the algorithms presented in this paper only require the knowledge of the important eigenfunctions of the graph Laplacian and do not require the computation of a “right hand side” that arises in more general TV minimization algorithms, such as split Bregman.

An unsupervised learning model can be constructed as a generalization of the piecewise constant Mumford-Shah model from image segmentation, applied to a graphical data model. We recall the the piecewise constant Mumford-Shah model [T. Chan and Vese \[2001\]](#), [Vese and T. F. Chan \[2002\]](#), and [Esedoğlu and Tsai \[2006\]](#) involves the identification of a contour  $\Phi$  that divides the image up into  $\hat{n}$  regions  $\Omega_r$ . The energy to minimize

is

$$E(\Phi, \{c_r\}_{r=1}^{\hat{n}}) = |\Phi| + \lambda \sum_{r=1}^{\hat{n}} \int_{\Omega_r} (u_0 - c_r)^2$$

where  $u_0$  is the observed image data,  $c_r$  denotes a constant approximation of the image in the set  $\Omega_r$  and  $|\Phi|$  denotes the length of the contour  $\Phi$ . The works [Hu, Sunu, and Bertozzi \[2015\]](#) and [Meng, Merkurjev, Koniges, and Bertozzi \[2017\]](#) present a generalization of this to graphical models. The examples studied are largely hyperspectral imagery however the idea could be applied to other high dimensional vectors. The data is used both to create the similarity graph and to solve the clustering problem because the constants will be chosen in the high dimensional space to approximate the high dimensional vectors within each class. This is different from the previous example in which the clustering can be computed outside of the high dimensional data space once the graph is known (or approximated) and the training data is known. More specifically, we consider the energy

$$E_2 = \frac{1}{2} |f|_{TV} + \lambda \sum_{r=1}^{\hat{n}} \sum_i f_r(n_i) \|u_0(n_i) - c_r\|^2,$$

where  $f$  is a simplex constrained vector value that indicates class assignment:

$$f : G \rightarrow \{0, 1\}^{\hat{n}}, \sum_{r=1}^{\hat{n}} f_r(n_i) = 1\}.$$

Specifically if  $f_r(n_i) = 1$  for some  $r$  then the data at node  $n_i$  belongs to the  $r$ -th class. For each  $f$  we have a partition of the graph into at most  $\hat{n}$  classes. The connection to the original piecewise constant Mumford-Shah model is that  $f_r$  is the characteristic function of the  $r$ -th class and thus  $\sum_i f_r(n_i) \|u_0(n_i) - c_r\|^2$  is analogous to the term  $\int_{\Omega_r} (u_0 - c_r)^2$  while the TV norm on graphs is the analogue of the length of the boundary in the Euclidean space problem.

**2.1 The MBO scheme on Graphs.** Rather than minimizing the GL functional, using an efficient convex splitting method such as in [Bertozzi and Flenner \[2016\]](#), we can use an even more efficient MBO method. Using the original Euclidean GL functional and classical PDE methods, [Esedoglu and Tsai \[2006\]](#) developed a simple algorithm for piecewise-constant image segmentation that alternated between evolution of the heat equation and thresholding. That paper built on even earlier work by [Merriman, Bence, and Osher \[1992\]](#) (MBO) for motion by mean curvature. Motivated by this work, the MBO computational scheme was extended to the graphical setting by [Merkurjev, Kostic, and Bertozzi \[2013\]](#) for binary classification and methods that build on binary classification such as bit-wise

greyscale classification for inpainting of greyscale images. The Graph MBO scheme for semi-supervised learning consists of the following two steps:

1. Heat equation with forcing term. Propagate using

$$\frac{u^{(n+1/2)} - u^{(n)}}{dt} = -L_u^{(n)} - \lambda(i)(u^{(n)} - u_0)$$

2. Threshold.

$$u^{(n+1)} = \begin{cases} 1 & \text{if } u^{(n+1/2)} \geq 0 \\ 0 & \text{if } u^{(n+1/2)} < 0. \end{cases}$$

The results in [Merkurjev, Kostic, and Bertozzi \[2013\]](#) showed significant speed-up in run time compared to the Ginzburg-Landau method developed here and also faster run times than the split-Bregman method applied to the Osher-Gilboa nonlocal means graph for the same datasets. Both the GL and MBO methods for binary learning were extended to the multiclass case in [Garcia-Cardona, Merkurjev, Bertozzi, Flenner, and Percus \[2014\]](#). The MBO scheme in particular is trivial to extend - the algorithm is the same except that the classes are defined taking the range of  $u$  in  $\hat{n}$  dimensions where  $\hat{n}$  is the number of classes and thresholding to the corners of the simplex. The MBO scheme is quite fast and in most cases finds the global minimum. For unusual problems that require a provably optimal solution, we have considered methods built around max flow and ADMM methods that are less efficient than MBO, but they can guarantee a global optimal solution for the binary, semi-supervised segmentation problem [Merkurjev, Bae, Bertozzi, and X.-C. Tai \[2015\]](#).

As an example of a high dimensional problem with multiple classes, consider the classification of hyperspectral pixels in a video sequence. [Figure 2](#) shows data from standoff detection of a glass plume using 128 spectra in the Long Wave Infrared (LWIR) from the Dugway Proving Ground. The graph weights are computed with spectral angle. The Nyström extension provides eigenfunctions of the graph Laplacian, which can run in Matlab in 2 minutes on a modest laptop. The actual classification runs in seconds. The Nyström method and the MBO scheme have recently been optimized on an exascale-ready platform at the National Energy Research Supercomputing Center (NERSC) [Meng, Koniges, He, S. Williams, Kurth, Cook, Deslippe, and Bertozzi \[2016\]](#).

Inspired by the work in [Esedoglu and Otto \[2015\]](#), one can translate the MBO scheme into a discrete time approximate graph cut minimization method. In [Hu, Sunu, and Bertozzi \[2015\]](#) and [van Gennip, Guillen, Osting, and Bertozzi \[2014\]](#) it is shown that the diffusion operator  $\Gamma_\tau = e^{-\tau L}$  where  $L$  is the graph Laplacian defined above and  $\tau$  is the timestep of the MBO scheme, then the discrete energy

$$E_{MBO}(u) = \frac{1}{\tau} \langle 1 - u, \Gamma_\tau u \rangle$$

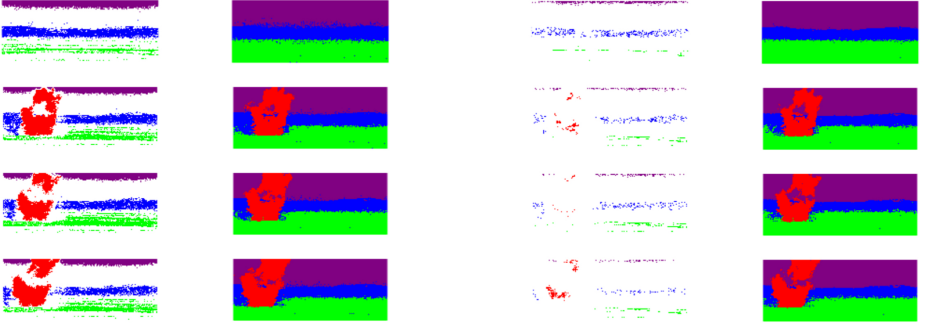


Figure 2: Feature vectors  $V_i$  are 128 dimensional hyperspectral pixels taken from a video sequence from standoff detection of gas plume in the Dugway Proving Ground. Shown are 4 of 7 video frames with  $N = 280,000$  graph nodes. The colors correspond to four classes: plume (red), sky (purple), foreground (green), mountain (blue). (left) 36% training data and the resulting classification - as in [Merkurjev, Sunu, and Bertozzi \[2014\]](#). (right) The same calculation with only 4% training data. In each case the training data is shown to the left of the fully segmented video. Code for this calculation is published online in [Meng, Merkurjev, Koniges, and Bertozzi \[2017\]](#)

decreases on each timestep and also approximates the graph TV energy.

**2.2 Volume Penalties.** In the case of unsupervised classification it is often desirable to have volume constraints. So rather than just minimizing the size of the graph cut, i.e.  $|u|_{TV}$ , one can put in a penalty that forces the size of the classes to be reasonably distributed. Two such normalizations are the ratio cut and the normalized cut, the problem is to find a subset  $S$  of the graph to minimize  $cut(S, \bar{S})R(S)$  where  $R$  is  $(1/|S| + 1/|\bar{S}|)$  for the ratio cut and  $(1/vol(S) + 1/vol(\bar{S}))$  for the normalized cut. Here the volume of the graph is the sum of the degrees of the all the vertices and the degree of the node is the sum of the weights connected to that node. Another normalization is the Cheeger cut in which  $R = (\min(|S|, |\bar{S}|))^{-1}$ . All three of these functionals are linear in the graph cut term and nonlinear in the volumetric constraints. The energy blows up as the size of  $S$  or  $\bar{S}$  goes to zero, thus ensuring a balance cut. There are several important papers related to clustering with volume penalties by Bresson, Szlam, Laurent, von Brecht. These works use other methods than the ones described above. In [Szlam and Bresson \[2010\]](#), study a relaxation of the Cheeger cut problem with connections between the energy of the relaxed problem and well studied energies in image processing. Authors of [Bresson, Laurent, Uminsky, and von Brecht \[2012\]](#) detail two procedures for the relaxed Cheeger cut problem. The

first algorithm is a steepest descent approach, and the second one is a modified inverse power method. In [Bresson, Laurent, Uminsky, and von Brecht \[2013\]](#), develop another version of the method shown in [Bresson, Laurent, Uminsky, and von Brecht \[2012\]](#) using a new adaptive stopping condition. The result is an algorithm that is monotonic and more efficient. The GL functional on graphs has been extended to these product-form volume penalized problems. The paper [Merkurjev, Bertozzi, Yan, and Lerman \[2017\]](#) uses a diffuse interface approach along with the graph Laplacian to solve the fully nonlinear ratio cut problem and the Cheeger cut problem. The results are shown to be very efficient with the efficiency partly achieved through the use of the Nyström extension method. The main idea is to approximate the cut term using the GL functional and then use PDE-based methods for gradient descent in a spectral approach. [Jacobs, Merkurjev, and Esedoglu \[2018\]](#) have a very efficient MBO-based method for solving the volume-constrained classification problem with different phases and with prescribed volume constraints and volume inequalities for the different phases. This work combines some of the best features of the MBO scheme in both Euclidean space and on graphs with a highly efficient algorithm of [Bertsekas \[1979\]](#) for the auction problem with volume constraints.

One of the challenges in machine learning is the case where the sizes of the classes are unknown. Volume constraints could perchance become incorporated as building blocks for solutions to complex data sorting problems, where the amount of data is so large that it becomes physically impossible for a human to verify all the results by inspection. An example of such large data currently under collection by law enforcement agencies around the world are video feeds from body worn cameras. The author and collaborators have been working with such a dataset provided by the Los Angeles Police Department and have developed classification methods based on the MBO scheme. The BW camera poses unusual challenges - typically the goal is to identify what is going on in the scene, both in terms of the wearer of the camera and his or her interaction with the scene. Thus the task requires understanding both the scene and the ego-motion, i.e. the motion of the individual to whom the camera is mounted. In [Meng, J. Sanchez, Morel, Bertozzi, and Brantingham \[2017\]](#), the authors develop an algorithm for the ego-motion classification, combining the MBO scheme for multi-class semi-supervised learning with an inverse compositional algorithm [Sánchez \[2016\]](#) to estimate transformations between successive frames. Thus the video is preprocessed to obtain an eight dimensional feature vector for each frame corresponding to the Left-Right; Up-Down; Rotational; and Forward-Backward motions of the camera wearer along with the frequencies of each of these motions. This is a gross reduction of the action of the video to a very low dimensional vector. These ideas have been extended by students at UCLA to higher dimensional feature vectors encoding both the egomotion and information from the scene [Akar, Chen, Dhillon, Song, and T. Zhou \[2017\]](#). Studies of the effect of class size are made possible by an extensive effort during a summer REU to hand classify sufficient video footage to provide ground truth for a

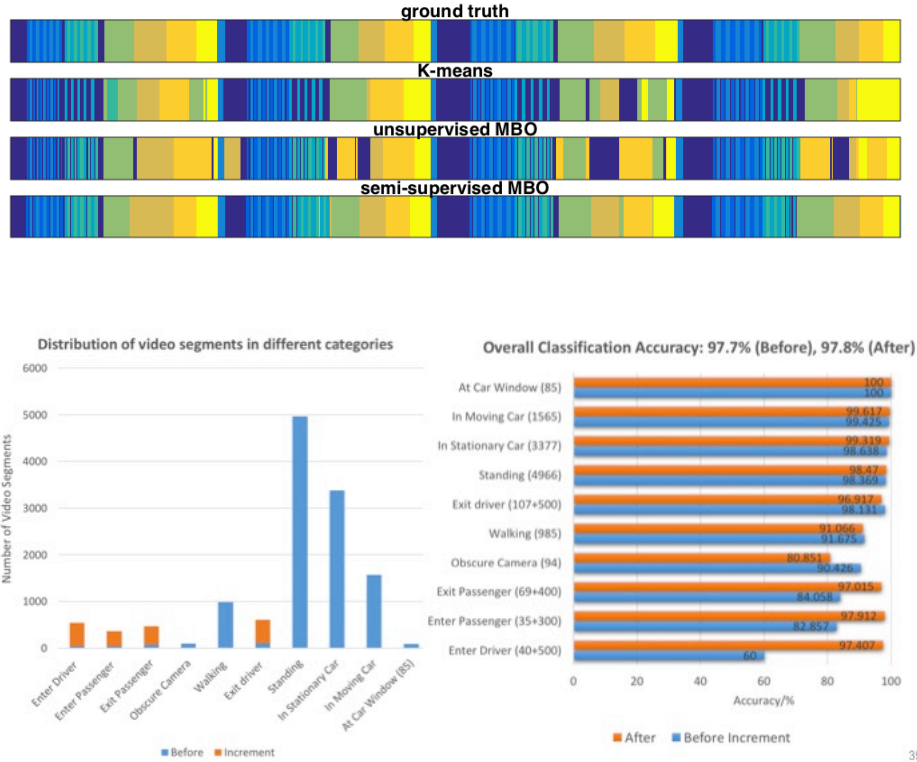


Figure 3: Top: Ego-motion classification results of the QUAD video [Meng, J. Sanchez, Morel, Bertozzi, and Brantingham \[2017\]](#). The 9 colors represent 9 different ego-motion classes: standing still (dark blue), turning left (moderate blue), turning right (light blue), looking up (dark green) and looking down (light green), jumping (bud green), stepping (aztec gold), walking (orange), runing (yellow). Copyright © 2018 Springer Nature. Published with permission of Springer Nature. Bottom: Semi-supervised MBO on LAPD Body Worn Cameras, using a complex motion-feature vector for each frame. Both examples use 10% of the data as training [Akar, Chen, Dhillon, Song, and T. Zhou \[2017\]](#).



larger study. Figure 3 shows results from Meng, J. Sanchez, Morel, Bertozzi, and Brantingham [2017] for the QUAD video in Baker and Matthews [2004] for various clustering algorithms described here and results on LAPD body worn camera data from Akar, Chen, Dhillion, Song, and T. Zhou [2017]. In the second example there are categories of activity with relatively small sample sizes and those are predominantly misclassified by this method. As a test, the smaller samples were augmented to an incremental level (shown in orange) by simply duplicating the video frame feature vectors, thereby increasing both the size of the classes and the size of the training data. The results show a marked increase in accuracy. Most research papers make use of scripted datasets with real-world data being relatively scarce for use in basic research (compared to its rate of capture in real-world applications). This case study shows the need for more algorithm development and theoretical results centered around the realities of real datasets. Privacy and proprietary reasons often hinder the use of such datasets in basic research and reproducibility can be hindered by the lack of public access to such data, nevertheless there is a strong societal need for more work to be done on real-world datasets and published in the scientific literature. Another important point related to this study is the fact that the data compression of entire video footage into low dimensional feature vectors (less than 100 dimensions per frame or group of frames) can serve as a tool for anonymizing sensitive data in order to develop computational algorithms on online computational platforms in shared workspaces, which can be forbidden to directly handle sensitive data. Such steps are imperative if one is to work with real-world data in an academic environment.

### 3 Uncertainty quantification (UQ) for graphical metrics

Semi-supervised learning combined both unlabeled data with labeled data; however, as the BWV example elucidates, in many applications there are so many unlabelled data points that one can not hand label everything. Moreover, there is a myriad of work in computer science and applied mathematics addressing the development of algorithms and a modest amount of work addressing performance of these methods in terms of convergence for problems such as clustering. For real-world applications, existing work does not address many obvious concerns - it is common to use methodologies ‘out of the box’ with measurements of performance of the methods based on ground truth information for toy/test problems but little information available regarding the likelihood of the results in general for real-world applications when ground truth is not available or when the existing ground truth is limited to a small percentage of training data. The graphical models and methods are particularly appropriate for the development of new mathematical methodology for uncertainty quantification, because of their organization around graph Laplacian matrices and nonlinear functionals that directly use these operators. In Blum and Chawla

[2001], using a graph min-cut problem for binary semi-supervised learning. This is equivalent to a maximum a posteriori (MAP) estimation on Bayesian posterior distribution for a Markov random field (MRF) over the discrete state space of binary labels [X. Zhu \[2005\]](#). Inference for multi-label discrete MRFs is typically intractable [Dahlhaus, Johnson, Papadimitriou, Seymour, and Yannakakis \[1992\]](#). Some approximate algorithms have been developed for the multi-label case [Y. Boykov, Veksler, and Zabih \[2001, 1998\]](#) and [Madry \[2010\]](#), with application to imaging tasks [Y. Y. Boykov and Jolly \[2001\]](#), [Berthod, Kato, Yu, and Zerubia \[1996\]](#), and [Li \[2012\]](#). In [X. Zhu, Ghahramani, Lafferty, et al. \[2003\]](#), relaxed the discrete state space to a continuous real-variable setting, and modeled the semi-supervised learning problem as a Gaussian random field. [D. Zhou, Bousquet, Lal, Weston, and Schölkopf \[2004\]](#) generalized the model to handle label noise, and also generalized it to the case of directed graphs [D. Zhou, Hofmann, and Schölkopf \[2004\]](#). We note that this earlier work of Zhou was a precursor to the nonlocal means graph developed by Buades Coll and Morel [Buades, Coll, and Morel \[2005\]](#) and further developed by Gilboa and Osher [Gilboa and Osher \[2007, 2008\]](#) that inspired some of the methods in the work of the Author and collaborators for the MBO scheme on graphs [Merkurjev, Kostic, and Bertozzi \[2013\]](#).

The probit classification method in [C. K. I. Williams and Rasmussen \[1996\]](#) uses the same prior as in [X. Zhu, Ghahramani, Lafferty, et al. \[2003\]](#) but the data takes on binary values, found from thresholding the underlying continuous variable, and thereby provides a link between the combinatorial and continuous state space approaches. The probit methodology is often implemented via MAP optimization – that is the posterior probability is maximized rather than sampled – or an approximation to the posterior is computed, in the neighborhood of the MAP estimator. In the context of MAP estimation, the graph-based terms act as a regularizer, in the form of the graph Dirichlet energy  $\frac{1}{2}\langle u, Lu \rangle$ , with  $L$  the symmetrized graph Laplacian. A formal framework for graph-based regularization can be found in [Belkin, Matveeva, and Niyogi \[2004\]](#) and [Belkin, Niyogi, and Sindhvani \[2006\]](#). More recently, other forms of regularization have been considered such as the graph wavelet regularization [Shuman, Faraji, and Vandergheynst \[2011\]](#) and [Hammond, Vandergheynst, and Gribonval \[2011\]](#).

The author and collaborators [Bertozzi, Luo, Stuart, and Zygalakis \[2017\]](#) have developed UQ methodologies for graph classification based on optimization over real-valued variables developed in the works discussed in [Bertozzi and Flenner \[2016\]](#). The UQ approach builds on the following ideas: (a) that a Bayesian formulation of the classification problem gives UQ automatically, (b) that fully Bayesian sampling is possible if one develops a methodology that scales well with respect to large graph size. The existing work on scalable algorithms for minimizing the graph GL functional is critical for (b). We have results for Gaussian noise models for binary classifiers that leverage several Bayesian models extended to classification on graphs; via the posterior distribution on the labels,

these methods automatically equip the classifications with measures of uncertainty. These models build on well-know Bayesian models in Euclidean space with fewer dimensions that arise in machine learning.

The probit classification [C. K. I. Williams and Rasmussen \[1996\]](#) nicely extends to graphical models. This involves defining a Gaussian measure through the graph Dirichlet energy and choosing a likelihood function involving thresholding a continuum latent variable plus noise. Bayes theorem provides a direct calculation of the posterior and its PDF. One can then compute the maximum a posteriori estimation (MAP) which is the minimizer of the negative of the log posterior, a convex function in the case of probit. The probit model may not be the most appropriate when the data is naturally segmented into groups, in which the variation is often understood to occur within the group. The next step is to build on several additional methods that have this structure; they are the level set method for Bayesian inverse problems [Iglesias, Lu, and Stuart \[2015\]](#), atomic noise models, and the Ginzburg-Landau optimization-based classifier [Bertozzi and Flenner \[2012\]](#) and [van Gennip and Bertozzi \[2012\]](#), which by virtue of its direct use of the Dirichlet energy, is tractable to generalize to a Bayesian setting. In all cases the posterior  $P(u|y)$  has the form

$$P(u|y) \propto \exp(-J(u)), \quad J(u) = \frac{1}{2c} \langle u, Lu \rangle + \Phi(u)$$

for some function  $\Phi$ , different for each of the four models - and for which the Ginzburg-Landau case, the independent variable is a real-valued relaxation of label space, rather than an underlying latent variable which may be thresholded by  $S(\cdot)$  into label space.) Here  $L$  is the graph Laplacian and  $c$  is a known scaling constant. The choice of scaling of  $L$  should be consistent with the scaling used for one of the learning methods (without UQ) discussed in the previous sections. Furthermore, the MAP estimator is the minimizer of  $J$ .  $\Phi$  is differentiable for the Ginzburg-Landau and probit models, but not for the level set and atomic noise models. We are interested in algorithms for both sampling and MAP estimation.

In [Bertozzi, Luo, Stuart, and Zygalakis \[2017\]](#) the authors develop efficient numerical methods, suited to large data-sets, for both MCMC-based sampling as well as gradient-based MAP estimation. In order to induce scalability with respect to size of the graph, we consider the pCN method described in [Cotter, G. O. Roberts, Stuart, and White \[2013\]](#) and introduced in the context of diffusions by Beskos in [Beskos, G. Roberts, Stuart, and Voss \[2008\]](#) and by Neal in the context of machine learning [Neal \[1998\]](#). The standard random walk Metropolis (RWM) algorithm suffers from the fact that the optimal proposal variance or stepsize scales inverse proportionally to the dimension of the state space [G. O. Roberts, Gelman, Gilks, et al. \[1997\]](#), which is the graph size  $N$  in this case. The pCN method was designed so that the proposal variance required to obtain a given acceptance

probability scales independently of the dimension of the state space, hence in practice giving faster convergence of the MCMC when compared with RWM. For graphs with a large number of nodes  $N$ , it is prohibitively costly to directly sample from the distribution  $\mu_0$ , since doing so involves knowledge of a complete eigen-decomposition of  $L$ . In machine learning classification tasks it is common to restrict the support of  $u$  to the eigenspace spanned by the first  $\ell$  eigenvectors with the smallest non-zero eigenvalues of  $L$  (hence largest precision) and this idea may be used to approximate the pCN method. The Author and collaborators have made use of both low rank [Fowlkes, Belongie, F. Chung, and Malik \[2004\]](#) approximations of nonsparse matrices and fast algorithms for computing the smallest non-zero eigenvalues of sparse matrices [Anderson \[2010\]](#). The upshot is a confidence score for the class assignment for binary classifiers, based on the node-wise posterior mean of the thresholded variable.

An example is shown in [Bertozzi, Luo, Stuart, and Zygalakis \[2017\]](#) with the MNIST database consists of 70,000 images of size  $28 \times 28$  pixels containing the handwritten digits 0 through 9; see [LeCun, Cortes, and Burges \[1998\]](#) for details. The nodes of the graph are the images and as feature vectors one uses the leading 50 principal components given by PCA; thus the feature vectors at each node have length  $d = 50$ . We construct a  $K$ -nearest neighbor graph with  $K = 20$  for each pair of digits considered. Namely, the weights  $a_{ij}$  are non-zero if and only if one of  $i$  or  $j$  is in the  $K$  nearest neighbors of the other. The non-zero weights are set using a local rescaling as in [Equation \(3\)](#). For more details see [Bertozzi, Luo, Stuart, and Zygalakis \[2017\]](#). The noise variance  $\gamma$  is set to 0.1, and 4% of fidelity points are chosen randomly from each class. The probit posterior is used to compute a node-wise posterior mean. [Figure 4](#) shows that nodes with scores posterior mean closer to the binary ground truth labels  $\pm 1$  look visually more uniform than nodes with score far from those labels. This illustrates that the posterior mean can differentiate between outliers and inliers that align with human perception.

There are a number of natural avenues to explore building on the work in [Bertozzi, Luo, Stuart, and Zygalakis \[ibid.\]](#); (a) there is a natural question of whether one works in label space, or a relaxation of it, as in GL, or with a latent variable as in probit - more investigation of the models on toy problems should elucidate this; (b) the models proposed above are rather simplistic and may not be best tuned to real datasets - it would be interesting to develop a preprocessing method to probe the data and to learn something about the data - preliminary results using probit as a preprocessing step for GL show some benefit to such hybrid methods; (c) these binary classification models will be extended to multiclass - the GL methodology is nicely extended in [Garcia-Cardona, Merkurjev, Bertozzi, Flenner, and Percus \[2014\]](#) but not in a Bayesian setting, whereas the other methods do not directly extend as easily although recursive methods can be helpful; (d) all of the methods described here involve a combination of graph Laplacian diffusion and thresholding analogous to the MBO scheme on graphs developed by the PI and collaborators [Merkurjev, Kostic, and](#)

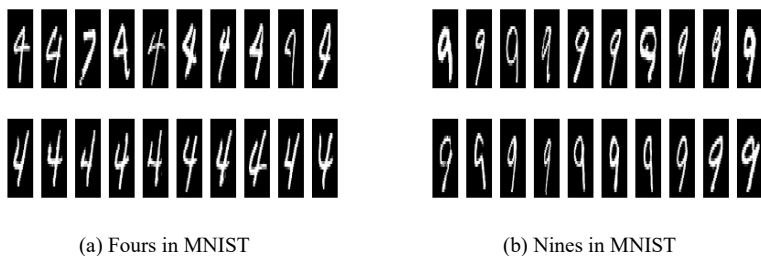


Figure 4: “Hard to classify” vs “easy to classify” nodes in the MNIST (4, 9) dataset under the probit model. Here the digit “4” is labeled +1 and “9” is labeled -1. The top (bottom) row of the left column corresponds to images that have the lowest (highest) values of the node-wise posterior mean out of all the “4” digit. The right column is organized in the same way for images with ground truth labels 9 except the top row now corresponds to the highest values of posterior mean. Higher posterior mean indicates higher confidence that the image is a 4 and not a “9”, hence the top row could be interpreted as images that are “hard to classify” by the current model, and vice versa for the bottom row. See [Bertozzi, Luo, Stuart, and Zygalakis \[2017\]](#) for more details.

[Bertozzi \[2013\]](#). Those algorithms also involve graph diffusion plus thresholding in a different way from the Bayesian statistical methods - and some measurement of similarity or difference should be made; (e) furthermore, one can consider unsupervised problems - for example the hybrid method in the paper [Hu, Sunu, and Bertozzi \[2015\]](#) that considers k-means plus the MBO scheme for clustering; (f) finally there are natural UQ questions that will arise from the other thrusts of the project. For example, for data fusion methods, the development of multimodal graphical models provides a natural context in which to extend the UQ methodology to these more complex data problems, providing not only insight into the results but also insight into the best choice of models for the fusion.

## 4 Network Analysis

The above discussion of uncertainty quantification was mainly directed at graphs that arise from machine learning problems involving similarity matrices that result from pairwise comparisons of high dimensional data. Another natural class of graphs are networks - for example social network graphs such as those that arise from social media, transportation networks, and other examples [M. E. J. Newman \[2010\]](#). Mathematical models and algorithms for structure in networks have led to a large body of work, for example in the

physics literature, that has largely happened independent of the work carried out in machine learning. There is a need to develop novel ideas in both areas and in some cases, especially with security applications, there is a need to have models that fit both machine learning (big data) problems and network problems.

There are several papers in the literature that connect network clustering to machine learning, and for brevity we mention a few methods here, including issues that arise when viewing network analysis methods in the context of machine learning: (a) in [Peel, Larremore, and Clauset \[2016\]](#) the authors consider metadata as ‘ground truth’ and prove a general “No Free Lunch” theorem for community detection, implying that no algorithm can perform better than others across all inputs; (b) Newman [M. E. J. Newman \[2013\]](#) considers spectral methods for three different problems - network modularity (discussed below), statistical inference, and normalized graph partitioning, concluding that algorithmically the spectral methods are the same for each class of problems; (c) Devooght et. al. consider random-walk based modularity applied to semi-supervised learning [Devooght, Mantrach, Kivimäki, Bersini, Jaimes, and Saerens \[2014\]](#) focusing on paths on the graph rather than edges. A review of clustering of graphs, including attributes (semi-supervision) from a network perspective is [Bothorel, Cruz, Magnani, and Micenkova \[2015\]](#). A recent review of community detection methods on networks can be found in [Fortunato and Hric \[2016\]](#).

A few years ago the Author and collaborators developed the first paper to directly connect network modularity optimization and total variation minimization on graphs, using the null model introduced by Newman and Girvan in [Girvan and M. E. J. Newman \[2004\]](#). To explain in more detail, the modularity of a network partition measures the fraction of total edge weight within communities versus what one might expect if edges were placed randomly according to some null model. More specifically, the objective is to **maximize** the modularity

$$Q = \frac{1}{2m} \sum_{ij} (w_{ij} - \gamma P_{ij}) \delta(g_i, g_j)$$

over all possible partitions where  $g_i$  is the group assignment for node  $i$ . Here  $P_{ij}$  is a probability null model (e.g.  $P_{ij} = k_i k_j / 2m$ ) where  $k_j = \sum_i w_{ij}$  and  $2m$  is the total volume of the graph ( $\sum_i k_i$ ) and  $\gamma$  is a resolution parameter. Our work [Hu, Laurent, Porter, and Bertozzi \[2013\]](#) shows that maximizing  $Q$  is equivalent to a graph cut problem that can be rewritten using the TV functional:

$$\text{Min}_{u: G \rightarrow V^{\hat{n}}} E(u) = |u|_{TV} - \gamma |u - m_2(u)|_{L_2}^2$$

for the case of  $\hat{n}$  classes where  $V^{\hat{n}}$  are the end nodes of the  $\hat{n}$ -dimensional simplex and  $m_2$  denotes a simple moment whose constraint can be introduced in a computationally tractable forcing term. Her  $u$  denotes the class assignment and takes vales on the corners

of the simplex. One can then use the above ideas to minimize this functional over all possible numbers of clusters  $\hat{n}$ . We note that our work also sheds new light on some of the other papers mentioned above. For example the TV-modularity connection is a direct relationship between the graph cuts and modularity, beyond the connection between spectral algorithms. Furthermore, the method used in [Hu, Laurent, Porter, and Bertozzi \[2013\]](#) builds on the graph heat equation, which is, roughly speaking, a mean field limit of a random walk dynamics. It uses directly the MBO scheme on graphs [Merkurjev, Kostic, and Bertozzi \[2013\]](#) from [Section 2.1](#) and multiclass methods for TV-minimization on graphs [Garcia-Cardona, Merkurjev, Bertozzi, Flenner, and Percus \[2014\]](#). The idea in these recent papers is to develop algorithms for graph clustering, in particular TV minimization, which is equivalent to graph cut minimization on weighted graphs when applied to partition functions. In the case of modularity optimization, the main idea is that maximizing the modularity functional, when applied to a fixed number of classes, is equivalent to minimization an energy for the assignment function, comprised of the graph total variation minus a second moment term. This opens the door to apply compressed sensing ideas to modularity optimization, a superior but computationally more complex method than spectral clustering. More can be done in this area and we propose to work on problems of direct relevance to multimodal graphs such as those that arise from composite information such as spatial nonlocal means, as in the example above, social networks, and latent information such as text-content topics from twitter.

The method is very scalable, which allows the algorithmic approach to go far beyond sparse network analysis, providing a new tool for analyzing large similarity graphs in machine learning. For example, the MNIST dataset [LeCun, Cortes, and Burges \[1998\]](#) of 70,000 handwritten digits, with tens of thousands of nodes, this approach is 10-100 times faster computationally than the GenLouvain algorithm [Jutla, Jeub, and Mucha \[n.d.\]](#) and produces comparable quality results, exceeding that of basic fast greedy algorithms such as [M. E. Newman \[2006\]](#) and [Blondel, Guillaume, Lambiotte, and Lefebvre \[2008b\]](#) and outperforming all other unsupervised clustering methods that we are aware of. What is most striking is the ability to correctly classify and identify the number of classes, in a fairly short amount of computational time. In general unsupervised clustering without prior knowledge of the number of classes is a very difficult problem for large datasets. So methodologies that are efficient enough to be useful for large data (including scalability) are needed. For example, in [Hu, van Gennip, Hunter, Bertozzi, and Porter \[2012\]](#) the GenLouvain code [Jutla, Jeub, and Mucha \[n.d.\]](#) was tested on the nonlocal means graph for a basic color image with excellent segmentation results for unsupervised clustering but with a run time that was neither practical nor scalable. Although for completeness one should compare with other methods such as the C++ implementations of Blondel of the Louvain method [Blondel, Guillaume, Lambiotte, and Lefebvre \[2008a\]](#). We note that even the soft clustering methods like Nonnegative Matrix Factorization and Latent

Dirichlet Allocation require the user to specify the number of classes and still have the restriction that they are built around a linear mixture model.

**4.1 Network analysis and machine learning.** An interesting line of inquiry is to unify the graphical models developed independently by the machine learning community and by the network science community for unsupervised learning. We believe that there is an opportunity to improve unsupervised learning algorithms (built on similarity graphs) for data science as well as to further understand the link between network structure and algorithm type. Starting from our earlier work on network modularity as a constrained multiclass graph cut problem, we address modularity as a constrained balanced cut problem in which convex methods can be used apart from the constraint. In a new work [Boyd, Bai, X. C. Tai, and Bertozzi \[2017\]](#) we have identified four different equivalent formulations of the modularity problem which we term soft balanced cut, penalized balanced cut, balanced total variation (TV) minimization, and penalized TV minimization.

**Theorem 1** (Equivalent forms of modularity [Boyd, Bai, X. C. Tai, and Bertozzi \[ibid.\]](#)). *For any subset  $S$  of the nodes of  $G$ , define  $\text{vol } S = \sum_{i \in S} k_i$ . Then the following optimization problems are all equivalent:*

(8)

$$\text{Std. form:} \quad \underset{\hat{n} \in \mathbb{N}, \{A_\ell\}_{\ell=1}^{\hat{n}} \in \Pi(G)}{\operatorname{argmax}} \quad \sum_{\ell=1}^{\hat{n}} \sum_{ij \in A_\ell} w_{ij} - \gamma \frac{k_i k_j}{2m},$$

(9)

$$\text{Bal. cut (I):} \quad \underset{\hat{n} \in \mathbb{N}, \{A_\ell\}_{\ell=1}^{\hat{n}} \in \Pi(G)}{\operatorname{argmin}} \quad \sum_{\ell=1}^{\hat{n}} \left( \text{Cut}(A_\ell, A_\ell^c) + \frac{\gamma}{2m} (\text{vol } A_\ell)^2 \right),$$

(10)

$$\text{Bal. cut (II):} \quad \underset{\hat{n} \in \mathbb{N}, \{A_\ell\}_{\ell=1}^{\hat{n}} \in \Pi(G)}{\operatorname{argmin}} \quad \sum_{\ell=1}^{\hat{n}} \left( \text{Cut}(A_\ell, A_\ell^c) + \frac{\gamma}{2m} \left( \text{vol } A_\ell - \frac{2m}{\hat{n}} \right)^2 \right) + \gamma \frac{2m}{\hat{n}}$$

(11)

$$\text{Bal. TV (I):} \quad \underset{\hat{n} \in \mathbb{N}, u \in \Pi(G)}{\operatorname{argmin}} \quad |u|_{TV} + \frac{\gamma}{2m} \left\| k^T u \right\|_2^2$$

(12)

$$\text{Bal. TV (II):} \quad \underset{\hat{n} \in \mathbb{N}, u \in \Pi(G)}{\operatorname{argmin}} \quad |u|_{TV} + \frac{\gamma}{2m} \left\| k^T u - \frac{2m}{\hat{n}} \right\|_2^2 + \gamma \frac{2m}{\hat{n}}.$$



Each of the preceding forms has a different interpretation. The original formulation of modularity was based on comparison with a statistical model and views communities as regions that are more connected than they would be if edges were totally random. The cut formulations represent modularity as favoring sparsely interconnected regions with balanced volumes, and the TV formulation seeks a piecewise-constant partition function  $u$  whose discontinuities have small perimeter, together with a balance-inducing quadratic penalty. The cut and TV forms come in pairs. The first form (labelled “I”) is simpler to write but harder to interpret, while the second (labelled “II”) has more terms, but the nature of the balance term is easy to understand, as it is minimized (for fixed  $\hat{n}$ ) when each community has volume  $\frac{2m}{\hat{n}}$ .

In addition to providing a new perspective on the modularity problem in general, this equivalence shows that modularity optimization can be viewed as minimizing a convex functional but subject to a binary constraint. These methodologies provide a direct connection between modularity and other balance cut problems such as the Cheeger or Ratio cut and a connection to convex optimization methods already developed for semi-supervised learning on graphs [Merkurjev, Bae, Bertozzi, and X.-C. Tai \[2015\]](#) and [Bae and Merkurjev \[2016\]](#). A significant emphasis on spectral algorithms exists in the literature on graph cut methods for networks, see e.g. [M. E. Newman \[2006\]](#) for spectral methods for modularity vs. other spectral methods applied to networks, and a large literature on accuracy of spectral approximations for the Cheeger cut (e.g. [Ghosh, Teng, Lerman, and Yan \[2014\]](#)). What distinguishes our approach from other efforts is the focus on non-network data using a network approach. There are many reasons to do this. At the forefront is the ability to do unsupervised clustering well, without knowing the number of clusters.

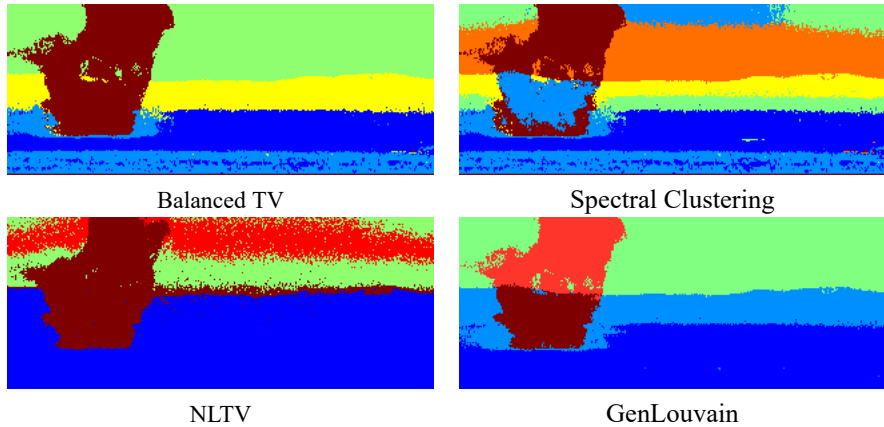


Figure 5: Segmentations [Boyd, Bai, X. C. Tai, and Bertozzi \[2017\]](#) of the plume hyperspectral video using different methods. The Balanced TV is the only method that has the whole plume into a single class without any extraneous pixels (NLTV method from [W. Zhu, Chayes, Tiard, S. Sanchez, Dahlberg, Bertozzi, Osher, Zosso, and Kuang \[2017\]](#)).

The ideas developed in [Boyd, Bai, X. C. Tai, and Bertozzi \[2017\]](#) show that, while modularity optimization is inherently nonconvex, that working with it as a constrained convex optimization problem produces results that are noticeably improved compared to prior methods that do not use such a formulation, including the formulation in [Hu, Laurent, Porter, and Bertozzi \[2013\]](#). Another relevant recent work is [Merkurjev, Bae, Bertozzi, and X.-C. Tai \[2015\]](#) that develops convex optimization methods to find global minimizers of graph cut problems for semi-supervised learning. This work is loosely related to [Boyd, Bai, X. C. Tai, and Bertozzi \[2017\]](#) and serves as a resource for the use of L1 compressed sensing methods and max flow methods for constrained cut problems. Regarding benchmark testing, we note that [Bazzi, Jeub, Arenas, Howison, and Porter \[2016\]](#) has developed a new class of benchmark networks that can be tested with algorithms in addition to the LFR benchmarks.

**4.2 Data fusion, multilayer graphs and networks.** There are many works in the literature for data fusion that do not use graphs - they require a specific connection between the information and are typically not flexible to extend to unrelated data fusion problems. One such example pan sharpening of remote sensing images in which a panchromatic sensor has higher spatial resolution than a multispectral sensor [Möller, Wittman, Bertozzi, and Burger \[2012\]](#). Graphical methods for data fusion are still in their infancy with a few ideas

in the literature but very little theoretical understanding of these approaches. Some examples from the works of the Author include (a) a homotopy parameter for social network vs. spatial embedding used to study Field Interview cards from Los Angeles [van Gennip, Hunter, et al. \[2013\]](#) (Figure 1 in this paper), (b) a variational model for MPLE for statistical density estimation of crime data [Woodworth, Mohler, Bertozzi, and Brantingham \[2014\]](#) that uses a nonlocal means-based graphical model for high spatial resolution housing density information as a regularization parameter, and (c) a threshold-based similarity graph for combined LIDAR and multispectral sensors [Iyer, Chanussot, and Bertozzi \[2017\]](#). The network science community has different methods for fusing network data compared to traditional methods used in sensor fusion. One might explore the similarities and differences between the network science models and the sensor fusion models and to examine and identify opportunities to bring ideas from one community into the other through the use of graphical models, along with related rigorous analytical results of relevance.

Another problem is to develop algorithms based on models for more complex networks - for example multi-layer modularity optimization as proposed by Porter and colleagues [Mucha, Richardson, Macon, Porter, and Onnela \[2010\]](#) (Science 2010) and more recent papers that build on that work (e.g. [Bazzi, Porter, S. Williams, McDonald, Fenn, and Howison \[2016\]](#) and [M. E. J. Newman and Peixoto \[2015\]](#)). The multilayer approach can give much better granularity of clustering in social networks however it is even more computationally prohibitive than regular modularity in the case of larger datasets (e.g. tens of thousands of nodes). Multilayer models are able to work with more complex similarity graphs, such as those that might arise from multimodal data, although little work has been done unifying these ideas. As an example, for the LAPD field interview cards studied in [van Gennip, Hunter, et al. \[2013\]](#), one might analyze what additional information might be encoded in a multilayer network structure compared to a parametric homotopy model on a single layer graph. For multilayer graphs, we expect TV minimization methods to handle structures within a layer, however different methods may be required when strong connections arise across layers. We expect that different issues may arise when considering such graphs for data fusion rather than complex network applications.

For multilayer graph models one could explore hybrid schemes that leverage the ultrafast segmentation that can be done for large clusters using something like MBO while using the combinatorial methods (e.g. Gen Louvain [Jutla, Jeub, and Mucha \[n.d.\]](#)) for the network structure that has some granularity. This is a main challenge when working with complex data such as the artificial LFR benchmarks [Lancichinetti, Fortunato, and Radicchi \[2008\]](#) that have a power law community distribution. One can also compare against the new benchmark graphs in [Bazzi, Jeub, Arenas, Howison, and Porter \[2016\]](#) using their code. Future work might involve a hybrid method that will have components of TV minimization methods such as the MBO scheme [Merkurjev, Kostic, and Bertozzi \[2013\]](#),

components of GenLouvain and possible post-processing steps such as Kernighan-Lin node-swapping [M. E. Newman \[2006\]](#), [Porter, Onnela, and Mucha \[2009\]](#), and [Richardson, Mucha, and Porter \[2009\]](#).

## 5 Final Comments

The Author and collaborators have developed rigorous analysis for the dynamics of the graphical MBO iteration scheme for semi-supervised learning [van Gennip, Guillen, Osting, and Bertozzi \[2014\]](#) and this work could be extended to unsupervised, multiclass, classification methods such as those that arise in network modularity. The Author and Luo have developed theoretical convergence estimates [Luo and Bertozzi \[2017\]](#) for the Ginzburg-Landau convex-splitting method for semi-supervised learning for various versions of the graph Laplacian discussed above. For example, for the standard graph Laplacian we have maximum norm convergence results for minimizers of the Ginzburg-Landau energy using a combination of  $L^2$ -energy estimates and maximum principle results for the Laplacian operator [Luo and Bertozzi \[ibid.\]](#). The GL energy is a non-convex functional, so those results prove convergence to a local minimizer rather than a global one and can require modest *a posteriori* estimates to guarantee convergence; these are ones that can be built directly into the code. One of the rigorous results proved in [Luo and Bertozzi \[ibid.\]](#) is that the *convergence and stability of the scheme are independent of the size of the graph, and of its sparseness*, an important feature for scalability of methods.

Another issue that is rarely discussed for either the semi-supervised or unsupervised cases, regarding similarity graphs, is whether to thin the graph before performing classification or to use the fully connected graph in connection with a low rank approximation of the matrix such as the Nyström extension, discussed above. Research is needed to develop rigorous estimates related to the thinning of the graph in conjunction with models for clustering data - for example we can take examples models built on the Gaussian priors in the previous section on UQ and develop estimates for what is lost from the matrix when removing edges with smaller weights, a common process using e.g. a k-nearest neighbor graph. This problem involves the role of the graph structure on optimization problems and can also benefit from existing results from the network literature.

**Acknowledgments.** The author thanks Z. Meng for help with [Figure 2](#) and Mason Porter for useful comments. This work discusses a body of research that would not have been possible without the perserverence and insight of many students and postdocs including Egil Bae, Zach Boyd, Julia Dobrosotskaya, Cristina Garcia-Cardona, Nestor Guillen, Huiyi Hu, Blake Hunter, Geoffrey Iyer, Tijana Kostic, Hao Li, Xiyang Luo, Zhaoyi Meng, Ekaterina

Merkurjev, Braxton Osting, Carola Schönlieb, Justin Sunu, and Yves van Gennip. Numerous ideas came from interactions with many colleagues including Chris Anderson, P. Jeffrey Brantingham, Tony Chan, Jocelyn Chanussot, Fan Chung, Arjuna Flenner, Thomas Laurent, Kristina Lerman, Stanley Osher, J. M. Morel, Mason Porter, George Tita, Andrew Stuart and Xue-Cheng Tai, and Luminita Vese.

## References

- O. Akar, H. Chen, A. Dhillon, A. Song, and T. Zhou (2017). “Body Worn Video”. Technical report from 2017 summer REU on classification of BWV from LAPD; Andrea L. Bertozzi, M. Haberland, H. Li, P. J. Brantingham, and M. Roper faculty mentors (cit. on pp. 3892–3894).
- Christopher R. Anderson (Sept. 2010). “[A Rayleigh-Chebyshev Procedure for Finding the Smallest Eigenvalues and Associated Eigenvectors of Large Sparse Hermitian Matrices](#)”. *J. Comput. Phys.* 229.19, pp. 7477–7487 (cit. on p. 3897).
- E. Bae and E. Merkurjev (2016). “Convex Variational Methods for Multiclass Data Segmentation on Graphs” (cit. on p. 3902).
- Simon Baker and Iain Matthews (Feb. 2004). “[Lucas-Kanade 20 Years On: A Unifying Framework](#)”. *International Journal of Computer Vision* 56.3, pp. 221–255 (cit. on p. 3894).
- M. Bazzi, L. G. S. Jeub, A. Arenas, S. D. Howison, and M. A. Porter (2016). “Generative benchmark models for mesoscale structure in multilayer networks” (cit. on pp. 3903, 3904).
- M. Bazzi, M. A. Porter, S. Williams, M. McDonald, D. J. Fenn, and S. D. Howison (2016). “Community Detection in Temporal Multilayer Networks, with an Application to Correlation Networks”. *Multiscale Model. Simul.* 14.1, pp. 1–41 (cit. on p. 3904).
- Mikhail Belkin, Irina Matveeva, and Partha Niyogi (2004). “Regularization and semi-supervised learning on large graphs”. In: *International Conference on Computational Learning Theory*. Springer, pp. 624–638 (cit. on p. 3895).
- Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani (2006). “Manifold regularization: A geometric framework for learning from labeled and unlabeled examples”. *Journal of machine learning research* 7.Nov, pp. 2399–2434 (cit. on p. 3895).
- Serge Belongie, Charless Fowlkes, Fan Chung, and Jitendra Malik (2002). “Spectral partitioning with indefinite kernels using the Nyström extension”. In: *European Conference on Computer Vision*, pp. 531–542 (cit. on p. 3888).
- Marc Berthod, Zoltan Kato, Shan Yu, and Josiane Zerubia (1996). “Bayesian image classification using Markov random fields”. *Image and vision computing* 14.4, pp. 285–295 (cit. on p. 3895).

- Andrea L. Bertozzi and Arjuna Flenner (2012). “Diffuse interface models on graphs for classification of high dimensional data”. *Multiscale Modeling & Simulation* 10.3, pp. 1090–1118 (cit. on pp. [3886](#), [3896](#)).
- (2016). “Diffuse Interface Models on Graphs for Classification of High Dimensional Data”. *SIAM Review* 58.2, pp. 293–328 (cit. on pp. [3886](#), [3889](#), [3895](#)).
- Andrea L. Bertozzi, X. Luo, A. M. Stuart, and K. C. Zygalakis (2017). “Uncertainty Quantification in the Classification of High Dimensional Data” (cit. on pp. [3895](#)–[3898](#)).
- Dimitri Bertsekas (1979). “A Distributed Algorithm for the Assignment Problem”. Technical report, MIT (cit. on p. [3892](#)).
- A. Beskos, G. Roberts, A. M. Stuart, and J. Voss (2008). “MCMC methods for diffusion bridges”. *Stochastics and Dynamics* 8.03, pp. 319–350 (cit. on p. [3896](#)).
- V. D. Blondel, J.-L. Guillaume, R. Lambiotte, and E. Lefebvre (2008a). *Louvain Method: Finding Communities in Large Networks* (cit. on p. [3900](#)).
- Vincent D Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre (2008b). “Fast unfolding of communities in large networks”. *J. Stat. Mech. Theory Exp.* 2008.10, P10008 (cit. on p. [3900](#)).
- Avrim Blum and Shuchi Chawla (2001). “Learning from labeled and unlabeled data using graph mincuts”. *Proc. 18th Int. Conf. Mach. Learning (ICML)* (cit. on p. [3894](#)).
- J. Bosch, S. Klamt, and M. Stoll (2016). “Generalizing diffuse interface methods on graphs: non-smooth potentials and hypergraphs” (cit. on p. [3887](#)).
- Cécile Bothorel, Juan David Cruz, Matteo Magnani, and Barbora Micenkova (2015). “Clustering attributed graphs: models, measures and methods”. *Network Science* 3.3, pp. 408–444 (cit. on p. [3899](#)).
- Z. Boyd, E. Bai, X. C. Tai, and Andrea L. Bertozzi (2017). “Simplified energy landscape for modularity using total variation” (cit. on pp. [3887](#), [3901](#), [3903](#)).
- Yuri Y Boykov and M-P Jolly (2001). “Interactive graph cuts for optimal boundary & region segmentation of objects in ND images”. In: *Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*. Vol. 1. IEEE, pp. 105–112 (cit. on p. [3895](#)).
- Yuri Boykov, Olga Veksler, and Ramin Zabih (1998). “Markov random fields with efficient approximations”. In: *Computer vision and pattern recognition, 1998. Proceedings. 1998 IEEE computer society conference on*. IEEE, pp. 648–655 (cit. on p. [3895](#)).
- (2001). “Fast approximate energy minimization via graph cuts”. *IEEE Transactions on pattern analysis and machine intelligence* 23.11, pp. 1222–1239 (cit. on p. [3895](#)).
- Xavier Bresson, Thomas Laurent, David Uminsky, and James H. von Brecht (2012). “Convergence and Energy Landscape for Cheeger Cut Clustering”. In: *Advances in Neural Information Processing Systems* 25. Ed. by F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger. Curran Associates, Inc., pp. 1385–1393 (cit. on pp. [3891](#), [3892](#)).

- Xavier Bresson, Thomas Laurent, David Uminsky, and James H. von Brecht (2013). “An Adaptive Total Variation Algorithm for Computing the Balanced Cut of a Graph” (cit. on p. 3892).
- Antoni Buades, Bartomeu Coll, and J-M Morel (2005). “A non-local algorithm for image denoising”. In: *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR '05)*. Vol. 2. IEEE, pp. 60–65 (cit. on pp. 3884, 3888, 3895).
- Luca Calatroni, Yves van Gennip, Carola-Bibiane Schönlieb, Hannah M. Rowland, and Arjuna Flenner (Feb. 2017). “Graph Clustering, Variational Image Segmentation Methods and Hough Transform Scale Detection for Object Measurement in Images”. *Journal of Mathematical Imaging and Vision* 57.2, pp. 269–291 (cit. on p. 3887).
- T. Chan and L. A. Vese (2001). “Active Contours without Edges”. *IEEE Trans. Image Process.* 10, pp. 266–277 (cit. on p. 3888).
- Fan R. K. Chung (1996). *Spectral Graph Theory*. American Mathematical Society (cit. on p. 3885).
- S. L. Cotter, G. O. Roberts, A. M. Stuart, and D. White (2013). “MCMC methods for functions: modifying old algorithms to make them faster.” *Statistical Science* 28.3, pp. 424–446 (cit. on p. 3896).
- Elias Dahlhaus, David S Johnson, Christos H Papadimitriou, Paul D Seymour, and Mihalis Yannakakis (1992). “The complexity of multiway cuts”. In: *Proceedings of the twenty-fourth annual ACM symposium on Theory of computing*. ACM, pp. 241–251 (cit. on p. 3895).
- Robin Devooght, Amin Mantrach, Ilkka Kivimäki, Hugues Bersini, Alejandro Jaimes, and Marco Saerens (2014). “Random Walks Based Modularity: Application to Semi-supervised Learning”. In: *Proceedings of the 23rd International Conference on World Wide Web*. WWW '14. New York, NY, USA: ACM, pp. 213–224 (cit. on p. 3899).
- Julia A. Dobrosotskaya and Andrea L. Bertozzi (2008). “A Wavelet-Laplace Variational Technique for Image Deconvolution and Inpainting”. *IEEE Trans. Imag. Proc.* 17.5, pp. 657–663 (cit. on p. 3887).
- (2010). “Wavelet analogue of the Ginzburg-Landau energy and its  $\Gamma$ –convergence”. *Int. Free Boundaries* 12.4, pp. 497–525 (cit. on p. 3887).
- Selim Esedoğlu and Felix Otto (2015). “Threshold Dynamics for Networks with Arbitrary Surface Tensions”. *Communications on Pure and Applied Mathematics* 68.5, pp. 808–864 (cit. on p. 3890).
- Selim Esedoğlu and Yen-Hsi Richard Tsai (2006). “Threshold dynamics for the piecewise constant Mumford-Shah functional”. *Journal of Computational Physics* 211.1, pp. 367–384 (cit. on pp. 3888, 3889).
- Santo Fortunato and Darko Hric (2016). “Community detection in networks: A user guide”. *Physics Reports* 659. Community detection in networks: A user guide, pp. 1–44 (cit. on p. 3899).

- Charless Fowlkes, Serge Belongie, Fan Chung, and Jitendra Malik (2004). “Spectral grouping using the Nyström method”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 26.2, pp. 214–225 (cit. on pp. 3888, 3897).
- Charless Fowlkes, Serge Belongie, and Jitendra Malik (2001). “Efficient spatiotemporal grouping using the Nyström method”. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. Vol. 1, pp. 1–231 (cit. on p. 3888).
- Cristina Garcia-Cardona, Ekaterina Merkurjev, Andrea L. Bertozzi, Arjuna Flenner, and Allon G Percus (2014). “Multiclass data segmentation using diffuse interface methods on graphs”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36.8, pp. 1600–1613 (cit. on pp. 3887, 3890, 3897, 3900).
- Yves van Gennip and Andrea L. Bertozzi (2012). “T-convergence of graph Ginzburg-Landau functionals”. *Advances in Differential Equations* 17.11–12, pp. 1115–1180 (cit. on pp. 3887, 3896).
- Yves van Gennip, Nestor Guillen, Braxton Osting, and Andrea L. Bertozzi (2014). “Mean curvature, threshold dynamics, and phase field theory on finite graphs”. *Milan J. of Math* 82.1, pp. 3–65 (cit. on pp. 3890, 3905).
- Yves van Gennip, Blake Hunter, et al. (2013). “Community detection using spectral clustering on sparse geosocial data”. *SIAM J. Appl. Math.* 73.1, pp. 67–83 (cit. on pp. 3885, 3886, 3904).
- R. Ghosh, S.-H. Teng, K. Lerman, and X. Yan (2014). “The interplay between dynamics and networks: centrality, communities, and Cheeger inequality”. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (cit. on p. 3902).
- Guy Gilboa and Stanley Osher (2007). “Nonlocal linear image regularization and supervised segmentation”. *Multiscale Modeling & Simulation* 6.2, pp. 595–630 (cit. on pp. 3888, 3895).
- (2008). “Nonlocal operators with applications to image processing”. *Multiscale Modeling & Simulation* 7.3, pp. 1005–1028 (cit. on pp. 3888, 3895).
- M. Girvan and M. E. J. Newman (2004). “Finding and evaluating community structure in networks”. *Phys. Rev. E*. 69 (cit. on p. 3899).
- Tom Goldstein and Stanley Osher (2009). “The split Bregman method for L1-regularized problems”. *SIAM Journal on Imaging Sciences* 2.2, pp. 323–343 (cit. on p. 3888).
- David K Hammond, Pierre Vandergheynst, and Rémi Gribonval (2011). “Wavelets on graphs via spectral graph theory”. *Applied and Computational Harmonic Analysis* 30.2, pp. 129–150 (cit. on p. 3895).
- H. Hu, Y. van Gennip, B. Hunter, Andrea L. Bertozzi, and M. A. Porter (2012). “Multislice Modularity Optimization in Community Detection and Image Segmentation”. *Proc. IEEE International Conference on Data Mining (Brussels), ICDM’12*, pp. 934–936 (cit. on p. 3900).



- Huiyi Hu, Thomas Laurent, Mason A. Porter, and Andrea L. Bertozzi (2013). “A Method Based on Total Variation for Network Modularity Optimization using the MBO Scheme”. *SIAM J. Appl. Math.* 73.6, pp. 2224–2246 (cit. on pp. [3887](#), [3899](#), [3900](#), [3903](#)).
- Huiyi Hu, Justin Sunu, and Andrea L. Bertozzi (2015). “Multi-class Graph Mumford-Shah Model for Plume Detection Using the MBO scheme”. In: *Energy Minimization Methods in Computer Vision and Pattern Recognition*. Springer, pp. 209–222 (cit. on pp. [3887](#), [3889](#), [3890](#), [3898](#)).
- Marco A Iglesias, Yulong Lu, and Andrew M Stuart (2015). “A Bayesian Level Set Method for Geometric Inverse Problems”. arXiv: [1504.00313](#) (cit. on p. [3896](#)).
- G. Iyer, J. Chanussot, and Andrea L. Bertozzi (2017). “A graph-based approach for feature extraction and segmentation in multimodal images”. *Proc. Int. Conf. Image Proc., Beijing*, pp. 3320–3324 (cit. on pp. [3887](#), [3904](#)).
- Matt Jacobs, Ekaterina Merkurjev, and Selim Esedoğlu (2018). “Auction dynamics: A volume constrained MBO scheme”. *Journal of Computational Physics* 354. Supplement C, pp. 288–310 (cit. on p. [3892](#)).
- Inderjit S. Jutla, Lucas G. S. Jeub, and Peter J. Mucha (n.d.). *A generalized Louvain method for community detectio implemented in MATLAB* (cit. on pp. [3900](#), [3904](#)).
- R. V. Kohn and P. Sternberg (1989). “Local minimisers and singular perturbations”. *Proc. Roy. Soc. Edinburgh Sect. A* 111, pp. 69–84 (cit. on p. [3887](#)).
- A. Lancichinetti, S. Fortunato, and F. Radicchi (2008). “Benchmark graphs for testing community detection algorithms”. *Phys. Rev. E* 78.04, p. 046110 (cit. on p. [3904](#)).
- Yann LeCun, Corinna Cortes, and Christopher JC Burges (1998). *The MNIST database of handwritten digits* (cit. on pp. [3897](#), [3900](#)).
- Stan Z Li (2012). *Markov random field modeling in computer vision*. Springer Science & Business Media (cit. on p. [3895](#)).
- X. Luo and Andrea L. Bertozzi (2017). “Convergence Analysis of the Graph Allen-Cahn Scheme”. *J. Stat. Phys.* 167.3, pp. 934–958 (cit. on p. [3905](#)).
- Ulrike von Luxburg (2007). “A tutorial on spectral clustering”. *Statistics and computing* 17.4, pp. 395–416 (cit. on pp. [3884](#), [3885](#)).
- Aleksander Madry (2010). “Fast approximation algorithms for cut-based problems in undirected graphs”. In: *Foundations of Computer Science (FOCS), 2010 51st Annual IEEE Symposium on*. IEEE, pp. 245–254 (cit. on p. [3895](#)).
- Z. Meng, E. Merkurjev, A. Koniges, and Andrea L. Bertozzi (2017). “Hyperspectral Image Classification Using Graph Clustering Methods”. *Image Processing Online (IPOL)* 7. published with code, pp. 218–245 (cit. on pp. [3887](#), [3889](#), [3891](#)).
- Zhaoyi Meng, Alice Koniges, Yun (Helen) He, Samuel Williams, Thorsten Kurth, Brandon Cook, Jack Deslippe, and Andrea L. Bertozzi (2016). “OpenMP Parallelization and Optimization of Graph-based Machine Learning Algorithm”. *Proc. 12th International Workshop on OpenMP (IWOMP)* (cit. on pp. [3887](#), [3890](#)).

- Zhaoyi Meng, Javier Sanchez, Jean-Michel Morel, Andrea L. Bertozzi, and P. Jeffrey Brantingham (2017). “Ego-motion Classification for Body-worn Videos”. accepted in the Proceedings of the 2016 Conference on Imaging, vision and learning based on optimization and PDEs, Bergen, Norway (cit. on pp. [3892–3894](#)).
- E. Merkurjev, E. Bae, Andrea L. Bertozzi, and X.-C. Tai (2015). “Global binary optimization on graphs for classification of high dimensional data”. *J. Math. Imag. Vis.* 52.3, pp. 414–435 (cit. on pp. [3890](#), [3902](#), [3903](#)).
- E. Merkurjev, Andrea L. Bertozzi, and F. Chung (2016). “A semi-supervised heat kernel pagerank MBO algorithm for data classification” (cit. on p. [3887](#)).
- E. Merkurjev, J. Sunu, and Andrea L. Bertozzi (2014). “Graph MBO method for multiclass segmentation of hyperspectral stand-off detection video”. In: *Proc. Int. Conf. Image Proc. (ICIP) Paris*. IEEE, pp. 689–693 (cit. on pp. [3887](#), [3888](#), [3891](#)).
- Ekaterina Merkurjev, Andrea L. Bertozzi, Xiaoran Yan, and Kristina Lerman (2017). “[Modified Cheeger and ratio cut methods using the Ginzburg-Landau functional for classification of high-dimensional data](#)”. *Inverse Problems* 33.7, p. 074003 (cit. on pp. [3887](#), [3892](#)).
- Ekaterina Merkurjev, Tijana Kostic, and Andrea L. Bertozzi (2013). “An MBO scheme on graphs for classification and image processing”. *SIAM Journal on Imaging Sciences* 6.4, pp. 1903–1930 (cit. on pp. [3887](#), [3889](#), [3890](#), [3895](#), [3897](#), [3900](#), [3904](#)).
- B. Merriman, J. Bence, and S. Osher (1992). “Diffusion generated motion by mean curvature”. *Proc. Comput. Crystal Growers Workshop*, pp. 73–83 (cit. on p. [3889](#)).
- Michael Möller, Todd Wittman, Andrea L. Bertozzi, and Martin Burger (2012). “[A Variational Approach for Sharpening High Dimensional Images](#)”. *SIAM Journal on Imaging Sciences* 5.1, pp. 150–178 (cit. on p. [3903](#)).
- Peter J Mucha, Thomas Richardson, Kevin Macon, Mason A Porter, and Jukka-Pekka Onnela (2010). “[Community structure in time-dependent, multiscale, and multiplex networks](#)”. *Science* 328.5980, pp. 876–878 (cit. on p. [3904](#)).
- R. Neal (1998). “[Regression and classification using Gaussian process priors](#)”. *Bayesian Statistics* 6, p. 475 (cit. on p. [3896](#)).
- M. E. J. Newman (Oct. 2013). “[Spectral methods for community detection and graph partitioning](#)”. *Phys. Rev. E* 88, p. 042822 (cit. on p. [3899](#)).
- Mark E J Newman (2010). *Networks: An Introduction*. Oxford, UK: Oxford University Press (cit. on p. [3898](#)).
- Mark E J Newman and Tiago P Peixoto (Aug. 2015). “[Generalized Communities in Networks](#)”. English. *Phys. Rev. Lett.* 115.8, p. 088701 (cit. on p. [3904](#)).
- Mark EJ Newman (2006). “Modularity and community structure in networks”. *Proc. Nat. Acad. Sci.* 103.23, pp. 8577–8582 (cit. on pp. [3900](#), [3902](#), [3905](#)).
- L. Peel, D. B. Larremore, and A. Clauset (2016). “The ground truth about metadata and community detection in networks” (cit. on p. [3899](#)).

- M. A. Porter, J.-P. Onnela, and P. J. Mucha (2009). “Communities in networks”. *Notices Amer. Math. Soc.* 56.9, pp. 1082–1097, 1164–1166 (cit. on p. 3905).
- Thomas Richardson, Peter J. Mucha, and Mason A. Porter (Sept. 2009). “Spectral tripartitioning of networks”. *Phys. Rev. E* 80, p. 036111 (cit. on p. 3905).
- Gareth O Roberts, Andrew Gelman, Walter R Gilks, et al. (1997). “Weak convergence and optimal scaling of random walk Metropolis algorithms”. *The annals of applied probability* 7.1, pp. 110–120 (cit. on p. 3896).
- J. Sánchez (2016). “The Inverse Compositional Algorithm for Parametric Registration”. *Image Processing On Line*, pp. 212–232 (cit. on p. 3892).
- Carola-Bibiane Schönlieb and Andrea L. Bertozzi (2011). “Unconditionally stable schemes for higher order inpainting”. *Comm. Math. Sci.* 9.2, pp. 413–457 (cit. on p. 3888).
- David I Shuman, Mohammadjavad Faraji, and Pierre Vandergheynst (2011). “Semi-supervised learning with spectral graph wavelets”. In: *Proceedings of the International Conference on Sampling Theory and Applications (SampTA)*. EPFL-CONF-164765 (cit. on p. 3895).
- Arthur Szlam and Xavier Bresson (2010). “Total Variation and Cheeger Cuts”. In: *Proceedings of the 27th International Conference on International Conference on Machine Learning*. ICML’10. USA: Omnipress, pp. 1039–1046 (cit. on p. 3891).
- M. Thorpe and F. Theil (2017). “Asymptotic Analysis of the Ginzburg-Landau Functional on Point Clouds”. to appear in the Proceedings of the Royal Society of Edinburgh Section A: Mathematics, 2017 (cit. on p. 3887).
- L. A. Vese and T. F. Chan (2002). “A multiphase level set framework for image segmentation using the Mumford-Shah model”. *Int. J. Comput. Vis.* 50, pp. 271–293 (cit. on p. 3888).
- Christopher K. I. Williams and Carl Edward Rasmussen (1996). *Gaussian Processes for Regression*. MIT (cit. on pp. 3895, 3896).
- J. T. Woodworth, G. O. Mohler, Andrea L. Bertozzi, and P. J. Brantingham (2014). “Non-local Crime Density Estimation Incorporating Housing Information”. *Phil. Trans. Roy. Soc. A* 372.2028 (cit. on p. 3904).
- L. P. Yaroslavsky (1985). *Digital Picture Processing. An Introduction*. Springer-Verlag (cit. on p. 3884).
- Lihi Zelnik-Manor and Pietro Perona (2004). “Self-tuning spectral clustering”. In: *Advances in neural information processing systems*, pp. 1601–1608 (cit. on p. 3885).
- Dengyong Zhou, Olivier Bousquet, Thomas Navin Lal, Jason Weston, and Bernhard Schölkopf (2004). “Learning with local and global consistency”. *Advances in neural information processing systems* 16.16, pp. 321–328 (cit. on p. 3895).
- Dengyong Zhou, Thomas Hofmann, and Bernhard Schölkopf (2004). “Semi-supervised learning on directed graphs”. In: *Advances in neural information processing systems*, pp. 1633–1640 (cit. on p. 3895).

- W. Zhu, V. Chayes, A. Tiard, S. Sanchez, D. Dahlberg, Andrea L. Bertozzi, S. Osher, D. Zosso, and D. Kuang (2017). “Unsupervised Classification in Hyperspectral Imagery With Nonlocal Total Variation and Primal-Dual Hybrid Gradient Algorithm”. *IEEE Transactions on Geoscience and Remote Sensing* 55.5, pp. 2786–2798 (cit. on pp. [3887](#), [3903](#)).
- Xiaojin Zhu (2005). “Semi-supervised learning literature survey”. Technical Report 1530, Computer Sciences, Univ. of Wisconsin-Madison (cit. on p. [3895](#)).
- Xiaojin Zhu, Zoubin Ghahramani, John Lafferty, et al. (2003). “Semi-supervised learning using Gaussian fields and harmonic functions”. In: *ICML*. Vol. 3, pp. 912–919 (cit. on p. [3895](#)).

Received 2017-12-02.

[ANDREA L. BERTOZZI](#)

DEPARTMENT OF MATHEMATICS

UCLA

[bertozzi@math.ucla.edu](mailto:bertozzi@math.ucla.edu)



# SCALABLE LOAD BALANCING IN NETWORKED SYSTEMS: UNIVERSALITY PROPERTIES AND STOCHASTIC COUPLING METHODS

MARK VAN DER BOOR, SEM C. BORST,  
JOHAN S. H. VAN LEEUWAARDEN AND DEBANKUR MUKHERJEE

## Abstract

We present an overview of scalable load balancing algorithms which provide favorable delay performance in large-scale systems, and yet only require minimal implementation overhead. Aimed at a broad audience, the paper starts with an introduction to the basic load balancing scenario – referred to as the *supermarket model* – consisting of a single dispatcher where tasks arrive that must immediately be forwarded to one of  $N$  single-server queues. The supermarket model is a dynamic counterpart of the classical balls-and-bins setup where balls must be sequentially distributed across bins.

A popular class of load balancing algorithms are power-of- $d$  or JSQ( $d$ ) policies, where an incoming task is assigned to a server with the shortest queue among  $d$  servers selected uniformly at random. As the name reflects, this class includes the celebrated Join-the-Shortest-Queue (JSQ) policy as a special case ( $d = N$ ), which has strong stochastic optimality properties and yields a mean waiting time that *vanishes* as  $N$  grows large for any fixed subcritical load. However, a nominal implementation of the JSQ policy involves a prohibitive communication burden in large-scale deployments. In contrast, a simple random assignment policy ( $d = 1$ ) does not entail any communication overhead, but the mean waiting time remains constant as  $N$  grows large for any fixed positive load.

In order to examine the fundamental trade-off between delay performance and implementation overhead, we consider an asymptotic regime where the diversity parameter  $d(N)$  depends on  $N$ . We investigate what growth rate of  $d(N)$  is required to match the optimal performance of the JSQ policy on fluid and diffusion scale, and achieve a vanishing waiting time in the limit. The results demonstrate that the asymptotics for the JSQ( $d(N)$ ) policy are insensitive to the exact growth rate of  $d(N)$ , as long as the latter is sufficiently fast, implying that the optimality of the JSQ policy

can asymptotically be preserved while dramatically reducing the communication overhead.

Stochastic coupling techniques play an instrumental role in establishing the asymptotic optimality and universality properties, and augmentations of the coupling constructions allow these properties to be extended to infinite-server settings and network scenarios. We additionally show how the communication overhead can be reduced yet further by the so-called Join-the-Idle-Queue (JIQ) scheme, leveraging memory at the dispatcher to keep track of idle servers.

## 1 Introduction

In the present paper we review scalable load balancing algorithms (LBAs) achieve excellent delay performance in large-scale systems and yet only involve low implementation overhead. LBAs play a critical role in distributing service requests or tasks (e.g. compute jobs, data base look-ups, file transfers) among servers or distributed resources in parallel-processing systems. The analysis and design of LBAs has attracted strong attention in recent years, mainly spurred by crucial scalability challenges arising in cloud networks and data centers with massive numbers of servers.

LBAs can be broadly categorized as static, dynamic, or some intermediate blend, depending on the amount of feedback or state information (e.g. congestion levels) that is used in allocating tasks. The use of state information naturally allows dynamic policies to achieve better delay performance, but also involves higher implementation complexity and a substantial communication burden. The latter issue is particularly pertinent in cloud networks and data centers with immense numbers of servers handling a huge influx of service requests. In order to capture the large-scale context, we examine scalability properties through the prism of asymptotic scalings where the system size grows large, and identify LBAs which strike an optimal balance between delay performance and implementation overhead in that regime.

The most basic load balancing scenario consists of  $N$  identical parallel servers and a dispatcher where tasks arrive that must immediately be forwarded to one of the servers. Tasks are assumed to have unit-mean exponentially distributed service requirements, and the service discipline at each server is supposed to be oblivious to the actual service requirements. In this canonical setup, the celebrated Join-the-Shortest-Queue (JSQ) policy has several strong stochastic optimality properties. In particular, the JSQ policy achieves the minimum mean overall delay among all non-anticipating policies that do not have any advance knowledge of the service requirements Ephremides, Varaiya, and Walrand [1980] and Winston [1977]. In order to implement the JSQ policy however, a dispatcher requires instantaneous knowledge of all the queue lengths, which may involve a prohibitive communication burden with a large number of servers  $N$ .

This poor scalability has motivated consideration of JSQ( $d$ ) policies, where an incoming task is assigned to a server with the shortest queue among  $d \geq 2$  servers selected uniformly at random. Note that this involves exchange of  $2d$  messages per task, irrespective of the number of servers  $N$ . Results in [Mitzenmacher \[2001\]](#) and [Vvedenskaya, Dobrushin, and Karpelevich \[1996\]](#) indicate that even sampling as few as  $d = 2$  servers yields significant performance enhancements over purely random assignment ( $d = 1$ ) as  $N$  grows large, which is commonly referred to as the “power-of-two” or “power-of-choice” effect. Specifically, when tasks arrive at rate  $\lambda N$ , the queue length distribution at each individual server exhibits super-exponential decay for any fixed  $\lambda < 1$  as  $N$  grows large, compared to exponential decay for purely random assignment.

As illustrated by the above, the diversity parameter  $d$  induces a fundamental trade-off between the amount of communication overhead and the delay performance. Specifically, a random assignment policy does not entail any communication burden, but the mean waiting time remains *constant* as  $N$  grows large for any fixed  $\lambda > 0$ . In contrast, a nominal implementation of the JSQ policy (without maintaining state information at the dispatcher) involves  $2N$  messages per task, but the mean waiting time *vanishes* as  $N$  grows large for any fixed  $\lambda < 1$ . Although JSQ( $d$ ) policies with  $d \geq 2$  yield major performance improvements over purely random assignment while reducing the communication burden by a factor  $O(N)$  compared to the JSQ policy, the mean waiting time *does not vanish* in the limit. Thus, no fixed value of  $d$  will provide asymptotically optimal delay performance. This is evidenced by results of [Gamarnik, Tsitsiklis, and Zubeldia \[2016\]](#) indicating that in the absence of any memory at the dispatcher the communication overhead per task *must increase* with  $N$  in order for any scheme to achieve a zero mean waiting time in the limit.

We will explore the intrinsic trade-off between delay performance and communication overhead as governed by the diversity parameter  $d$ , in conjunction with the relative load  $\lambda$ . The latter trade-off is examined in an asymptotic regime where not only the overall task arrival rate is assumed to grow with  $N$ , but also the diversity parameter is allowed to depend on  $N$ . We write  $\lambda(N)$  and  $d(N)$ , respectively, to explicitly reflect that, and investigate what growth rate of  $d(N)$  is required, depending on the scaling behavior of  $\lambda(N)$ , in order to achieve a zero mean waiting time in the limit. We establish that the fluid-scale and diffusion-scale limiting processes are insensitive to the exact growth rate of  $d(N)$ , as long as the latter is sufficiently fast, and in particular coincide with the limiting processes for the JSQ policy. This reflects a remarkable universality property and demonstrates that the optimality of the JSQ policy can asymptotically be preserved while dramatically lowering the communication overhead.

We will extend the above-mentioned universality properties to network scenarios where the  $N$  servers are assumed to be inter-connected by some underlying graph topology  $G_N$ . Tasks arrive at the various servers as independent Poisson processes of rate  $\lambda$ , and each incoming task is assigned to whichever server has the shortest queue among the one where



it appears and its neighbors in  $G_N$ . In case  $G_N$  is a clique, each incoming task is assigned to the server with the shortest queue across the entire system, and the behavior is equivalent to that under the JSQ policy. The above-mentioned stochastic optimality properties of the JSQ policy thus imply that the queue length process in a clique will be ‘better’ than in an arbitrary graph  $G_N$ . We will establish sufficient conditions for the fluid-scaled and diffusion-scaled versions of the queue length process in an arbitrary graph to be equivalent to the limiting processes in a clique as  $N \rightarrow \infty$ . The conditions reflect similar universality properties as described above, and in particular demonstrate that the optimality of a clique can asymptotically be preserved while markedly reducing the number of connections, provided the graph  $G_N$  is suitably random.

While a zero waiting time can be achieved in the limit by sampling only  $d(N) = o(N)$  servers, the amount of communication overhead in terms of  $d(N)$  must still grow with  $N$ . This may be explained from the fact that a large number of servers need to be sampled for each incoming task to ensure that at least one of them is found idle with high probability. As alluded to above, this can be avoided by introducing memory at the dispatcher, in particular maintaining a record of vacant servers, and assigning tasks to idle servers, if there are any. This so-called Join-the-Idle-Queue (JIQ) scheme [Badonnel and Burgess \[2008\]](#) and [Lu, Xie, Kliot, Geller, Larus, and Greenberg \[2011\]](#) has gained huge popularity recently, and can be implemented through a simple token-based mechanism generating at most one message per task. As established by [Stolyar \[2015\]](#), the fluid-scaled queue length process under the JIQ scheme is equivalent to that under the JSQ policy as  $N \rightarrow \infty$ , and this result can be shown to extend the diffusion-scaled queue length process. Thus, the use of memory allows the JIQ scheme to achieve asymptotically optimal delay performance with minimal communication overhead. In particular, ensuring that tasks are assigned to idle servers whenever available is sufficient to achieve asymptotic optimality, and using any additional queue length information yields no meaningful performance benefits on the fluid or diffusion levels.

Stochastic coupling techniques play an instrumental role in the proofs of the above-described universality and asymptotic optimality properties. A direct analysis of the queue length processes under a  $\text{JSQ}(d(N))$  policy, in a load balancing graph  $G_N$ , or under the JIQ scheme is confronted with unsurmountable obstacles. As an alternative route, we leverage novel stochastic coupling constructions to relate the relevant queue length processes to the corresponding processes under a JSQ policy, and show that the deviation between these two is asymptotically negligible under mild assumptions on  $d(N)$  or  $G_N$ .

While the stochastic coupling schemes provide a remarkably effective and overarching approach, they defy a systematic recipe and involve some degree of ingenuity and customization. Indeed, the specific coupling arguments that we develop are not only different from those that were originally used in establishing the stochastic optimality properties of

the JSQ policy, but also differ in critical ways between a  $\text{JSQ}(d(N))$  policy, a load balancing graph  $G_N$ , and the JIQ scheme. Yet different coupling constructions are devised for model variants with infinite-server dynamics that we will discuss in [Section 4](#).

The remainder of the paper is organized as follows. In [Section 2](#) we discuss a wide spectrum of LBAs and evaluate their scalability properties. In [Section 3](#) we introduce some useful preliminaries, review fluid and diffusion limits for the JSQ policy as well as  $\text{JSQ}(d)$  policies with a fixed value of  $d$ , and explore the trade-off between delay performance and communication overhead as function of the diversity parameter  $d$ . In particular, we establish asymptotic universality properties for  $\text{JSQ}(d)$  policies, which are extended to systems with server pools and network scenarios in [Sections 4](#) and [5](#), respectively. In [Section 6](#) we establish asymptotic optimality properties for the JIQ scheme. We discuss somewhat related redundancy policies and alternative scaling regimes and performance metrics in [Section 7](#).

## 2 Scalability spectrum

In this section we review a wide spectrum of LBAs and examine their scalability properties in terms of the delay performance vis-a-vis the associated implementation overhead in large-scale systems.

**2.1 Basic model.** Throughout this section and most of the paper, we focus on a basic scenario with  $N$  parallel single-server infinite-buffer queues and a single dispatcher where tasks arrive as a Poisson process of rate  $\lambda(N)$ , as depicted in [Figure 2](#). Arriving tasks cannot be queued at the dispatcher, and must immediately be forwarded to one of the servers. This canonical setup is commonly dubbed the *supermarket model*. Tasks are assumed to have unit-mean exponentially distributed service requirements, and the service discipline at each server is supposed to be oblivious to the actual service requirements.

In [Section 4](#) we consider some model variants with  $N$  server pools and possibly finite buffers and in [Section 5](#) we will treat network generalizations of the above model.

**2.2 Asymptotic scaling regimes.** An exact analysis of the delay performance is quite involved, if not intractable, for all but the simplest LBAs. Numerical evaluation or simulation are not straightforward either, especially for high load levels and large system sizes. A common approach is therefore to consider various limit regimes, which not only provide mathematical tractability and illuminate the fundamental behavior, but are also natural in view of the typical conditions in which cloud networks and data centers operate. One can distinguish several asymptotic scalings that have been used for these purposes: (i) In the classical heavy-traffic regime,  $\lambda(N) = \lambda N$  with a fixed number of servers  $N$

and a relative load  $\lambda$  that tends to one in the limit. (ii) In the conventional large-capacity or many-server regime, the relative load  $\lambda(N)/N$  approaches a constant  $\lambda < 1$  as the number of servers  $N$  grows large. (iii) The popular Halfin-Whitt regime [Halfin and Whitt \[1981\]](#) combines heavy traffic with a large capacity, with

$$(2-1) \quad \frac{N - \lambda(N)}{\sqrt{N}} \rightarrow \beta > 0 \text{ as } N \rightarrow \infty,$$

so the relative capacity slack behaves as  $\beta/\sqrt{N}$  as the number of servers  $N$  grows large. (iv) The so-called non-degenerate slow-down regime [Atar \[2012\]](#) involves  $N - \lambda(N) \rightarrow \gamma > 0$ , so the relative capacity slack shrinks as  $\gamma/N$  as the number of servers  $N$  grows large.

The term non-degenerate slow-down refers to the fact that in the context of a centralized multi-server queue, the mean waiting time in regime (iv) tends to a strictly positive constant as  $N \rightarrow \infty$ , and is thus of similar magnitude as the mean service requirement. In contrast, in regimes (ii) and (iii), the mean waiting time decays exponentially fast in  $N$  or is of the order  $1/\sqrt{N}$ , respectively, as  $N \rightarrow \infty$ , while in regime (i) the mean waiting time grows arbitrarily large relative to the mean service requirement.

In the present paper we will focus on scalings (ii) and (iii), and occasionally also refer to these as fluid and diffusion scalings, since it is natural to analyze the relevant queue length process on fluid scale ( $1/N$ ) and diffusion scale ( $1/\sqrt{N}$ ) in these regimes, respectively. We will not provide a detailed account of scalings (i) and (iv), which do not capture the large-scale perspective and do not allow for low delays, respectively, but we will briefly revisit these regimes in [Section 7](#).

**2.3 Random assignment:  $N$  independent M/M/1 queues.** One of the most basic LBAs is to assign each arriving task to a server selected uniformly at random. In that case, the various queues collectively behave as  $N$  independent M/M/1 queues, each with arrival rate  $\lambda(N)/N$  and unit service rate. In particular, at each of the queues, the total number of tasks in stationarity has a geometric distribution with parameter  $\lambda(N)/N$ . By virtue of the PASTA property, the probability that an arriving task incurs a non-zero waiting time is  $\lambda(N)/N$ . The mean number of waiting tasks (excluding the possible task in service) at each of the queues is  $\frac{\lambda(N)^2}{N(N - \lambda(N))}$ , so the total mean number of waiting tasks is  $\frac{\lambda(N)^2}{N - \lambda(N)}$ , which by Little's law implies that the mean waiting time of a task is  $\frac{\lambda(N)}{N - \lambda(N)}$ . In particular, when  $\lambda(N) = N\lambda$ , the probability that a task incurs a non-zero waiting time is  $\lambda$ , and the mean waiting time of a task is  $\frac{\lambda}{1 - \lambda}$ , independent of  $N$ , reflecting the independence of the various queues.

A slightly better LBA is to assign tasks to the servers in a Round-Robin manner, dispatching every  $N$ -th task to the same server. In the large-capacity regime where  $\lambda(N) =$

$N\lambda$ , the inter-arrival time of tasks at each given queue will then converge to a constant  $1/\lambda$  as  $N \rightarrow \infty$ . Thus each of the queues will behave as an D/M/1 queue in the limit, and the probability of a non-zero waiting time and the mean waiting time will be somewhat lower than under purely random assignment. However, both the probability of a non-zero waiting time and the mean waiting time will still tend to strictly positive values and not vanish as  $N \rightarrow \infty$ .

**2.4 Join-the-Shortest Queue (JSQ).** Under the Join-the-Shortest-Queue (JSQ) policy, each arriving task is assigned to the server with the currently shortest queue (ties are broken arbitrarily). In the basic model described above, the JSQ policy has several strong stochastic optimality properties, and yields the ‘most balanced and smallest’ queue process among all non-anticipating policies that do not have any advance knowledge of the service requirements Ephremides, Varaiya, and Walrand [1980] and Winston [1977]. Specifically, the JSQ policy minimizes the joint queue length vector in a stochastic majorization sense, and in particular stochastically minimizes the total number of tasks in the system, and hence the mean overall delay. In order to implement the JSQ policy however, a dispatcher requires instantaneous knowledge of the queue lengths at all the servers. A nominal implementation would involve exchange of  $2N$  messages per task, and thus yield a prohibitive communication burden in large-scale systems.

**2.5 Join-the-Smallest-Workload (JSW): centralized M/M/N queue.** Under the Join-the-Smallest-Workload (JSW) policy, each arriving task is assigned to the server with the currently smallest workload. Note that this is an anticipating policy, since it requires advance knowledge of the service requirements of all the tasks in the system. Further observe that this policy (myopically) minimizes the waiting time for each incoming task, and mimics the operation of a centralized  $N$ -server queue with a FCFS discipline. The equivalence with a centralized  $N$ -server queue yields a strong optimality property of the JSW policy: The vector of joint workloads at the various servers observed by each incoming task is smaller in the Schur convex sense than under any alternative admissible policy Foss and Chernova [2001].

The equivalence with a centralized FCFS queue means that there cannot be any idle servers while tasks are waiting. In our setting with Poisson arrivals and exponential service requirements, it can therefore be shown that the total number of tasks under the JSW policy is stochastically smaller than under the JSQ policy. At the same time, it means that the total number of tasks under the JSW policy behaves as a birth-death process, which renders it far more tractable than the JSQ policy. Specifically, given that all the servers are busy, the total number of waiting tasks is geometrically distributed with parameter  $\lambda(N)/N$ . Thus the total mean number of waiting tasks is  $\Pi_W(N, \lambda(N)) \frac{\lambda(N)}{N - \lambda(N)}$ , and the mean waiting

time is  $\Pi_W(N, \lambda(N)) \frac{1}{N - \lambda(N)}$ , with  $\Pi_W(N, \lambda(N))$  denoting the probability of all servers being occupied and a task incurring a non-zero waiting time. This immediately shows that the mean waiting time is smaller by at least a factor  $\lambda(N)$  than for the random assignment policy considered in [Section 2.3](#).

In the large-capacity regime  $\lambda(N) = N\lambda$ , it can be shown that the probability  $\Pi_W(N, \lambda(N))$  of a non-zero waiting time decays exponentially fast in  $N$ , and hence so does the mean waiting time. In the Halfin-Whitt heavy-traffic regime (2-1), the probability  $\Pi_W(N, \lambda(N))$  of a non-zero waiting time converges to a finite constant  $\Pi_W^*(\beta)$ , implying that the mean waiting time of a task is of the order  $1/\sqrt{N}$ , and thus vanishes as  $N \rightarrow \infty$ .

**2.6 Power-of- $d$  load balancing (JSQ( $d$ )).** As mentioned above, the achilles heel of the JSQ policy is its excessive communication overhead in large-scale systems. This poor scalability has motivated consideration of so-called JSQ( $d$ ) policies, where an incoming task is assigned to a server with the shortest queue among  $d$  servers selected uniformly at random. Results in [Mitzenmacher \[2001\]](#) and [Vvedenskaya, Dobrushin, and Karpelevich \[1996\]](#) indicate that even sampling as few as  $d = 2$  servers yields significant performance enhancements over purely random assignment ( $d = 1$ ) as  $N \rightarrow \infty$ . Specifically, in the fluid regime where  $\lambda(N) = \lambda N$ , the probability that there are  $i$  or more tasks at a given queue is proportional to  $\lambda^{\frac{d^i - 1}{d - 1}}$  as  $N \rightarrow \infty$ , and thus exhibits super-exponential decay as opposed to exponential decay for the random assignment policy considered in [Section 2.3](#).

As illustrated by the above, the diversity parameter  $d$  induces a fundamental trade-off between the amount of communication overhead and the performance in terms of queue lengths and delays. A rudimentary implementation of the JSQ policy ( $d = N$ , without replacement) involves  $O(N)$  communication overhead per task, but it can be shown that the probability of a non-zero waiting time and the mean waiting *vanish* as  $N \rightarrow \infty$ , just like in a centralized queue. Although JSQ( $d$ ) policies with a fixed parameter  $d \geq 2$  yield major performance improvements over purely random assignment while reducing the communication burden by a factor  $O(N)$  compared to the JSQ policy, the probability of a non-zero waiting time and the mean waiting time *do not vanish* as  $N \rightarrow \infty$ .

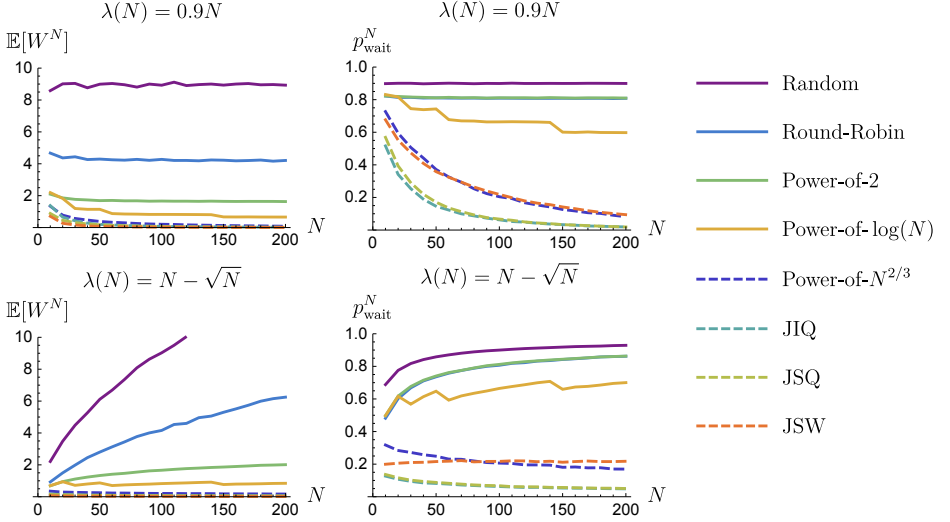
In [Section 3.5](#) we will explore the intrinsic trade-off between delay performance and communication overhead as function of the diversity parameter  $d$ , in conjunction with the relative load. We will examine an asymptotic regime where not only the total task arrival rate  $\lambda(N)$  is assumed to grow with  $N$ , but also the diversity parameter is allowed to depend on  $N$ . As will be demonstrated, the optimality of the JSQ policy ( $d(N) = N$ ) can be preserved, and in particular a vanishing waiting time can be achieved in the limit as  $N \rightarrow \infty$ , even when  $d(N) = o(N)$ , thus dramatically lowering the communication overhead.

**2.7 Token-based strategies: Join-the-Idle-Queue (JIQ).** While a zero waiting time can be achieved in the limit by sampling only  $d(N) = o(N)$  servers, the amount of communication overhead in terms of  $d(N)$  must still grow with  $N$ . This can be countered by introducing memory at the dispatcher, in particular maintaining a record of vacant servers, and assigning tasks to idle servers as long as there are any, or to a uniformly at random selected server otherwise. This so-called Join-the-Idle-Queue (JIQ) scheme [Badonnell and Burgess \[2008\]](#) and [Lu, Xie, Kliot, Geller, Larus, and Greenberg \[2011\]](#) has received keen interest recently, and can be implemented through a simple token-based mechanism. Specifically, idle servers send tokens to the dispatcher to advertise their availability, and when a task arrives and the dispatcher has tokens available, it assigns the task to one of the corresponding servers (and disposes of the token). Note that a server only issues a token when a task completion leaves its queue empty, thus generating at most one message per task. Surprisingly, the mean waiting time and the probability of a non-zero waiting time vanish under the JIQ scheme in both the fluid and diffusion regimes, as we will further discuss in [Section 6](#). Thus, the use of memory allows the JIQ scheme to achieve asymptotically optimal delay performance with minimal communication overhead.

**2.8 Performance comparison.** We now present some simulation experiments that we have conducted to compare the above-described LBAs in terms of delay performance. Specifically, we evaluate the mean waiting time and the probability of a non-zero waiting time in both a fluid regime ( $\lambda(N) = 0.9N$ ) and a diffusion regime ( $\lambda(N) = N - \sqrt{N}$ ). The results are shown in [Figure 1](#). We are especially interested in distinguishing two classes of LBAs – ones delivering a mean waiting time and probability of a non-zero waiting time that vanish asymptotically, and ones that fail to do so – and relating that dichotomy to the associated overhead.

**JSQ, JIQ, and JSW..** JSQ, JIQ and JSW evidently have a vanishing waiting time in both the fluid and the diffusion regime as discussed in [Sections 2.4, 2.5 and 2.7](#). The optimality of JSW as mentioned in [Section 2.5](#) can also be clearly observed.

However, there is a significant difference between JSW and JSQ/JIQ in the diffusion regime. We observe that the probability of a non-zero waiting time *approaches a positive constant* for JSW, while it *vanishes* for JSQ/JIQ. In other words, the mean of all positive waiting times is of a larger order of magnitude in JSQ/JIQ compared to JSW. Intuitively, this is clear since in JSQ/JIQ, when a task is placed in a queue, it waits for at least a residual service time. In JSW, which is equivalent to the M/M/N queue, a task that cannot start service immediately, joins a queue that is collectively drained by all the  $N$  servers



**Figure 1:** Simulation results for mean waiting time  $\mathbb{E}[W^N]$  and probability of a non-zero waiting time  $p_{\text{wait}}^N$ , for both a fluid regime and a diffusion regime.

**Random and Round-Robin.** The mean waiting time does not vanish for Random and Round-Robin in the fluid regime, as already mentioned in [Section 2.3](#). Moreover, the mean waiting time grows without bound in the diffusion regime for these two schemes. This is because the system can still be decomposed, and the loads of the individual M/M/1 and D/M/1 queues tend to 1.

**JSQ(d) policies.** Three versions of  $\text{JSQ}(d)$  are included in the figures;  $d(N) = 2 \not\rightarrow \infty$ ,  $d(N) = \lfloor \log(N) \rfloor \rightarrow \infty$  and  $d(N) = N^{2/3}$  for which  $\frac{d(N)}{\sqrt{N \log(N)}} \rightarrow \infty$ . Note that the graph for  $d(N) = \lfloor \log(N) \rfloor$  shows sudden jumps when  $d(N)$  increases by 1. The variants for which  $d(N) \rightarrow \infty$  have a vanishing waiting time in the fluid regime, while  $d = 2$  does not. The latter observation is a manifestation of the results of [Gamarnik, Tsitsiklis, and Zubeldia \[2016\]](#) mentioned in the introduction, since  $\text{JSQ}(d)$  uses no memory and the overhead per task does not increase with  $N$ . Furthermore, it follows that  $\text{JSQ}(d)$  policies outperform Random and Round-Robin, while  $\text{JSQ}/\text{JIQ}/\text{JSW}$  are better in terms of mean waiting time.

In order to succinctly capture the results and observed dichotomy in [Figure 1](#), we provide an overview of the delay performance of the various LBAs and the associated overhead in [Table 1](#), where  $q_i^*$  denotes the stationary fraction of servers with  $i$  or more tasks.

Scheme	Queue length	Waiting time (fixed $\lambda < 1$ )	Waiting time ( $1 - \lambda \sim 1/\sqrt{N}$ )	Over- head per task
Random	$q_i^* = \lambda^i$	$\frac{\lambda}{1-\lambda}$	$\Theta(\sqrt{N})$	0
JSQ( $d$ )	$q_i^* = \lambda^{\frac{d^i-1}{d-1}}$	$\Theta(1)$	$\Omega(\log N)$	$2d$
$d(N) \rightarrow \infty$	same as JSQ	same as JSQ	??	$2d(N)$
$\frac{d(N)}{\sqrt{N} \log(N)} \rightarrow \infty$	same as JSQ	same as JSQ	same as JSQ	$2d(N)$
JSQ	$q_1^* = \lambda, q_2^* = o(1)$	$o(1)$	$\Theta(1/\sqrt{N})$	$2N$
JIQ	same as JSQ	same as JSQ	same as JSQ	$\leq 1$

**Table 1:** Queue length distribution, waiting times and communication overhead for various LBAs.

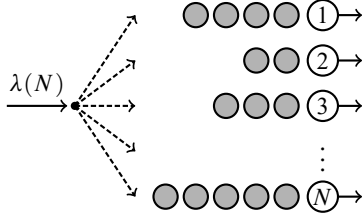
### 3 JSQ( $d$ ) policies and universality properties

In this section we first introduce some useful preliminary concepts, then review fluid and diffusion limits for the JSQ policy as well as JSQ( $d$ ) policies with a fixed value of  $d$ , and finally discuss universality properties when the diversity parameter  $d(N)$  is being scaled with  $N$ .

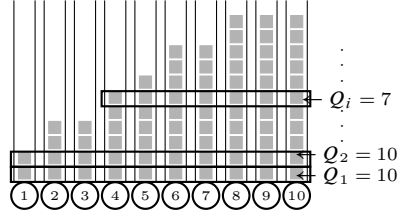
As described in the previous section, we focus on a basic scenario where all the servers are homogeneous, the service requirements are exponentially distributed, and the service discipline at each server is oblivious of the actual service requirements. In order to obtain a Markovian state description, it therefore suffices to only track the number of tasks, and in fact we do not need to keep record of the number of tasks at each individual server, but only count the number of servers with a given number of tasks. Specifically, we represent the state of the system by a vector  $\mathbf{Q}(t) := (Q_1(t), Q_2(t), \dots)$ , with  $Q_i(t)$  denoting the number of servers with  $i$  or more tasks at time  $t$ , including the possible task in service,  $i = 1, 2, \dots$ . Note that if we represent the queues at the various servers as (vertical) stacks, and arrange these from left to right in non-descending order, then the value of  $Q_i$  corresponds to the width of the  $i$ -th (horizontal) row, as depicted in the schematic diagram in [Figure 3](#).

In order to examine the asymptotic behavior when the number of servers  $N$  grows large, we consider a sequence of systems indexed by  $N$ , and attach a superscript  $N$  to the associated state variables.





**Figure 2:** Tasks arrive at the dispatcher as a Poisson process of rate  $\lambda(N)$ , and are forwarded to one of the  $N$  servers according to some specific load balancing algorithm.



**Figure 3:** The value of  $Q_i$  represents the width of the  $i$ -th row, when the servers are arranged in non-decreasing order of their queue lengths.

The fluid-scaled occupancy state is denoted by  $\mathbf{q}^N(t) := (q_1^N(t), q_2^N(t), \dots)$ , with  $q_i^N(t) = Q_i^N(t)/N$  representing the fraction of servers in the  $N$ -th system with  $i$  or more tasks as time  $t$ ,  $i = 1, 2, \dots$ . Let  $\mathcal{S} = \{\mathbf{q} \in [0, 1]^\infty : q_i \leq q_{i-1} \forall i = 2, 3, \dots\}$  be the set of all possible fluid-scaled states. Whenever we consider fluid limits, we assume the sequence of initial states is such that  $\mathbf{q}^N(0) \rightarrow \mathbf{q}^\infty \in \mathcal{S}$  as  $N \rightarrow \infty$ .

The diffusion-scaled occupancy state is defined as  $\bar{\mathbf{Q}}^N(t) = (\bar{Q}_1^N(t), \bar{Q}_2^N(t), \dots)$ , with

$$(3-1) \quad \bar{Q}_1^N(t) = -\frac{N - Q_1^N(t)}{\sqrt{N}}, \quad \bar{Q}_i^N(t) = \frac{Q_i^N(t)}{\sqrt{N}}, \quad i = 2, 3, \dots$$

Note that  $-\bar{Q}_1^N(t)$  corresponds to the number of vacant servers, normalized by  $\sqrt{N}$ . The reason why  $Q_1^N(t)$  is centered around  $N$  while  $Q_i^N(t)$ ,  $i = 2, 3, \dots$ , are not, is because for the scalable LBAs that we pursue, the fraction of servers with exactly one task tends to one, whereas the fraction of servers with two or more tasks tends to zero as  $N \rightarrow \infty$ .

**3.1 Fluid limit for JSQ( $d$ ) policies.** We first consider the fluid limit for JSQ( $d$ ) policies with an arbitrary but fixed value of  $d$  as characterized by Mitzenmacher [2001] and Vvedenskaya, Dobrushin, and Karpelevich [1996].

The sequence of processes  $\{\mathbf{q}^N(t)\}_{t \geq 0}$  has a weak limit  $\{\mathbf{q}(t)\}_{t \geq 0}$  that satisfies the system of differential equations

$$(3-2) \quad \frac{dq_i(t)}{dt} = \lambda[(q_{i-1}(t))^d - (q_i(t))^d] - [q_i(t) - q_{i+1}(t)], \quad i = 1, 2, \dots$$

The fluid-limit equations may be interpreted as follows. The first term represents the rate of increase in the fraction of servers with  $i$  or more tasks due to arriving tasks that are

assigned to a server with exactly  $i - 1$  tasks. Note that the latter occurs in fluid state  $\mathbf{q} \in \mathcal{S}$  with probability  $q_{i-1}^d - q_i^d$ , i.e., the probability that all  $d$  sampled servers have  $i - 1$  or more tasks, but not all of them have  $i$  or more tasks. The second term corresponds to the rate of decrease in the fraction of servers with  $i$  or more tasks due to service completions from servers with exactly  $i$  tasks, and the latter rate is given by  $q_i - q_{i+1}$ .

The unique fixed point of (3-2) for any  $d \geq 2$  is obtained as

$$(3-3) \quad q_i^* = \lambda^{\frac{d^i - 1}{d - 1}}, \quad i = 1, 2, \dots$$

It can be shown that the fixed point is asymptotically stable in the sense that  $\mathbf{q}(t) \rightarrow \mathbf{q}^*$  as  $t \rightarrow \infty$  for any initial fluid state  $\mathbf{q}^\infty$  with  $\sum_{i=1}^{\infty} q_i^\infty < \infty$ . The fixed point reveals that the stationary queue length distribution at each individual server exhibits super-exponential decay as  $N \rightarrow \infty$ , as opposed to exponential decay for a random assignment policy. It is worth observing that this involves an interchange of the many-server ( $N \rightarrow \infty$ ) and stationary ( $t \rightarrow \infty$ ) limits. The justification is provided by the asymptotic stability of the fixed point along with a few further technical conditions.

**3.2 Fluid limit for JSQ policy.** We now turn to the fluid limit for the ordinary JSQ policy, which rather surprisingly was not rigorously established until fairly recently in [Mukherjee, Borst, van Leeuwen, and Whiting \[2016c\]](#), leveraging martingale functional limit theorems and time-scale separation arguments [Hunt and Kurtz \[1994\]](#).

In order to state the fluid limit starting from an arbitrary fluid-scaled occupancy state, we first introduce some additional notation. For any fluid state  $\mathbf{q} \in \mathcal{S}$ , denote by  $m(\mathbf{q}) = \min\{i : q_{i+1} < 1\}$  the minimum queue length among all servers. Now if  $m(\mathbf{q}) = 0$ , then define  $p_0(m(\mathbf{q})) = 1$  and  $p_i(m(\mathbf{q})) = 0$  for all  $i = 1, 2, \dots$ . Otherwise, in case  $m(\mathbf{q}) > 0$ , define

$$(3-4) \quad p_i(\mathbf{q}) = \begin{cases} \min\{(1 - q_{m(\mathbf{q})+1})/\lambda, 1\} & \text{for } i = m(\mathbf{q}) - 1, \\ 1 - p_{m(\mathbf{q})-1}(\mathbf{q}) & \text{for } i = m(\mathbf{q}), \end{cases}$$

and  $p_i(\mathbf{q}) = 0$  otherwise. The coefficient  $p_i(\mathbf{q})$  represents the instantaneous fraction of incoming tasks assigned to servers with a queue length of exactly  $i$  in the fluid state  $\mathbf{q} \in \mathcal{S}$ .

Any weak limit of the sequence of processes  $\{\mathbf{q}^N(t)\}_{t \geq 0}$  is given by the deterministic system  $\{\mathbf{q}(t)\}_{t \geq 0}$  satisfying the following system of differential equations

$$(3-5) \quad \frac{d^+ q_i(t)}{dt} = \lambda p_{i-1}(\mathbf{q}(t)) - (q_i(t) - q_{i+1}(t)), \quad i = 1, 2, \dots,$$

where  $d^+/dt$  denotes the right-derivative.

The unique fixed point  $\mathbf{q}^* = (q_1^*, q_2^*, \dots)$  of the dynamical system in (3-5) is given by

$$(3-6) \quad q_i^* = \begin{cases} \lambda, & i = 1, \\ 0, & i = 2, 3, \dots \end{cases}$$

The fixed point in (3-6), in conjunction with an interchange of limits argument, indicates that in stationarity the fraction of servers with a queue length of two or larger under the JSQ policy is negligible as  $N \rightarrow \infty$ .

**3.3 Diffusion limit for JSQ policy.** We next describe the diffusion limit for the JSQ policy in the Halfin-Whitt heavy-traffic regime (2-1), as recently derived by [Eschenfeldt and Gamarnik \[2015\]](#).

*For suitable initial conditions, the sequence of processes  $\{\bar{\mathbf{Q}}^N(t)\}_{t \geq 0}$  as in (3-1) converges weakly to the limit  $\{\bar{\mathbf{Q}}(t)\}_{t \geq 0}$ , where  $(\bar{Q}_1(t), \bar{Q}_2(t), \dots)$  is the unique solution to the following system of SDEs*

$$(3-7) \quad \begin{aligned} d\bar{Q}_1(t) &= \sqrt{2}dW(t) - \beta dt - \bar{Q}_1(t)dt + \bar{Q}_2(t)dt - dU_1(t), \\ d\bar{Q}_2(t) &= dU_1(t) - (\bar{Q}_2(t) - \bar{Q}_3(t))dt, \\ d\bar{Q}_i(t) &= -(\bar{Q}_i(t) - \bar{Q}_{i+1}(t))dt, \quad i \geq 3, \end{aligned}$$

for  $t \geq 0$ , where  $W(\cdot)$  is the standard Brownian motion and  $U_1(\cdot)$  is the unique nondecreasing nonnegative process satisfying  $\int_0^\infty \mathbb{1}_{[\bar{Q}_1(t) < 0]} dU_1(t) = 0$ .

The above diffusion limit implies that the mean waiting time under the JSQ policy is of a similar order  $O(1/\sqrt{N})$  as in the corresponding centralized M/M/N queue. Hence, we conclude that despite the distributed queueing operation a suitable load balancing policy can deliver a similar combination of excellent service quality and high resource utilization in the Halfin-Whitt regime (2-1) as in a centralized queueing arrangement. It is important though to observe a subtle but fundamental difference in the distributional properties due to the distributed versus centralized queueing operation. In the ordinary M/M/N queue a fraction  $\Pi_W^*(\beta)$  of the customers incur a non-zero waiting time as  $N \rightarrow \infty$ , but a non-zero waiting time is only of length  $1/(\beta\sqrt{N})$  in expectation. In contrast, under the JSQ policy, the fraction of tasks that experience a non-zero waiting time is only of the order  $O(1/\sqrt{N})$ . However, such tasks will have to wait for the duration of a residual service time, yielding a waiting time of the order  $O(1)$ .

**3.4 Heavy-traffic limits for JSQ(d) policies.** Finally, we briefly discuss the behavior of JSQ(d) policies for fixed  $d$  in a heavy-traffic regime where  $(N - \lambda(N))/\eta(N) \rightarrow \beta > 0$  as  $N \rightarrow \infty$  with  $\eta(N)$  a positive function diverging to infinity. Note that the case  $\eta(N) = \sqrt{N}$  corresponds to the Halfin-Whitt heavy-traffic regime (2-1). While a

complete characterization of the occupancy process for fixed  $d$  has remained elusive so far, significant partial results were recently obtained by [Eschenfeldt and Gamarnik \[2016\]](#). In order to describe the transient asymptotics, we introduce the following rescaled processes  $\bar{Q}_i^N(t) := (N - Q_i^N(t))/\eta(N)$ ,  $i = 1, 2, \dots$

Then, for suitable initial states, on any finite time interval,  $\{\bar{\mathbf{Q}}^N(t)\}_{t \geq 0}$  converges weakly to a deterministic system  $\{\bar{\mathbf{Q}}(t)\}_{t \geq 0}$  that satisfies the following system of ODEs

$$(3-8) \quad \frac{d\bar{Q}_i(t)}{dt} = -d[\bar{Q}_i(t) - \bar{Q}_{i-1}(t)] - [\bar{Q}_i(t) - \bar{Q}_{i+1}(t)], \quad i = 1, 2, \dots,$$

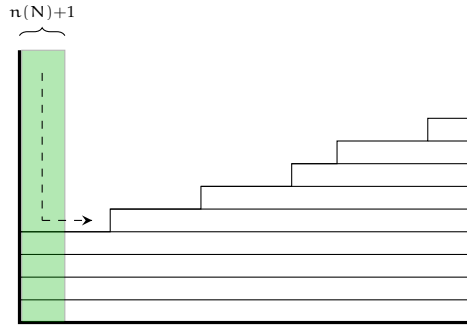
with the convention that  $\bar{Q}_0(t) \equiv 0$ .

It is noteworthy that the scaled occupancy process loses its diffusive behavior for fixed  $d$ . It is further shown in [Eschenfeldt and Gamarnik \[ibid.\]](#) that with high probability the steady-state fraction of queues with length at least  $\log_d(N/\eta(N)) - \omega(1)$  tasks approaches unity, which in turn implies that with high probability the steady-state delay is at least  $\log_d(N/\eta(N)) - O(1)$  as  $N \rightarrow \infty$ . The diffusion approximation of the JSQ( $d$ ) policy in the Halfin-Whitt regime (2-1), starting from a different initial scaling, has been studied by [Budhiraja and Friedlander \[2017\]](#). Recently, [Ying \[2017\]](#) introduced a broad framework involving Stein's method to analyze the rate of convergence of the scaled steady-state occupancy process of the JSQ(2) policy when  $\eta(N) = N^\alpha$  with  $\alpha > 0.8$ . The results in [Ying \[ibid.\]](#) establish that in steady state, most of the queues are of size  $\log_2(N/\eta(N)) + O(1)$ , and thus the steady-state delay is of order  $\log_2(N/\eta(N))$ .

**3.5 Universality properties.** We now further explore the trade-off between delay performance and communication overhead as a function of the diversity parameter  $d$ , in conjunction with the relative load. The latter trade-off will be examined in an asymptotic regime where not only the total task arrival rate  $\lambda(N)$  grows with  $N$ , but also the diversity parameter depends on  $N$ , and we write  $d(N)$ , to explicitly reflect that. We will specifically investigate what growth rate of  $d(N)$  is required, depending on the scaling behavior of  $\lambda(N)$ , in order to asymptotically match the optimal performance of the JSQ policy and achieve a zero mean waiting time in the limit. The results presented in this section are based on [Mukherjee, Borst, van Leeuwen, and Whiting \[2016c\]](#), unless specified otherwise.

**Theorem 3.1.** (Universality fluid limit for JSQ( $d(N)$ )) *If  $d(N) \rightarrow \infty$  as  $N \rightarrow \infty$ , then the fluid limit of the JSQ( $d(N)$ ) scheme coincides with that of the ordinary JSQ policy given by the dynamical system in (3-5). Consequently, the stationary occupancy states converge to the unique fixed point in (3-6).*

**Theorem 3.2.** (Universality diffusion limit for JSQ( $d(N)$ )) *If  $d(N)/(\sqrt{N} \log N) \rightarrow \infty$ , then for suitable initial conditions the weak limit of the sequence of processes  $\{\bar{\mathbf{Q}}^{d(N)}(t)\}_{t \geq 0}$*



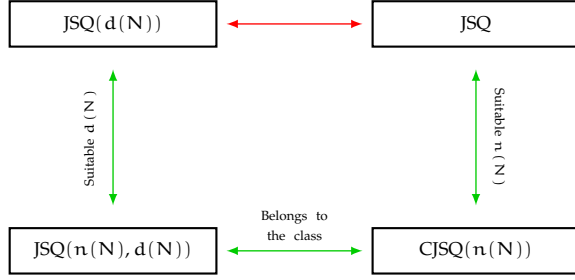
**Figure 4:** CJSQ( $n(N)$ ) scheme. High-level view of the CJSQ( $n(N)$ ) class of schemes, where as in Figure 3, the servers are arranged in nondecreasing order of their queue lengths, and the arrival must be assigned through the left tunnel.

coincides with that of the ordinary JSQ policy, and in particular, is given by the system of SDEs in (3-7).

The above universality properties indicate that the JSQ overhead can be lowered by almost a factor  $O(N)$  and  $O(\sqrt{N}/\log N)$  while retaining fluid- and diffusion-level optimality, respectively. In other words, Theorems 3.1 and 3.2 thus reveal that it is sufficient for  $d(N)$  to grow at any rate and faster than  $\sqrt{N} \log N$  in order to observe similar scaling benefits as in a corresponding centralized M/M/N queue on fluid scale and diffusion scale, respectively. The stated conditions are in fact close to necessary, in the sense that if  $d(N)$  is uniformly bounded and  $d(N)/(\sqrt{N} \log N) \rightarrow 0$  as  $N \rightarrow \infty$ , then the fluid-limit and diffusion-limit paths of the system occupancy process under the JSQ( $d(N)$ ) scheme differ from those under the ordinary JSQ policy, respectively. In particular, if  $d(N)$  is uniformly bounded, the mean steady-state delay does not vanish asymptotically as  $N \rightarrow \infty$ .

**High-level proof idea.** The proofs of both Theorems 3.1 and 3.2 rely on a stochastic coupling construction to bound the difference in the queue length processes between the JSQ policy and a scheme with an arbitrary value of  $d(N)$ . This S-coupling (‘S’ stands for server-based) is then exploited to obtain the fluid and diffusion limits of the JSQ( $d(N)$ ) policy under the conditions stated in Theorems 3.1 and 3.2.

A direct comparison between the JSQ( $d(N)$ ) scheme and the ordinary JSQ policy is not straightforward, which is why the CJSQ( $n(N)$ ) class of schemes is introduced as an intermediate scenario to establish the universality result. Just like the JSQ( $d(N)$ ) scheme, the schemes in the class CJSQ( $n(N)$ ) may be thought of as “sloppy” versions of the JSQ policy, in the sense that tasks are not necessarily assigned to a server with the shortest



**Figure 5:** Asymptotic equivalence relations. The equivalence structure is depicted for various intermediate load balancing schemes to facilitate the comparison between the  $\text{JSQ}(d(N))$  scheme and the ordinary JSQ policy.

queue length but to one of the  $n(N) + 1$  lowest ordered servers, as graphically illustrated in Figure 4. In particular, for  $n(N) = 0$ , the class only includes the ordinary JSQ policy. Note that the  $\text{JSQ}(d(N))$  scheme is guaranteed to identify the lowest ordered server, but only among a randomly sampled subset of  $d(N)$  servers. In contrast, a scheme in the  $\text{CJSQ}(n(N))$  class only guarantees that one of the  $n(N) + 1$  lowest ordered servers is selected, but across the entire pool of  $N$  servers. It may be shown that for sufficiently small  $n(N)$ , any scheme from the class  $\text{CJSQ}(n(N))$  is still ‘close’ to the ordinary JSQ policy. It can further be proved that for sufficiently large  $d(N)$  relative to  $n(N)$  we can construct a scheme called  $\text{JSQ}(n(N), d(N))$ , belonging to the  $\text{CJSQ}(n(N))$  class, which differs ‘negligibly’ from the  $\text{JSQ}(d(N))$  scheme. Therefore, for a ‘suitable’ choice of  $d(N)$  the idea is to produce a ‘suitable’  $n(N)$ . This proof strategy is schematically represented in Figure 5.

In order to prove the stochastic comparisons among the various schemes, the many-server system is described as an ensemble of stacks, in a way that two different ensembles can be ordered. This stack formulation has also been considered in the literature for establishing the stochastic optimality properties of the JSQ policy Sparaggis, Towsley, and Cassandra [1994], Towsley [1995], and Towsley, Sparaggis, and Cassandra [1992]. However, it is only through the stack arguments developed in Mukherjee, Borst, van Leeuwen, and Whiting [2016c] that the comparison results can be extended to any scheme from the class CJSQ.

## 4 Blocking and infinite-server dynamics

The basic scenario that we have focused on so far involved single-server queues. In this section we turn attention to a system with parallel server pools, each with  $B$  servers, where

$B$  can possibly be infinite. As before, tasks must immediately be forwarded to one of the server pools, but also directly start execution or be discarded otherwise. The execution times are assumed to be exponentially distributed, and do not depend on the number of other tasks receiving service simultaneously. The current scenario will be referred to as ‘infinite-server dynamics’, in contrast to the earlier single-server queueing dynamics.

As it turns out, the JSQ policy has similar stochastic optimality properties as in the case of single-server queues, and in particular stochastically minimizes the cumulative number of discarded tasks [Sparaggis, Towsley, and Cassandras \[1993\]](#), [Johri \[1989\]](#), [Menich \[1987\]](#), and [Menich and Serfozo \[1991\]](#). However, the JSQ policy also suffers from a similar scalability issue due to the excessive communication overhead in large-scale systems, which can be mitigated through JSQ( $d$ ) policies. Results of [Turner \[1998\]](#) and recent papers by [Karthik, Mukhopadhyay, and Mazumdar \[2017\]](#), [Mukhopadhyay, Karthik, Mazumdar, and Guillemin \[2015\]](#), [Mukhopadhyay, Mazumdar, and Guillemin \[2015\]](#), and [Xie, Dong, Lu, and Srikant \[2015\]](#) indicate that JSQ( $d$ ) policies provide similar “power-of-choice” gains for loss probabilities. It may be shown though that the optimal performance of the JSQ policy cannot be matched for any fixed value of  $d$ .

Motivated by these observations, we explore the trade-off between performance and communication overhead for infinite-server dynamics. We will demonstrate that the optimal performance of the JSQ policy can be asymptotically retained while drastically reducing the communication burden, mirroring the universality properties described in [Section 3.5](#) for single-server queues. The results presented in the remainder of the section are extracted from [Mukherjee, Borst, van Leeuwen, and Whiting \[2016a\]](#), unless indicated otherwise.

**4.1 Fluid limit for JSQ policy.** As in [Section 3.2](#), for any fluid state  $\mathbf{q} \in \mathcal{S}$ , denote by  $m(\mathbf{q}) = \min\{i : q_{i+1} < 1\}$  the minimum queue length among all servers. Now if  $m(\mathbf{q}) = 0$ , then define  $p_0(m(\mathbf{q})) = 1$  and  $p_i(m(\mathbf{q})) = 0$  for all  $i = 1, 2, \dots$ . Otherwise, in case  $m(\mathbf{q}) > 0$ , define

$$(4-1) \quad p_i(\mathbf{q}) = \begin{cases} \min\{m(\mathbf{q})(1 - q_{m(\mathbf{q})+1})/\lambda, 1\} & \text{for } i = m(\mathbf{q}) - 1, \\ 1 - p_{m(\mathbf{q})-1}(\mathbf{q}) & \text{for } i = m(\mathbf{q}), \end{cases}$$

and  $p_i(\mathbf{q}) = 0$  otherwise. As before, the coefficient  $p_i(\mathbf{q})$  represents the instantaneous fraction of incoming tasks assigned to servers with a queue length of exactly  $i$  in the fluid state  $\mathbf{q} \in \mathcal{S}$ .

Any weak limit of the sequence of processes  $\{\mathbf{q}^N(t)\}_{t \geq 0}$  is given by the deterministic system  $\{\mathbf{q}(t)\}_{t \geq 0}$  satisfying the following of differential equations

$$(4-2) \quad \frac{d^+ q_i(t)}{dt} = \lambda p_{i-1}(\mathbf{q}(t)) - i(q_i(t) - q_{i+1}(t)), \quad i = 1, 2, \dots,$$

where  $d^+ / dt$  denotes the right-derivative.

Equations (4-1) and (4-2) are to be contrasted with Equations (3-4) and (3-5). While the form of (4-1) and the evolution equations (4-2) of the limiting dynamical system remains similar to that of (3-4) and (3-5), respectively, an additional factor  $m(\mathbf{q})$  appears in (4-1) and the rate of decrease in (4-2) now becomes  $i(q_i - q_{i+1})$ , reflecting the infinite-server dynamics.

Let  $K := \lfloor \lambda \rfloor$  and  $f := \lambda - K$  denote the integral and fractional parts of  $\lambda$ , respectively. It is easily verified that, assuming  $\lambda < B$ , the unique fixed point of the dynamical system in (4-2) is given by

$$(4-3) \quad q_i^* = \begin{cases} 1 & i = 1, \dots, K \\ f & i = K + 1 \\ 0 & i = K + 2, \dots, B, \end{cases}$$

and thus  $\sum_{i=1}^B q_i^* = \lambda$ . This is consistent with the results in [Mukhopadhyay, Karthik, Mazumdar, and Guillemin \[2015\]](#), [Mukhopadhyay, Mazumdar, and Guillemin \[2015\]](#), and [Xie, Dong, Lu, and Srikant \[2015\]](#) for fixed  $d$ , where taking  $d \rightarrow \infty$  yields the same fixed point. The fixed point in (4-3), in conjunction with an interchange of limits argument, indicates that in stationarity the fraction of server pools with at least  $K + 2$  and at most  $K - 1$  active tasks is negligible as  $N \rightarrow \infty$ .

**4.2 Diffusion limit for JSQ policy.** As it turns out, the diffusion-limit results may be qualitatively different, depending on whether  $f = 0$  or  $f > 0$ , and we will distinguish between these two cases accordingly. Observe that for any assignment scheme, in the absence of overflow events, the total number of active tasks evolves as the number of jobs in an M/M/ $\infty$  system, for which the diffusion limit is well-known. For the JSQ policy, it can be established that the total number of server pools with  $K - 2$  or less and  $K + 2$  or more tasks is negligible on the diffusion scale. If  $f > 0$ , the number of server pools with  $K - 1$  tasks is negligible as well, and the dynamics of the number of server pools with  $K$  or  $K + 1$  tasks can then be derived from the known diffusion limit of the total number of tasks mentioned above. In contrast, if  $f = 0$ , the number of server pools with  $K - 1$  tasks is not negligible on the diffusion scale, and the limiting behavior is qualitatively different, but can still be characterized. We refer to [Mukherjee, Borst, van Leeuwen, and Whiting \[2016a\]](#) for further details.

**4.3 Universality of JSQ(d) policies in infinite-server scenario.** As in [Section 3.5](#), we now further explore the trade-off between performance and communication overhead as a function of the diversity parameter  $d(N)$ , in conjunction with the relative load. We will specifically investigate what growth rate of  $d(N)$  is required, depending on the scaling



behavior of  $\lambda(N)$ , in order to asymptotically match the optimal performance of the JSQ policy.

**Theorem 4.1.** (Universality fluid limit for JSQ( $d(N)$ )) *If  $d(N) \rightarrow \infty$  as  $N \rightarrow \infty$ , then the fluid limit of the JSQ( $d(N)$ ) scheme coincides with that of the ordinary JSQ policy given by the dynamical system in (4-2). Consequently, the stationary occupancy states converge to the unique fixed point in (4-3).*

In order to state the universality result on diffusion scale, define in case  $f > 0$ ,  $f(N) := \lambda(N) - K(N)$ ,

$$\begin{aligned}\bar{Q}_i^{d(N)}(t) &:= \frac{N - Q_i^{d(N)}(t)}{\sqrt{N}} \quad (i \leq K), \\ \bar{Q}_{K+1}^{d(N)}(t) &:= \frac{Q_{K+1}^{d(N)}(t) - f(N)}{\sqrt{N}}, \\ \bar{Q}_i^{d(N)}(t) &:= \frac{Q_i^{d(N)}(t)}{\sqrt{N}} \geq 0 \quad (i \geq K+2),\end{aligned}$$

and otherwise, if  $f = 0$ , assume  $(KN - \lambda(N))/\sqrt{N} \rightarrow \beta \in \mathbb{R}$  as  $N \rightarrow \infty$ , and define

$$\begin{aligned}\hat{Q}_{K-1}^{d(N)}(t) &:= \sum_{i=1}^{K-1} \frac{N - Q_i^{d(N)}(t)}{\sqrt{N}}, \\ \hat{Q}_K^{d(N)}(t) &:= \frac{N - Q_K^{d(N)}(t)}{\sqrt{N}}, \\ \hat{Q}_i^{d(N)}(t) &:= \frac{Q_i^{d(N)}(t)}{\sqrt{N}} \geq 0 \quad (i \geq K+1).\end{aligned}$$

**Theorem 4.2** (Universality diffusion limit for JSQ( $d(N)$ )). *Assume*

$$d(N)/(\sqrt{N} \log N) \rightarrow \infty$$

*Under suitable initial conditions*

(i) *If  $f > 0$ , then  $\bar{Q}_i^{d(N)}(\cdot)$  converges to the zero process for  $i \neq K+1$ , and  $\bar{Q}_{K+1}^{d(N)}(\cdot)$  converges weakly to the Ornstein-Uhlenbeck process satisfying the SDE  $d\bar{Q}_{K+1}(t) = -\bar{Q}_{K+1}(t)dt + \sqrt{2\lambda}dW(t)$ , where  $W(\cdot)$  is the standard Brownian motion.*

(ii) *If  $f = 0$ , then  $\hat{Q}_{K-1}^{d(N)}(\cdot)$  converges weakly to the zero process, and  $(\hat{Q}_K^{d(N)}(\cdot), \hat{Q}_{K+1}^{d(N)}(\cdot))$  converges weakly to  $(\hat{Q}_K(\cdot), \hat{Q}_{K+1}(\cdot))$ , described by the unique*

solution of the following system of SDEs:

$$\begin{aligned} d\hat{Q}_K(t) &= \sqrt{2K}dW(t) - (\hat{Q}_K(t) + K\hat{Q}_{K+1}(t))dt + \beta dt + dV_1(t) \\ d\hat{Q}_{K+1}(t) &= dV_1(t) - (K+1)\hat{Q}_{K+1}(t)dt, \end{aligned}$$

where  $W(\cdot)$  is the standard Brownian motion, and  $V_1(\cdot)$  is the unique nondecreasing process satisfying  $\int_0^t \mathbb{1}_{[\hat{Q}_K(s) \geq 0]} dV_1(s) = 0$ .

Given the asymptotic results for the JSQ policy in [Sections 4.1](#) and [4.2](#), the proofs of the asymptotic results for the  $\text{JSQ}(d(N))$  scheme in [Theorems 4.1](#) and [4.2](#) involve establishing a universality result which shows that the limiting processes for the  $\text{JSQ}(d(N))$  scheme are ‘ $g(N)$ -alike’ to those for the ordinary JSQ policy for suitably large  $d(N)$ . Loosely speaking, if two schemes are  $g(N)$ -alike, then in some sense, the associated system occupancy states are indistinguishable on  $g(N)$ -scale.

The next theorem states a sufficient criterion for the  $\text{JSQ}(d(N))$  scheme and the ordinary JSQ policy to be  $g(N)$ -alike, and thus, provides the key vehicle in establishing the universality result.

**Theorem 4.3.** *Let  $g : \mathbb{N} \rightarrow \mathbb{R}_+$  be a function diverging to infinity. Then the JSQ policy and the  $\text{JSQ}(d(N))$  scheme are  $g(N)$ -alike, with  $g(N) \leq N$ , if (i)  $d(N) \rightarrow \infty$  for  $g(N) = O(N)$ , (ii)  $d(N) \left( \frac{N}{g(N)} \log \left( \frac{N}{g(N)} \right) \right)^{-1} \rightarrow \infty$  for  $g(N) = o(N)$ .*

The proof of [Theorem 4.3](#) relies on a novel coupling construction, called T-coupling (‘T’ stands for task-based), which will be used to (lower and upper) bound the difference of occupancy states of two arbitrary schemes. This T-coupling [Mukherjee, Borst, van Leeuwen, and Whiting \[2016a\]](#) is distinct from and inherently stronger than the S-coupling used in [Section 3.5](#) in the single-server queueing scenario. Note that in the current infinite-server scenario, the departures of the ordered server pools cannot be coupled, mainly since the departure rate at the  $m^{\text{th}}$  ordered server pool, for some  $m = 1, 2, \dots, N$ , depends on its number of active tasks. The T-coupling is also fundamentally different from the coupling constructions used in establishing the weak majorization results in [Winston \[1977\]](#), [Sparaggis, Towsley, and Cassandras \[1994\]](#), [Towsley \[1995\]](#), [Towsley, Sparaggis, and Cassandras \[1992\]](#), and [Weber \[1978\]](#) in the context of the ordinary JSQ policy in the single-server queueing scenario, and in [Sparaggis, Towsley, and Cassandras \[1993\]](#), [Johri \[1989\]](#), [Menich \[1987\]](#), and [Menich and Serfozo \[1991\]](#) in the scenario of state-dependent service rates.

## 5 Universality of load balancing in networks

In this section we return to the single-server queueing dynamics, and extend the universal properties to network scenarios, where the  $N$  servers are assumed to be inter-connected

by some underlying graph topology  $G_N$ . Tasks arrive at the various servers as independent Poisson processes of rate  $\lambda$ , and each incoming task is assigned to whichever server has the smallest number of tasks among the one where it arrives and its neighbors in  $G_N$ . Thus, in case  $G_N$  is a clique, each incoming task is assigned to the server with the shortest queue across the entire system, and the behavior is equivalent to that under the JSQ policy. The stochastic optimality properties of the JSQ policy thus imply that the queue length process in a clique will be better balanced and smaller (in a majorization sense) than in an arbitrary graph  $G_N$ .

Besides the prohibitive communication overhead discussed earlier, a further scalability issue of the JSQ policy arises when executing a task involves the use of some data. Storing such data for all possible tasks on all servers will typically require an excessive amount of storage capacity. These two burdens can be effectively mitigated in sparser graph topologies where tasks that arrive at a specific server  $i$  are only allowed to be forwarded to a subset of the servers  $\mathfrak{N}_i$ . For the tasks that arrive at server  $i$ , queue length information then only needs to be obtained from servers in  $\mathfrak{N}_i$ , and it suffices to store replicas of the required data on the servers in  $\mathfrak{N}_i$ . The subset  $\mathfrak{N}_i$  containing the peers of server  $i$  can be naturally viewed as its neighbors in some graph topology  $G_N$ . In this section we focus on the results in [Mukherjee, Borst, and van Leeuwaarden \[2017\]](#) for the case of undirected graphs, but most of the analysis can be extended to directed graphs.

The above model has been studied in [Gast \[2015\]](#) and [Turner \[1998\]](#), focusing on certain fixed-degree graphs and in particular ring topologies. The results demonstrate that the flexibility to forward tasks to a few neighbors, or even just one, with possibly shorter queues significantly improves the performance in terms of the waiting time and tail distribution of the queue length. This resembles the “power-of-choice” gains observed for JSQ( $d$ ) policies in complete graphs. However, the results in [Gast \[2015\]](#) and [Turner \[1998\]](#) also establish that the performance sensitively depends on the underlying graph topology, and that selecting from a fixed set of  $d - 1$  neighbors typically does not match the performance of re-sampling  $d - 1$  alternate servers for each incoming task from the entire population, as in the power-of- $d$  scheme in a complete graph.

If tasks do not get served and never depart but simply accumulate, then the scenario described above amounts to a so-called balls-and-bins problem on a graph. Viewed from that angle, a close counterpart of our setup is studied in [Kenthapadi and Panigrahy \[2006\]](#), where in our terminology each arriving task is routed to the shortest of  $d \geq 2$  randomly selected neighboring queues.

The key challenge in the analysis of load balancing on arbitrary graph topologies is that one needs to keep track of the evolution of number of tasks at each vertex along with their corresponding neighborhood relationship. This creates a major problem in constructing a tractable Markovian state descriptor, and renders a direct analysis of such processes highly intractable. Consequently, even asymptotic results for load balancing processes on

an arbitrary graph have remained scarce so far. The approach in [Mukherjee, Borst, and van Leeuwaarden \[2017\]](#) is radically different, and aims at comparing the load balancing process on an arbitrary graph with that on a clique. Specifically, rather than analyzing the behavior for a given class of graphs or degree value, the analysis explores for what types of topologies and degree properties the performance is asymptotically similar to that in a clique. The proof arguments in [Mukherjee, Borst, and van Leeuwaarden \[ibid.\]](#) build on the stochastic coupling constructions developed in [Section 3.5](#) for JSQ( $d$ ) policies. Specifically, the load balancing process on an arbitrary graph is viewed as a ‘sloppy’ version of that on a clique, and several other intermediate sloppy versions are constructed.

Let  $Q_i(G_N, t)$  denote the number of servers with queue length at least  $i$  at time  $t$ ,  $i = 1, 2, \dots$ , and let the fluid-scaled variables  $q_i(G_N, t) := Q_i(G_N, t)/N$  be the corresponding fractions. Also, in the Halfin-Whitt heavy-traffic regime (2-1), define the centered and diffusion-scaled variables  $\bar{Q}_1(G_N, t) := -(N - Q_1(G_N, t))/\sqrt{N}$  and  $\bar{Q}_i(G_N, t) := Q_i(G_N, t)/\sqrt{N}$  for  $i = 2, 3, \dots$ , analogous to (3-1).

The next definition introduces two notions of *asymptotic optimality*.

**Definition 5.1** (Asymptotic optimality). *A graph sequence  $\mathbf{G} = \{G_N\}_{N \geq 1}$  is called ‘asymptotically optimal on  $N$ -scale’ or ‘ $N$ -optimal’, if for any  $\lambda < 1$ , the scaled occupancy process  $(q_1(G_N, \cdot), q_2(G_N, \cdot), \dots)$  converges weakly, on any finite time interval, to the process  $(q_1(\cdot), q_2(\cdot), \dots)$  given by (3-5).*

*Moreover, a graph sequence  $\mathbf{G} = \{G_N\}_{N \geq 1}$  is called ‘asymptotically optimal on  $\sqrt{N}$ -scale’ or ‘ $\sqrt{N}$ -optimal’, if in the Halfin-Whitt heavy-traffic regime (2-1), on any finite time interval, the process  $(\bar{Q}_1(G_N, \cdot), \bar{Q}_2(G_N, \cdot), \dots)$  converges weakly to the process  $(\bar{Q}_1(\cdot), \bar{Q}_2(\cdot), \dots)$  given by (3-7).*

Intuitively speaking, if a graph sequence is  $N$ -optimal or  $\sqrt{N}$ -optimal, then in some sense, the associated occupancy processes are indistinguishable from those of the sequence of cliques on  $N$ -scale or  $\sqrt{N}$ -scale. In other words, on any finite time interval their occupancy processes can differ from those in cliques by at most  $o(N)$  or  $o(\sqrt{N})$ , respectively.

**5.1 Asymptotic optimality criteria for deterministic graph sequences.** We now develop a criterion for asymptotic optimality of an arbitrary deterministic graph sequence on different scales. We first introduce some useful notation, and two measures of *well-connectedness*. Let  $G = (V, E)$  be any graph. For a subset  $U \subseteq V$ , define  $\text{COM}(U) := |V \setminus N[U]|$  to be the set of all vertices that are disjoint from  $U$ , where  $N[U] := U \cup \{v \in V : \exists u \in U \text{ with } (u, v) \in E\}$ . For any fixed  $\varepsilon > 0$  define

$$(5-1) \quad \text{DIS}_1(G, \varepsilon) := \sup_{U \subseteq V, |U| \geq \varepsilon |V|} \text{COM}(U), \quad \text{DIS}_2(G, \varepsilon) := \sup_{U \subseteq V, |U| \geq \varepsilon \sqrt{|V|}} \text{COM}(U).$$

The next theorem provides sufficient conditions for asymptotic optimality on  $N$ -scale and  $\sqrt{N}$ -scale in terms of the above two well-connectedness measures.

**Theorem 5.2.** *For any graph sequence  $\mathbf{G} = \{G_N\}_{N \geq 1}$ , (i)  $\mathbf{G}$  is  $N$ -optimal if for any  $\varepsilon > 0$ ,  $\text{DIS}_1(G_N, \varepsilon)/N \rightarrow 0$  as  $N \rightarrow \infty$ . (ii)  $\mathbf{G}$  is  $\sqrt{N}$ -optimal if for any  $\varepsilon > 0$ ,  $\text{DIS}_2(G_N, \varepsilon)/\sqrt{N} \rightarrow 0$  as  $N \rightarrow \infty$ .*

The next corollary is an immediate consequence of [Theorem 5.2](#).

**Corollary 5.3.** *Let  $\mathbf{G} = \{G_N\}_{N \geq 1}$  be any graph sequence. Then (i) If  $d_{\min}(G_N) = N - o(N)$ , then  $\mathbf{G}$  is  $N$ -optimal, and (ii) If  $d_{\min}(G_N) = N - o(\sqrt{N})$ , then  $\mathbf{G}$  is  $\sqrt{N}$ -optimal.*

We now provide a sketch of the main proof arguments for [Theorem 5.2](#) as used in [Mukherjee, Borst, and van Leeuwen \[2017\]](#), focusing on the proof of  $N$ -optimality. The proof of  $\sqrt{N}$ -optimality follows along similar lines. First of all, it can be established that if a system is able to assign each task to a server in the set  $\mathcal{S}^N(n(N))$  of the  $n(N)$  nodes with shortest queues, where  $n(N)$  is  $o(N)$ , then it is  $N$ -optimal. Since the underlying graph is not a clique however (otherwise there is nothing to prove), for any  $n(N)$  not every arriving task can be assigned to a server in  $\mathcal{S}^N(n(N))$ . Hence, a further stochastic comparison property is proved in [Mukherjee, Borst, and van Leeuwen \[ibid.\]](#) implying that if on any finite time interval of length  $t$ , the number of tasks  $\Delta^N(t)$  that are not assigned to a server in  $\mathcal{S}^N(n(N))$  is  $o_P(N)$ , then the system is  $N$ -optimal as well. The  $N$ -optimality can then be concluded when  $\Delta^N(t)$  is  $o_P(N)$ , which is demonstrated in [Mukherjee, Borst, and van Leeuwen \[ibid.\]](#) under the condition that  $\text{DIS}_1(G_N, \varepsilon)/N \rightarrow 0$  as  $N \rightarrow \infty$  as stated in [Theorem 5.2](#).

**5.2 Asymptotic optimality of random graph sequences.** Next we investigate how the load balancing process behaves on random graph topologies. Specifically, we aim to understand what types of graphs are asymptotically optimal in the presence of randomness (i.e., in an average-case sense). [Theorem 5.4](#) below establishes sufficient conditions for asymptotic optimality of a sequence of inhomogeneous random graphs. Recall that a graph  $G' = (V', E')$  is called a supergraph of  $G = (V, E)$  if  $V = V'$  and  $E \subseteq E'$ .

**Theorem 5.4.** *Let  $\mathbf{G} = \{G_N\}_{N \geq 1}$  be a graph sequence such that for each  $N$ ,  $G_N = (V_N, E_N)$  is a super-graph of the inhomogeneous random graph  $G'_N$  where any two vertices  $u, v \in V_N$  share an edge with probability  $p_{uv}^N$ .*

- (i) *If  $\inf \{p_{uv}^N : u, v \in V_N\}$  is  $\omega(1/N)$ , then  $\mathbf{G}$  is  $N$ -optimal.*
- (ii) *If  $\inf \{p_{uv}^N : u, v \in V_N\}$  is  $\omega(\log(N)/\sqrt{N})$ , then  $\mathbf{G}$  is  $\sqrt{N}$ -optimal.*

The proof of [Theorem 5.4](#) relies on [Theorem 5.2](#). Specifically, if  $G_N$  satisfies conditions (i) and (ii) in [Theorem 5.4](#), then the corresponding conditions (i) and (ii) in [Theorem 5.2](#) hold.

As an immediate corollary to [Theorem 5.4](#) we obtain an optimality result for the sequence of Erdős–Rényi random graphs.

**Corollary 5.5.** *Let  $\mathbf{G} = \{G_N\}_{N \geq 1}$  be a graph sequence such that for each  $N$ ,  $G_N$  is a super-graph of  $\text{ER}_N(p(N))$ , and  $d(N) = (N - 1)p(N)$ . Then (i) If  $d(N) \rightarrow \infty$  as  $N \rightarrow \infty$ , then  $\mathbf{G}$  is  $N$ -optimal. (ii) If  $d(N)/(\sqrt{N} \log N) \rightarrow \infty$  as  $N \rightarrow \infty$ , then  $\mathbf{G}$  is  $\sqrt{N}$ -optimal.*

The growth rate condition for  $N$ -optimality in [Corollary 5.5](#) (i) is not only sufficient, but necessary as well. Thus informally speaking,  $N$ -optimality is achieved under the minimum condition required as long as the underlying topology is suitably random.

## 6 Token-based load balancing

While a zero waiting time can be achieved in the limit by sampling only  $d(N) = o(N)$  servers as [Sections 3.5, 4](#) and [5](#) showed, even in network scenarios, the amount of communication overhead in terms of  $d(N)$  must still grow with  $N$ . As mentioned earlier, this can be avoided by introducing memory at the dispatcher, in particular maintaining a record of only vacant servers, and assigning tasks to idle servers, if there are any, or to a uniformly at random selected server otherwise. This so-called Join-the-Idle-Queue (JIQ) scheme [Badonnel and Burgess \[2008\]](#) and [Lu, Xie, Kliot, Geller, Larus, and Greenberg \[2011\]](#) can be implemented through a simple token-based mechanism generating at most one message per task. Remarkably enough, even with such low communication overhead, the mean waiting time and the probability of a non-zero waiting time vanish under the JIQ scheme in both the fluid and diffusion regimes, as we will discuss in the next two sections.

**6.1 Asymptotic optimality of JIQ scheme.** We first consider the fluid limit of the JIQ scheme. Let  $q_i^N(\infty)$  be a random variable denoting the process  $q_i^N(\cdot)$  in steady state. It was proved in [Stolyar \[2015\]](#) for the JIQ scheme (under very broad conditions),

$$(6-1) \quad q_1^N(\infty) \rightarrow \lambda, \quad q_i^N(\infty) \rightarrow 0 \quad \text{for all } i \geq 2, \quad \text{as } N \rightarrow \infty.$$

The above equation in conjunction with the PASTA property yields that the steady-state probability of a non-zero wait vanishes as  $N \rightarrow \infty$ , thus exhibiting asymptotic optimality of the JIQ scheme on fluid scale.

We now turn to the diffusion limit of the JIQ scheme.

**Theorem 6.1.** (Diffusion limit for JIQ) *In the Halfin-Whitt heavy-traffic regime (2-1), under suitable initial conditions, the weak limit of the sequence of centered and diffusion-scaled occupancy process in (3-1) coincides with that of the ordinary JSQ policy given by the system of SDEs in (3-7).*

The above theorem implies that for suitable initial states, on any finite time interval, the occupancy process under the JIQ scheme is indistinguishable from that under the JSQ policy. The proof of [Theorem 6.1](#) relies on a coupling construction as described in greater detail in [Mukherjee, Borst, van Leeuwaarden, and Whiting \[2016b\]](#). The idea is to compare the occupancy processes of two systems following JIQ and JSQ policies, respectively. Comparing the JIQ and JSQ policies is facilitated when viewed as follows: (i) If there is an idle server in the system, both JIQ and JSQ perform similarly, (ii) Also, when there is no idle server and only  $O(\sqrt{N})$  servers with queue length two, JSQ assigns the arriving task to a server with queue length one. In that case, since JIQ assigns at random, the probability that the task will land on a server with queue length two and thus JIQ acts differently than JSQ is  $O(1/\sqrt{N})$ . Since on any finite time interval the number of times an arrival finds all servers busy is at most  $O(\sqrt{N})$ , all the arrivals except an  $O(1)$  of them are assigned in exactly the same manner in both JIQ and JSQ, which then leads to the same scaling limit for both policies.

**6.2 Multiple dispatchers.** So far we have focused on a basic scenario with a single dispatcher. Since it is not uncommon for LBAs to operate across multiple dispatchers though, we consider in this section a scenario with  $N$  parallel identical servers as before and  $R \geq 1$  dispatchers. (We will assume the number of dispatchers to remain fixed as the number of servers grows large, but a further natural scenario would be for the number of dispatchers  $R(N)$  to scale with the number of servers as considered by [Mitzenmacher \[2016\]](#), who analyzes the case  $R(N) = rN$  for some constant  $r$ , so that the relative load of each dispatcher is  $\lambda r$ .) Tasks arrive at dispatcher  $r$  as a Poisson process of rate  $\alpha_r \lambda N$ , with  $\alpha_r > 0$ ,  $r = 1, \dots, R$ ,  $\sum_{r=1}^R \alpha_r = 1$ , and  $\lambda$  denoting the task arrival rate per server. For conciseness, we denote  $\alpha = (\alpha_1, \dots, \alpha_R)$ , and without loss of generality we assume that the dispatchers are indexed such that  $\alpha_1 \geq \alpha_2 \geq \dots \geq \alpha_R$ .

When a server becomes idle, it sends a token to one of the dispatchers selected uniformly at random, advertising its availability. When a task arrives at a dispatcher which has tokens available, one of the tokens is selected, and the task is immediately forwarded to the corresponding server.

We distinguish two scenarios when a task arrives at a dispatcher which has no tokens available, referred to as the *blocking* and *queueing* scenario respectively. In the blocking scenario, the incoming task is blocked and instantly discarded. In the queueing scenario, the arriving task is forwarded to one of the servers selected uniformly at random. If the

selected server happens to be idle, then the outstanding token at one of the other dispatchers is revoked.

In the queueing scenario we assume  $\lambda < 1$ , which is not only necessary but also sufficient for stability. Denote by  $B(R, N, \lambda, \alpha)$  the steady-state blocking probability of an arbitrary task in the blocking scenario. Also, denote by  $W(R, N, \lambda, \alpha)$  a random variable with the steady-state waiting-time distribution of an arbitrary task in the queueing scenario.

Scenarios with multiple dispatchers have received limited attention in the literature, and the scant papers that exist [Lu, Xie, Kliot, Geller, Larus, and Greenberg \[2011\]](#), [Mitzenmacher \[2016\]](#), and [Stolyar \[2017\]](#) almost exclusively assume that the loads at the various dispatchers are strictly equal. In these cases the fluid limit, for suitable initial states, is the same as that for a single dispatcher, and in particular the fixed point is the same, hence, the JIQ scheme continues to achieve asymptotically optimal delay performance with minimal communication overhead. As one of the few exceptions, [van der Boor, Borst, and van Leeuwen \[2017b\]](#) allows the loads at the various dispatchers to be different.

**Results for blocking scenario.** For the blocking scenario, it is established in [van der Boor, Borst, and van Leeuwen \[ibid.\]](#) that,

$$B(R, N, \lambda, \alpha) \rightarrow \max\{1 - R\alpha_R, 1 - 1/\lambda\} \quad \text{as } N \rightarrow \infty.$$

This result shows that in the many-server limit the system performance in terms of blocking is either determined by the relative load of the least-loaded dispatcher, or by the aggregate load. This indirectly reveals that, somewhat counter-intuitively, it is the least-loaded dispatcher that throttles tokens and leaves idle servers stranded, thus acting as bottleneck.

**Results for queueing scenario.** For the queueing scenario, it is shown in [van der Boor, Borst, and van Leeuwen \[ibid.\]](#) that, for fixed  $\lambda < 1$

$$\mathbb{E}[W(R, N, \lambda, \alpha)] \rightarrow \frac{\lambda_2(R, \lambda, \alpha)}{1 - \lambda_2(R, \lambda, \alpha)} \quad \text{as } N \rightarrow \infty,$$

where  $\lambda_2(R, \lambda, \alpha) = 1 - \frac{1 - \lambda \sum_{i=1}^{r^*} \alpha_i}{1 - \lambda r^*/R}$ , with  $r^* = \sup \{r | \alpha_r > \frac{1}{R} \frac{1 - \lambda \sum_{i=1}^r \alpha_i}{1 - \lambda r/R}\}$ , may be interpreted as the rate at which tasks are forwarded to randomly selected servers.

When the arrival rates at all dispatchers are strictly equal, i.e.,  $\alpha_1 = \dots = \alpha_R = 1/R$ , the above results indicate that the stationary blocking probability and the mean waiting time asymptotically vanish as  $N \rightarrow \infty$ , which is in agreement with the observations in [Stolyar \[2017\]](#) mentioned above. However, when the arrival rates at the various dispatchers are not perfectly equal, so that  $\alpha_R < 1/R$ , the blocking probability and mean



waiting time are strictly positive in the limit, even for arbitrarily low overall load and an arbitrarily small degree of skewness in the arrival rates. Thus, the ordinary JIQ scheme fails to achieve asymptotically optimal performance for heterogeneous dispatcher loads.

In order to counter the above-described performance degradation for asymmetric dispatcher loads, [van der Boor, Borst, and van Leeuwaarden \[2017b\]](#) proposes two enhancements. Enhancement A uses a non-uniform token allotment: When a server becomes idle, it sends a token to dispatcher  $r$  with probability  $\beta_r$ . Enhancement B involves a token exchange mechanism: Any token is transferred to a uniformly randomly selected dispatcher at rate  $\nu$ . Note that the token exchange mechanism only creates a constant communication overhead per task as long as the rate  $\nu$  does not depend on the number of servers  $N$ , and thus preserves the scalability of the basic JIQ scheme.

The above enhancements can achieve asymptotically optimal performance for suitable values of the  $\beta_r$  parameters and the exchange rate  $\nu$ . Specifically, the stationary blocking probability in the blocking scenario and the mean waiting time in the queueing scenario asymptotically vanish as  $N \rightarrow \infty$ , upon using Enhancement A with  $\beta_r = \alpha_r$  or Enhancement B with  $\nu \geq \frac{\lambda}{1-\lambda}(\alpha_1 R - 1)$ .

## 7 Redundancy policies and alternative scaling regimes

In this section we discuss somewhat related redundancy policies and alternative scaling regimes and performance metrics.

**Redundancy- $d$  policies.** So-called redundancy- $d$  policies involve a somewhat similar operation as JSQ( $d$ ) policies, and also share the primary objective of ensuring low delays [Ananthanarayanan, Ghodsi, Shenker, and Stoica \[2013\]](#) and [Vulimiri, Godfrey, Mittal, Sherry, Ratnasamy, and Shenker \[2013\]](#). In a redundancy- $d$  policy,  $d \geq 2$  candidate servers are selected uniformly at random (with or without replacement) for each arriving task, just like in a JSQ( $d$ ) policy. Rather than forwarding the task to the server with the shortest queue however, replicas are dispatched to all sampled servers.

Two common options can be distinguished for abortion of redundant clones. In the first variant, as soon as the first replica starts service, the other clones are abandoned. In this case, a task gets executed by the server which had the smallest workload at the time of arrival (and which may or may not have had the shortest queue length) among the sampled servers. This may be interpreted as a power-of- $d$  version of the Join-the-Smallest Workload (JSW) policy discussed in [Section 2.5](#). In the second option the other clones of the task are not aborted until the first replica has completed service (which may or may not have been the first replica to start service). While a task is only handled by one of the servers in the former case, it may be processed by several servers in the latter case.

**Conventional heavy traffic.** It is also worth mentioning some asymptotic results for the classical heavy-traffic regime as described in [Section 2.2](#) where the number of servers  $N$  is fixed and the relative load tends to one in the limit. The papers [Foschini and Salz \[1978\]](#), [Reiman \[1984\]](#), and [Zhang, Hsu, and Wang \[1995\]](#) establish diffusion limits for the JSQ policy in a sequence of systems with Markovian characteristics as in our basic model set-up, but where in the  $K$ -th system the arrival rate is  $K\lambda + \hat{\lambda}\sqrt{K}$ , while the service rate of the  $i$ -th server is  $K\mu_i + \hat{\mu}_i\sqrt{K}$ ,  $i = 1, \dots, N$ , with  $\lambda = \sum_{i=1}^N \mu_i$ , inducing critical load as  $K \rightarrow \infty$ . It is proved that for suitable initial conditions the queue lengths are of the order  $O(\sqrt{K})$  over any finite time interval and exhibit a state-space collapse property.

[Atar, Keslassy, and Mendelson \[n.d.\]](#) investigate a similar scenario, and establish diffusion limits for three policies: the JSQ( $d$ ) policy, the redundancy- $d$  policy (where the redundant clones are abandoned as soon as the first replica starts service), and a combined policy called Replicate-to-Shortest-Queues (RSQ) where  $d$  replicas are dispatched to the  $d$ -shortest queues.

**Non-degenerate slowdown.** Asymptotic results for the so-called non-degenerate slowdown regime described in [Section 2.2](#) where  $N - \lambda(N) \rightarrow \gamma > 0$  as the number of servers  $N$  grows large, are scarce. [Gupta and Walton \[2017\]](#) characterize the diffusion-scaled queue length process under the JSQ policy in this asymptotic regime. They further compare the diffusion limit for the JSQ policy with that for a centralized queue as described above as well as several LBAs such as the JIQ scheme and a refined version called Idle-One-First (IIF), where a task is assigned to a server with exactly one task if no idle server is available and to a randomly selected server otherwise.

It is proved that the diffusion limit for the JIQ scheme is no longer asymptotically equivalent to that for the JSQ policy in this asymptotic regime, and the JIQ scheme fails to achieve asymptotic optimality in that respect, as opposed to the behavior in the large-capacity and Halfin-Whitt regimes discussed in [Section 2.7](#). In contrast, the IIF scheme does preserve the asymptotic equivalence with the JSQ policy in terms of the diffusion-scaled queue length process, and thus retains asymptotic optimality in that sense.

**Sparse-feedback regime.** As described in [Section 2.7](#), the JIQ scheme involves a communication overhead of at most one message per task, and yet achieves optimal delay performance in the fluid and diffusion regimes. However, even just one message per task may still be prohibitive, especially when tasks do not involve big computational tasks, but small data packets which require little processing.

Motivated by the above issues, [van der Boor, Borst, and van Leeuwen \[2017a\]](#) proposes a novel class of LBAs which also leverage memory at the dispatcher, but allow the communication overhead to be seamlessly adapted and reduced below that of the JIQ

scheme. Specifically, in the proposed schemes, the various servers provide occasional queue status notifications to the dispatcher, either in a synchronous or asynchronous fashion. The dispatcher uses these reports to maintain queue estimates, and forwards incoming tasks to the server with the lowest queue estimate. The results in [van der Boor, Borst, and van Leeuwaarden \[2017a\]](#) demonstrate that the proposed schemes markedly outperform JSQ( $d$ ) policies with the same number of  $d \geq 1$  messages per task and they can achieve a vanishing waiting time in the limit when the update frequency exceeds  $\lambda/(1 - \lambda)$ . In case servers only report zero queue lengths and suppress updates for non-zero queues, the update frequency required for a vanishing waiting time can in fact be lowered to just  $\lambda$ , matching the one message per task involved in the JIQ scheme.

**Scaling of maximum queue length.** So far we have focused on the asymptotic behavior of LBAs in terms of the number of servers with a certain queue length, either on fluid scale or diffusion scale, in various regimes as  $N \rightarrow \infty$ . A related but different performance metric is the maximum queue length  $M(N)$  among all servers as  $N \rightarrow \infty$ . [Luczak and McDiarmid \[2006\]](#) showed that for fixed  $d \geq 2$  the steady-state maximum queue length  $M(N)$  under the JSQ( $d$ ) policy is given by  $\log(\log(N))/\log(d) + O(1)$  and is concentrated on at most two adjacent values, whereas for purely random assignment ( $d = 1$ ), it scales as  $\log(N)/\log(1/\lambda)$  and does not concentrate on a bounded range of values. This is yet a further manifestation of the “power-of-choice” effect.

The maximum queue length  $M(N)$  is the central performance metric in balls-and-bins models where arriving items (balls) do not get served and never depart but simply accumulate in bins, and (stationary) queue lengths are not meaningful. In fact, the very notion of randomized load balancing and power-of- $d$  strategies was introduced in a balls-and-bins setting in the seminal paper by [Azar, Broder, Karlin, and Upfal \[1999\]](#).

## References

- Ganesh Ananthanarayanan, Ali Ghodsi, Scott Shenker, and Ion Stoica (2013). “Effective Straggler Mitigation: Attack of the Clones.” In: *NSDI*. Vol. 13, pp. 185–198 (cit. on p. 3938).
- R. Atar, I. Keslassy, and G. Mendelson (n.d.). *Randomized load balancing in heavy traffic*. Preprint (cit. on p. 3939).
- Rami Atar (2012). “A diffusion regime with nondegenerate slowdown”. *Oper. Res.* 60.2, pp. 490–500. MR: 2935073 (cit. on p. 3916).
- Yossi Azar, Andrei Z. Broder, Anna R. Karlin, and Eli Upfal (1999). “Balanced allocations”. *SIAM J. Comput.* 29.1, pp. 180–200. MR: 1710347 (cit. on p. 3940).

- Remi Badonnel and Mark Burgess (2008). “Dynamic pull-based load balancing for autonomic servers”. In: *Proc. IEEE/IFIP*. IEEE, pp. 751–754 (cit. on pp. [3914](#), [3919](#), [3935](#)).
- Mark van der Boor, Sem C. Borst, and Johan S. H. van Leeuwen (2017a). *Hyper-scalable JSQ with sparse feedback*. Preprint (cit. on pp. [3939](#), [3940](#)).
- (2017b). “Load balancing in large-scale systems with multiple dispatchers”. In: *INFOCOM 2017-IEEE Conference on Computer Communications, IEEE*. IEEE, pp. 1–9 (cit. on pp. [3937](#), [3938](#)).
- Amarjit Budhiraja and Eric Friedlander (June 2017). “Diffusion Approximations for Load Balancing Mechanisms in Cloud Storage Systems”. arXiv: [1706.09914](#) (cit. on p. [3925](#)).
- A. Ephremides, P. Varaiya, and J. Walrand (1980). “A simple dynamic routing problem”. *IEEE Trans. Automat. Control* 25.4, pp. 690–693. MR: [583444](#) (cit. on pp. [3912](#), [3917](#)).
- Patrick Eschenfeldt and David Gamarnik (Feb. 2015). “Join the Shortest Queue with Many Servers. The Heavy Traffic Asymptotics”. arXiv: [1502.00999](#) (cit. on p. [3924](#)).
- (Oct. 2016). “Supermarket Queueing System in the Heavy Traffic Regime. Short Queue Dynamics”. arXiv: [1610.03522](#) (cit. on p. [3925](#)).
- G Foschini and JACK Salz (1978). “A basic dynamic routing problem and diffusion”. *IEEE Transactions on Communications* 26.3, pp. 320–327 (cit. on p. [3939](#)).
- S. G. Foss and N. I. Chernova (2001). “On the optimality of the FCFS discipline in multi-server systems and queueing networks”. *Sibirsk. Mat. Zh.* 42.2, pp. 434–450, iii. MR: [1833168](#) (cit. on p. [3917](#)).
- David Gamarnik, John N Tsitsiklis, and Martin Zubeldia (2016). “Delay, memory, and messaging tradeoffs in distributed service systems”. In: *Proc. SIGMETRICS '16*, pp. 1–12 (cit. on pp. [3913](#), [3920](#)).
- Nicolas Gast (2015). “The Power of Two Choices on Graphs: the Pair-Approximation is Accurate”. In: *MAMA workshop '15* (cit. on p. [3932](#)).
- Varun Gupta and Neil Walton (July 2017). “Load Balancing in the Non-Degenerate Slow-down Regime”. arXiv: [1707.01969](#) (cit. on p. [3939](#)).
- Shlomo Halfin and Ward Whitt (1981). “Heavy-traffic limits for queues with many exponential servers”. *Oper. Res.* 29.3, pp. 567–588. MR: [629195](#) (cit. on p. [3916](#)).
- P. J. Hunt and T. G. Kurtz (1994). “Large loss networks”. *Stochastic Process. Appl.* 53.2, pp. 363–378. MR: [1302919](#) (cit. on p. [3923](#)).
- Pravin K. Johri (1989). “Optimality of the shortest line discipline with state-dependent service rates”. *European J. Oper. Res.* 41.2, pp. 157–161. MR: [1010313](#) (cit. on pp. [3928](#), [3931](#)).
- A. Karthik, Arpan Mukhopadhyay, and Ravi R. Mazumdar (2017). “Choosing among heterogeneous server clouds”. *Queueing Syst.* 85.1-2, pp. 1–29. MR: [3604116](#) (cit. on p. [3928](#)).

- Krishnamurthy Kenthapadi and Rina Panigrahy (2006). “Balanced allocation on graphs”. In: *Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithms*. ACM, New York, pp. 434–443. MR: [2368840](#) (cit. on p. [3932](#)).
- Yi Lu, Qiaomin Xie, Gabriel Kliot, Alan Geller, James R Larus, and Albert Greenberg (2011). “Join-Idle-Queue: A novel load balancing algorithm for dynamically scalable web services”. *Performance Evaluation* 68.11, pp. 1056–1071 (cit. on pp. [3914](#), [3919](#), [3935](#), [3937](#)).
- Malwina J. Luczak and Colin McDiarmid (2006). “On the maximum queue length in the supermarket model”. *Ann. Probab.* 34.2, pp. 493–527. MR: [2223949](#) (cit. on p. [3940](#)).
- Ronald Menich (1987). “Optimality of shortest queue routing for dependent service stations”. In: *Decision and Control, 1987. 26th IEEE Conference on*. Vol. 26. IEEE, pp. 1069–1072 (cit. on pp. [3928](#), [3931](#)).
- Ronald Menich and Richard F. Serfozo (1991). “Optimality of routing and servicing in dependent parallel processing systems”. *Queueing Systems Theory Appl.* 9.4, pp. 403–418. MR: [1137934](#) (cit. on pp. [3928](#), [3931](#)).
- Michael Mitzenmacher (2001). “The power of two choices in randomized load balancing”. *IEEE Transactions on Parallel and Distributed Systems* 12.10, pp. 1094–1104 (cit. on pp. [3913](#), [3918](#), [3922](#)).
- (2016). “Analyzing distributed Join-Idle-Queue: A fluid limit approach”. In: *Communication, Control, and Computing (Allerton), 2016 54th Annual Allerton Conference on*. IEEE, pp. 312–318 (cit. on pp. [3936](#), [3937](#)).
- Debankur Mukherjee, Sem C. Borst, and Johan S. H. van Leeuwaarden (July 2017). “Asymptotically Optimal Load Balancing Topologies”. arXiv: [1707.05866](#) (cit. on pp. [3932](#)–[3934](#)).
- Debankur Mukherjee, Sem C. Borst, Johan S. H. van Leeuwaarden, and Philip A. Whiting (Dec. 2016a). “Asymptotic Optimality of Power-of- $d$  Load Balancing in Large-Scale Systems”. arXiv: [1612.00722](#) (cit. on pp. [3928](#), [3929](#), [3931](#)).
- (2016b). “Universality of load balancing schemes on the diffusion scale”. *J. Appl. Probab.* 53.4, pp. 1111–1124. MR: [3581245](#) (cit. on p. [3936](#)).
  - (Dec. 2016c). “Universality of power-of- $d$  load balancing in many-server systems”. arXiv: [1612.00723](#) (cit. on pp. [3923](#), [3925](#), [3927](#)).
- Arpan Mukhopadhyay, A Karthik, Ravi R Mazumdar, and Fabrice Guillemin (2015). “Mean field and propagation of chaos in multi-class heterogeneous loss models”. *Performance Evaluation* 91, pp. 117–131 (cit. on pp. [3928](#), [3929](#)).
- Arpan Mukhopadhyay, Ravi R Mazumdar, and Fabrice Guillemin (2015). “The power of randomized routing in heterogeneous loss systems”. In: *Teletraffic Congress (ITC 27), 2015 27th International*. IEEE, pp. 125–133 (cit. on pp. [3928](#), [3929](#)).

- Martin I. Reiman (1984). “Some diffusion approximations with state space collapse”. In: *Modelling and performance evaluation methodology (Paris, 1983)*. Vol. 60. Lect. Notes Control Inf. Sci. Springer, Berlin, pp. 209–240. MR: [893658](#) (cit. on p. [3939](#)).
- Panayotis D. Sparaggis, D. Towsley, and C. G. Cassandras (1993). “Extremal properties of the shortest/longest nonfull queue policies in finite-capacity systems with state-dependent service rates”. *J. Appl. Probab.* 30.1, pp. 223–236. MR: [1206364](#) (cit. on pp. [3928](#), [3931](#)).
- Panayotis D. Sparaggis, Don Towsley, and Christos G. Cassandras (1994). “Sample path criteria for weak majorization”. *Adv. in Appl. Probab.* 26.1, pp. 155–171. MR: [1260308](#) (cit. on pp. [3927](#), [3931](#)).
- Alexander L. Stolyar (2015). “Pull-based load distribution in large-scale heterogeneous service systems”. *Queueing Syst.* 80.4, pp. 341–361. MR: [3367704](#) (cit. on pp. [3914](#), [3935](#)).
- (2017). “Pull-based load distribution among heterogeneous parallel servers: the case of multiple routers”. *Queueing Syst.* 85.1-2, pp. 31–65. MR: [3604117](#) (cit. on p. [3937](#)).
- D. Towsley (1995). “Application of majorization to control problems in queueing systems”. In: *Scheduling theory and its applications*. Ed. by P. Chr tienne, E. G. Coffman, J. K. Lenstra, and Z. Liu. Wiley, Chichester, pp. 295–311. MR: [1376619](#) (cit. on pp. [3927](#), [3931](#)).
- Don Towsley, Panayotis D. Sparaggis, and Christos G. Cassandras (1992). “Optimal routing and buffer allocation for a class of finite capacity queueing systems”. *IEEE Trans. Automat. Control* 37.9, pp. 1446–1451. MR: [1183112](#) (cit. on pp. [3927](#), [3931](#)).
- Stephen R. E. Turner (1998). “The effect of increasing routing choice on resource pooling”. *Probab. Engrg. Inform. Sci.* 12.1, pp. 109–124. MR: [1492143](#) (cit. on pp. [3928](#), [3932](#)).
- Ashish Vulimiri, Philip Brighten Godfrey, Radhika Mittal, Justine Sherry, Sylvia Ratnasamy, and Scott Shenker (2013). “Low latency via redundancy”. In: *Proceedings of the ninth ACM conference on Emerging networking experiments and technologies*. ACM, pp. 283–294 (cit. on p. [3938](#)).
- N. D. Vvedenskaya, R. L. Dobrushin, and F. I. Karpelevich (1996). “A queueing system with a choice of the shorter of two queues—an asymptotic approach”. *Problemy Peredachi Informatsii* 32.1, pp. 20–34. MR: [1384927](#) (cit. on pp. [3913](#), [3918](#), [3922](#)).
- Richard R. Weber (1978). “On the optimal assignment of customers to parallel servers”. *J. Appl. Probability* 15.2, pp. 406–413. MR: [0518586](#) (cit. on p. [3931](#)).
- Wayne Winston (1977). “Optimality of the shortest line discipline”. *J. Appl. Probability* 14.1, pp. 181–189. MR: [0428516](#) (cit. on pp. [3912](#), [3917](#), [3931](#)).
- Qiaomin Xie, Xiaobo Dong, Yi Lu, and Rayadurgam Srikant (2015). “Power of d choices for large-scale bin packing: A loss model”. *Proc. SIGMETRICS ’15* 43.1, pp. 321–334 (cit. on pp. [3928](#), [3929](#)).

- Lei Ying (2017). “Stein’s Method for Mean Field Approximations in Light and Heavy Traffic Regimes”. *Proceedings of the ACM on Measurement and Analysis of Computing Systems* 1.1, p. 12 (cit. on p. 3925).
- Hanqin Zhang, Guang-Hui Hsu, and Rongxin Wang (1995). “Heavy traffic limit theorems for a sequence of shortest queueing systems”. *Queueing Systems Theory Appl.* 21.1-2, pp. 217–238. MR: 1372056 (cit. on p. 3939).

Received 2017-11-30.

MARK VAN DER BOOR  
EINDHOVEN UNIVERSITY OF TECHNOLOGY, THE NETHERLANDS  
[m.v.d.boor@tue.nl](mailto:m.v.d.boor@tue.nl)

SEM C. BORST  
EINDHOVEN UNIVERSITY OF TECHNOLOGY, THE NETHERLANDS  
and  
NOKIA BELL LABS, MURRAY HILL, NJ, USA  
[s.c.borst@tue.nl](mailto:s.c.borst@tue.nl)

JOHAN S. H. VAN LEEUWAARDEN  
EINDHOVEN UNIVERSITY OF TECHNOLOGY, THE NETHERLANDS  
[j.s.h.v.leeuwaarden@tue.nl](mailto:j.s.h.v.leeuwaarden@tue.nl)

DEBANKUR MUKHERJEE  
EINDHOVEN UNIVERSITY OF TECHNOLOGY, THE NETHERLANDS  
[d.mukherjee@tue.nl](mailto:d.mukherjee@tue.nl)

# MATHEMATICAL MODELS OF COLLECTIVE DYNAMICS AND SELF-ORGANIZATION

PIERRE DEGOND

## Abstract

In this paper, we begin by reviewing a certain number of mathematical challenges posed by the modelling of collective dynamics and self-organization. Then, we focus on two specific problems, first, the derivation of fluid equations from particle dynamics of collective motion and second, the study of phase transitions and the stability of the associated equilibria.

**Data statement:** No new data were collected in the course of this research.

**Conflict of interest:** The authors declare that they have no conflict of interest.

## 1 Overview

Fascinating examples of collective motion can be observed in nature, such as insect swarms [Bazazi, Buhl, Hale, Anstey, Sword, Simpson, and Couzin \[2008\]](#) and [Khuong, Theraulaz, Jost, Perna, and Gautrais \[2011\]](#), bird flocks [Lukeman, Li, and Edelstein-Keshet \[2010\]](#), fish schools [Aoki \[1982\]](#), [Degond and Motsch \[2008b, 2011\]](#), [Domeier and Colin \[1997\]](#), [Gautrais, Jost, Soria, Campo, Motsch, Fournier, Blanco, and Theraulaz \[2009\]](#), and [Gautrais, Ginelli, Fournier, Blanco, Soria, Chaté, and Theraulaz \[2012\]](#), or in social phenomena, such as the spontaneous formation of lanes in pedestrian crowds [Moussaïd](#)

---

The author acknowledges support by the Engineering and Physical Sciences Research Council (EPSRC) under grants no. EP/M006883/1 and EP/P013651/1, by the Royal Society and the Wolfson Foundation through a Royal Society Wolfson Research Merit Award no. WM130048 and by the National Science Foundation (NSF) under grant no. RNMS11-07444 (KI-Net). PD is on leave from CNRS, Institut de Mathématiques de Toulouse, France. Works mentioned in this article have been realized in collaboration with many people. I wish to acknowledge more particularly A. Frouvelle, J-G. Liu, S. Merino-Aceituno, S. Motsch and A. Trescases for their decisive contributions.

*MSC2010:* primary 35Q92; secondary 82C22, 82C70, 92D50.

*Keywords:* Body attitude coordination, collective motion, Vicsek model, generalized collision invariant, rotation group, phase transitions, order parameter.



et al. [2012]. Similarly, at the microscopic scale, collective bacterial migration is frequently observed [Czirók, Ben-Jacob, Cohen, and Vicsek \[1996\]](#) and collective cell migration occurs during organism development [Shraiman \[2005\]](#) or healing [Poujade, Grasland-Mongrain, Hertzog, Jouanneau, Chavrier, Ladoux, Buguin, and Silberzan \[2007\]](#). Such systems of many autonomous agents locally interacting with each other are able to generate large-scale structures of sizes considerably exceeding the perception range of the agents. These large-scale structures are not directly encoded in the interaction rules between the individuals, which are usually fairly simple. They spontaneously emerge when a large number of individuals collectively interact [Vicsek and Zafeiris \[2012\]](#). This is referred to as “emergence”.

Emergence is a sort of bifurcation, or phase transition. In physics, phase transitions are dramatic changes of the system state consecutive to very small changes of some parameters, such as the temperature. In self-organized systems, the role of temperature is played by the noise level associated to the random component of the motion of the agents. For instance, in road traffic, the presence of drivers with erratic behavior can induce the formation of stop-and-go waves leading to a transition from fluid to congested traffic. Here, an increase of temperature (the random behavior of some agents) leads to a sudden blockage of the system. This is an example of the so-called “freezing-by-heating” phenomenon [Helbing, Farkas, and Vicsek \[2000\]](#) also observed in pedestrian crowds and a signature of the paradoxical and unconventional behavior of self-organized systems.

Another parameter which may induce phase transitions is the density of individuals. An increase of this density is very often associated with an increase of the order of the system [Vicsek, Czirók, Ben-Jacob, Cohen, and Shochet \[1995\]](#). For instance, the spontaneous lane formation in pedestrian crowds only appears when the density is high enough. This increase of order with the density is another paradoxical phenomenon in marked contrast with what is observed in more classical physical systems where an increase of density is generally associated with an increase of temperature, i.e. of disorder (this can be observed when pumping air into a bicycle tire: after using it, the pump core has heated up).

The passage between two different phases is called a critical state. In physical systems, critical states appear only for well-chosen ranges of parameters. For instance, at ambient pressure, liquid water passes to the gaseous state at the temperature of 100 °C. In self-organized systems, by contrast, critical states are extremely robust: they appear almost systematically, whatever the initial conditions of the system. In dynamical systems terms, the critical state is an attractor. The presence of critical states which are attractors of the dynamics is called “Self-Organized Criticality” [Bak, Tang, and Wiesenfeld \[1987\]](#) and its study is important in physics.

We shall focus on models of collective dynamics and self-organization that provide a prediction from an initial state of the system. These are stated as Cauchy problems for appropriate systems of differential equations. The modelling of self-organization meets important scientific and societal challenges. There are environmental and societal stakes: for instance, better understanding the behavior of a gregarious species can lead to improved conservation policies ; modelling human crowds improves the security, efficiency and profitability of public areas ; understanding collective cell migration opens new paradigms in cancer treatment or regenerative medicine. There are also technological stakes: robotists use social interaction mechanisms to geer fleets of robots or drones ; architects study social insect nests to look for new sustainable architecture ideas.

Large systems of interacting agents (aka particles) are modelled at different levels of detail. The most detailed models are particle models (aka individual-based or agent-based models). They describe the position and state of any single agent (particle) of the system as it evolves in time through its interactions with the other agents and the environment. This leads to large coupled systems of ordinary or stochastic differential equations (see an example in [Vicsek, Czirók, Ben-Jacob, Cohen, and Shochet \[1995\]](#)). When the number of particles is large, these systems are computationally intensive as their cost increases polynomially with the number of particles. Additionally their output is not directly exploitable as we are only interested in statistical averages (e.g. the pressure in a gas) and requires some post-processing which can generate errors.

For this reason, continuum models are often preferred [Toner and Tu \[1998\]](#). They consist of partial differential equations for averaged quantities such as the mean density or mean velocity of the agents. However, in the literature, a rigorous and systematic link between particle and continuum models is rarely found. Yet, establishing such a link is important. Indeed, often, the microscopic behavior of the agents is not well-known and is the actual target. On the other hand, large-scale structures are more easily accessible to experiments and can be used to calibrate continuum models. But to uncover the underlying individual behavior requires the establishment of a rigorous correspondence between the two types of models. Our goal is precisely to provide methodologies to establish this correspondence.

To derive continuum models from particle models rigorously requires a coarse-graining methodology. There are two steps of coarsening. The first step consists of deriving a “kinetic model”, which provides the time evolution of the probability distribution of the agents in position and state spaces. The equation for this kinetic distribution can be derived from the particle model, however not in closed form unless one assumes a strong hypothesis named “propagation of chaos” which means statistical independence between the particles. This hypothesis is generally wrong but admittedly, becomes asymptotically valid as the particle number tends to infinity. To prove such a result is a very difficult task and until recently [Gallagher, Saint-Raymond, and Texier \[2013\]](#) and [Mischler and](#)

[Mouhot \[2013\]](#), the only available result one was due to Lanford for the Boltzmann model [Lanford \[1976\]](#). Kinetic models are differential or integro-differential equations posed on a large dimensional space such as the Boltzmann or Fokker-Planck equations.

The second step of coarsening consists of reducing the description of the system to a few macroscopic averages (or moments) such as the density or the mean velocity as functions of position and time. The resulting fluid models are systems of nonlinear partial differential equations such as the Euler or Navier-Stokes equations. Fluid models are derived by averaging out the state variable of kinetic models (such as the particle velocity) to only keep track of the spatio-temporal dependence. Here again, a closure assumption is needed, by which one postulates a known shape of the distribution function as functions of its fluid moments. It can be justified in the hydrodynamic regime when the kinetic phenomena precisely bring the distribution function close to the postulated one. Providing a rigorous framework to these approaches is the core subject of “kinetic theory”, whose birthdate is the statement of his 6th problem by Hilbert in his 1900 ICM address. Since then, kinetic theory has undergone impressive developments, with Field’s medals awarded to P. L. Lions and C. Villani for works in this theory.

It is therefore appealing to apply kinetic theory methods to collective dynamics and self-organization. However, this has proved more delicate than anticipated and fascinating new mathematical questions have emerged from these difficulties. A first difficulty is that kinetic models may lose validity as propagation of chaos may simply be not true. Indeed, self-organization supposes the build-up of correlations between the particles. It is not clear that these correlations disappear with the number of particles tending to infinity. We have indeed proved (with E. Carlen and B. Wennberg [Carlen, Degond, and Wennberg \[2013\]](#)) in a simple collective dynamics model that propagation of chaos may break down at large temporal scales. Are there new models that can replace the defective kinetic equations when propagation of chaos breaks down ? Some phenomenological answers have been proposed but to the best of our knowledge, no mathematical theory is available yet.

A second difficulty arises at the passage between kinetic and fluid models. In classical physics, a fundamental concept is that of conservation law (such as mass, momentum or energy conservations). These conservation laws are satisfied at particle level and so, are transferred to the macroscopic scale and serve as corner stone in the derivation of fluid equations. By contrast, biological or social systems are open systems which exchange momentum and energy with the outside world and have no reason to satisfy such conservation laws. This is a major difficulties as acknowledged in Vicsek’s review [Vicsek and Zafeiris \[2012\]](#). In a series of works initiated in [Degond and Motsch \[2008a\]](#), we have overcome this problem and shown that some weaker conservation laws which we named “generalized collision invariants (GCI)” prevail. They enabled us to derive fluid models showing new and intriguing properties. Their mathematical study is still mostly open. We will provide more details in [Section 2](#).

The third difficulty is linked to the ubiquity of phase transitions in self-organized systems. This puts a strong constraint on fluid models which must be able to correctly describe the various phases and their interfaces. Complex phenomena like hysteresis [Couzin, Krause, James, Ruxton, and Franks \[2002\]](#), which results from the presence of multiple stable equilibria and involves the time-history of the system, must also be correctly rendered. However, different phases are described by types of fluid models. For instance, in symmetry-breaking phase transitions, the disordered phase is described by a parabolic equation while the ordered phase is described by a hyperbolic equation [Degond, Frouvelle, and Liu \[2013, 2015\]](#). At the critical state, these two phases co-exist and should be related by transmission conditions through phase boundaries. These transmission conditions are still unknown. More about phase transitions can be found in [Section 3](#) and references [Barbero and Degond \[2014\]](#) and [Frouvelle and Liu \[2012\]](#). Convergence to swarming states for the Cucker-Smale model [Cucker and Smale \[2007\]](#) has been extensively studied in the mathematical literature [Carrillo, Fornasier, Rosado, and Toscani \[2010\]](#), [Ha and Liu \[2009\]](#), [Ha and Tadmor \[2008\]](#), [Motsch and Tadmor \[2011\]](#), and [Shen \[2007/08\]](#), as well as for related models [Chuang, D’Orsogna, Marthaler, Bertozzi, and Chayes \[2007\]](#).

We have used symmetry-breaking phase transitions in a surprising context: to design automatized fertility tests for ovine sperm samples [Creppy, Plouraboué, Praud, Druart, Cazin, Yu, and Degond \[2016\]](#). Other types of phase transition play important roles. One of them is the packing transition which occurs when finite size particles reach densities at which they are in contact with each other. This transition occurs for instance in cancer tumors [Leroy-Lerêtre, Dimarco, Cazales, Boizeau, Ducommun, Lobjois, and Degond \[2017\]](#), crowds [Degond and Hua \[2013\]](#) and [Degond, Hua, and Navoret \[2011\]](#), road traffic [Berthelin, Degond, Delitala, and Rascle \[2008\]](#), herds [Degond, Navoret, Bon, and Sanchez \[2010\]](#) or tissue self-organization [Peurichard, Delebecque, Lorsignol, Barreau, Rouquette, Descombes, Casteilla, and Degond \[2017\]](#). Another example is the transition from a continuum to a network, and is at play for instance in the emergence of ant-trail networks [Boissard, Degond, and Motsch \[2013\]](#) and [Haskovec, Markowich, Perthame, and Schlottbom \[2016\]](#). For such systems, many challenges remain such as the derivation of macroscopic models.

In the forthcoming sections, we will focus on two specific aspects: the derivation of fluid models in spite of the lack of conservations relations ([Section 2](#)) and the investigation of phase transitions ([Section 3](#)).

## 2 Derivation of fluid models

**2.1 The Vicsek model.** We start with the description of particle models of collective behavior. As an example, we introduce the Vicsek model [Vicsek, Czirók, Ben-Jacob, Cohen, and Shochet \[1995\]](#) (see related models in [Bertin, Droz, and Grégoire \[2009\]](#), [Degond, Manhart, and Yu \[2017\]](#), and [Ginelli, Peruani, Bär, and Chaté \[2010\]](#)). It considers systems of self-propelled particles moving with constant speed (here supposed equal to 1 for notational simplicity) and interacting with their neighbors through local alignment. Such a model describes the dynamics of bird flocks and fish schools [Vicsek and Zafeiris \[2012\]](#). It is written in the form of the following stochastic differential system:

$$(2-1) \quad dX_i(t) = V_i(t)dt,$$

$$(2-2) \quad dV_i(t) = P_{V_i(t)^\perp} \circ (F_i(t) dt + \sqrt{2\tau} dB_t^i),$$

$$(2-3) \quad F_i(t) = \nu U_i(t), \quad U_i(t) = \frac{J_i(t)}{|J_i(t)|}, \quad J_i(t) = \sum_{j \mid |X_j(t) - X_i(t)| \leq R} V_j(t).$$

Here,  $X_i(t) \in \mathbb{R}^d$  is the position of the  $i$ -th particle (with  $i \in \{1, \dots, N\}$ ),  $V_i(t) \in \mathbb{S}^{d-1}$  is its velocity direction.  $B_t^i$  are standard independent Brownian motions in  $\mathbb{R}^d$  describing idiosyncratic noise i.e. noise specific to each agent and  $\sqrt{2\tau}$  is a constant and uniform noise intensity.  $F_i$  is the alignment force acting on the particles: it is proportional to the mean orientation  $U_i(t) \in \mathbb{S}^{d-1}$  of the agents around agent  $i$ , with a constant and uniform multiplication factor  $\nu$  encoding the alignment force intensity.  $U_i(t)$  itself is obtained by normalizing the total momentum  $J_i(t)$  of the agents belonging to a ball of radius  $R$  centered at the position  $X_i(t)$  of agent  $i$ . The normalization of  $J_i(t)$  (i.e. its division by  $|J_i(t)|$  where  $|\cdot|$  denotes the euclidean norm) makes only sense if  $J_i(t) \neq 0$ , which we assume here. The projection  $P_{V_i(t)^\perp}$  onto  $\{V_i(t)\}^\perp$  is there to maintain  $V_i(t)$  of unit norm and is a matrix given by  $P_{V_i^\perp} = \text{Id} - V_i \otimes V_i$  where  $\text{Id}$  is the identity matrix of  $\mathbb{R}^d$  and  $\otimes$  denotes the tensor product. The Stochastic Differential [Equation \(2-2\)](#) is understood in the Stratonovich sense, hence the symbol  $\circ$ , so that the noise term provides a Brownian motion on the sphere  $\mathbb{S}^{d-1}$  [Hsu \[2002\]](#). [Equation \(2-2\)](#) models two antagonist effects acting on the particles: the alignment force (the first term) which has a focusing effect and the noise (the second term) which has a defocusing effect. The original model proposed in [Vicsek, Czirók, Ben-Jacob, Cohen, and Shochet \[1995\]](#) is a time-discretized variant of this model.

Next, we present the kinetic model corresponding to this discrete model. It is written:

$$(2-4) \quad \partial_t f + \nabla_x \cdot (vf) = \nabla_v \cdot \left( - (P_{v\perp} F_f) f + \tau \nabla_v f \right),$$

$$(2-5) \quad F_f(x, t) = v U_f(x, t), \quad U_f(x, t) = \frac{J_f(x, t)}{|J_f(x, t)|},$$

$$(2-6) \quad J_f(x, t) = \int_{|y-x| \leq R} \int_{\mathbb{S}^{d-1}} f(y, w, t) w dw dy,$$

where  $f = f(x, v, t)$  is the particle distribution function and is a function of the position  $x \in \mathbb{R}^d$ , velocity  $v \in \mathbb{S}^{d-1}$  and time  $t > 0$ ,  $\nabla_v$  stands for the nabla operator on the sphere  $\mathbb{S}^{d-1}$  and  $P_{v\perp}$  is the projection operator on  $\{v\}^\perp$ .  $f(x, v, t)$  represents the probability density of particles in the  $(x, v)$  space. The left-hand side of (2-4) describes motion of the particles in physical space with speed  $v$ , while the right-hand side models the contributions of the alignment force  $F_f$  and of velocity diffusion (with diffusion coefficient  $\tau$ ) induced by Brownian noise at the particle level. The construction of the force term follows the same principles as for the discrete model, with  $F_f(x, t)$ ,  $U_f(x, t)$ ,  $J_f(x, t)$  replacing  $F_i(t)$ ,  $U_i(t)$ ,  $J_i(t)$ . The sum of the velocities over neighboring particles in the computation of the momentum (2-3) is replaced by integrals of the velocity weighted by  $f$ , with spatial integration domain being the ball centered at  $x$  and of radius  $R$ , and velocity integration domain being the whole sphere  $\mathbb{S}^{d-1}$  (Equation (2-6)). Analysis of this model can be found in Figalli, Kang, and Morales [2018] and Gamba and Kang [2016]. The passage from (2-1)-(2-3) to (2-4)-(2-5) is shown in Bolley, Cañizo, and Carrillo [2012], in the variant where  $J_i(t)$  is directly used in (2-2) instead of  $F_i(t)$ . In the case presented here, the control of  $J_i(t)$  away from zero presents additional difficulties which haven't been solved yet.

The macroscopic equations describe a large spatio-temporal scale regime. This regime is modelled by a time and space rescaling in (2-4)-(2-5) involving a small parameter  $\varepsilon \ll 1$  describing the ratio between the micro and the macro scales, which leads to

$$(2-7) \quad \varepsilon(\partial_t f^\varepsilon + \nabla_x \cdot (vf^\varepsilon)) = \nabla_v \cdot \left( - (P_{v\perp} F_{f^\varepsilon}) f^\varepsilon + \tau \nabla_v f^\varepsilon \right),$$

$$(2-8) \quad F_{f^\varepsilon}(x, t) = v u_f(x, t), \quad u_f(x, t) = \frac{j_f(x, t)}{|j_f(x, t)|},$$

$$(2-9) \quad j_f(x, t) = \int_{\mathbb{S}^{d-1}} f(x, w, t) w dw.$$

The scale change brings a factor  $\varepsilon$  in front of the terms at the left-hand side of (2-7) describing the motion of the particles in position space. It also localizes the integral describing the momentum of particles which now only involves an integration with respect to the velocity  $w$  of the distribution at the same location  $x$  as the particle onto which the force

applies (see Equation (2-8)). This is due to the interaction radius  $R$  being of order  $\varepsilon$  in this regime. The expansion of  $J_f$  in powers of  $\varepsilon$  leads to (2-8) up to terms of order  $\varepsilon^2$  which are neglected here as not contributing to the final macroscopic model at the end. The macroscopic model is obtained as the limit  $\varepsilon \rightarrow 0$  of this perturbation problem.

Before stating the result, we introduce the “von Mises Fisher (VMF)” distribution of orientation  $u$  and concentration parameter  $\kappa$  where  $u$  is an arbitrary vector in  $\mathbb{S}^{d-1}$  and  $\kappa \in [0, \infty)$ . This distribution denoted by  $M_{\kappa u}$  is such that for all  $v \in \mathbb{S}^{d-1}$ :

$$(2-10) \quad M_{\kappa u}(v) = \frac{1}{Z} \exp(\kappa u \cdot v),$$

where  $u \cdot v$  is the euclidean inner product of  $u$  and  $v$  and  $Z$  is a normalization constant only depending on  $\kappa$ . In Degond and Motsch [2008a], we proved the following formal theorem

**Theorem 2.1.** *If the solution  $f^\varepsilon$  of (2-7), (2-8) has a limit  $f^0$  when  $\varepsilon \rightarrow 0$ , it is given by*

$$(2-11) \quad f^0(x, v, t) = \rho(x, t) M_{\kappa u(x, t)}(v),$$

where  $\kappa = v/\tau$  and the pair  $(\rho, u)$  satisfies the following “self-organized hydrodynamic” (SOH) model:

$$(2-12) \quad \partial_t \rho + c_1 \nabla_x \cdot (\rho u) = 0,$$

$$(2-13) \quad \rho(\partial_t u + c_2(u \cdot \nabla_x u)) + \tau P_{u^\perp} \nabla_x \rho = 0,$$

$$(2-14) \quad |u| = 1,$$

with the coefficients  $c_1, c_2$  depending on  $v$  and  $\tau$  and  $P_{u^\perp}$  being the projection onto  $\{u\}^\perp$ .

The VMF distribution provides a way to extend the concept of Gaussian distribution to statistical distributions defined on the sphere. The orientation  $u$  describes the mean orientation of the particles while  $1/\kappa$  measures the dispersion of the particles around this mean. When  $\kappa$  is close to zero, the VMF is close to a uniform distribution while when it is large, it is close to a Dirac delta at  $u$ . The theorem states that at large scales, the distribution function approaches a VMF distribution weighted by the local density  $\rho$ . However, both  $\rho$  and the orientation  $u$  of the VMF depend on position and space and they are determined by solving the SOH model.

The SOH model is akin to the compressible Euler equations of gas dynamics, but with some important differences. First, the mean orientation  $u$  is constrained to lie on the sphere as (2-14) shows. The presence of the projection  $P_{u^\perp}$  in (2-13) guarantees that it is the case as soon as the initial orientation  $u|_{t=0}$  belongs to the sphere. The presence of  $P_{u^\perp}$

makes the system belong to the class of non-conservative hyperbolic problems, which are notoriously difficult (we can show that the model is hyperbolic). Finally, the convection terms in the two equations are multiplied by different coefficients  $c_1 \neq c_2$ , while they are the same in standard gas dynamics. This is a signature of a non-Galilean invariant dynamics. Indeed, as the particles are supposed to move with speed 1, there is a preferred frame in which this speed is measured. In any other Galilean frame this property will be lost. The mathematical properties of the SOH model are open, except for a local existence result in [Degond, Liu, Motsch, and Panferov \[2013\]](#). A rigorous proof of [Theorem 2.1](#) has been given in [Jiang, Xiong, and Zhang \[2016\]](#).

To understand how [Theorem 2.1](#) can be proved, we write (2-7) as

$$(2-15) \quad \partial_t f^\varepsilon + \nabla_x \cdot (v f^\varepsilon) = \frac{1}{\varepsilon} Q(f^\varepsilon)$$

$$(2-16) \quad Q(f) = \nabla_v \cdot \left( - (P_{v^\perp} F_f) f + \tau \nabla_v f \right),$$

with  $F_f$  given by (2-8), (2-9). It is readily seen that  $Q(f)$  can be written as

$$(2-17) \quad Q(f) = \mathbb{Q}(f; u_f),$$

where  $u_f$  is the mean orientation associated with  $f$  and is given by (2-8) and where for any  $u \in \mathbb{S}^{d-1}$ ,

$$(2-18) \quad \mathbb{Q}(f; u)(v) = \tau \nabla_v \cdot \left( M_{ku}(v) \nabla_v \left( \frac{f(v)}{M_{ku}(v)} \right) \right).$$

We note that for a given  $u \in \mathbb{S}^{d-1}$ , the operator  $\mathbb{Q}(\cdot; u)$  is linear. However, this is not the linearization of  $Q$  around  $\rho M_{ku}$  as extra terms coming from the variation of  $u_f$  with respect to  $f$  would appear.

By formally letting  $\varepsilon \rightarrow 0$  in (2-15), we get that  $f^0$  is a solution of  $Q(f^0) = 0$ . It is an easy matter to show that this implies the existence of two functions  $\rho(x, t)$  and  $u(x, t)$  with values in  $[0, \infty)$  and  $\mathbb{S}^{d-1}$  respectively such that (2-11) holds. Indeed, from (2-18) and Green's formula, we get

$$(2-19) \quad \int \mathbb{Q}(f; u)(v) \frac{f(v)}{M_{ku}(v)} dv = -d \int M_{ku}(v) \left| \nabla_v \left( \frac{f(v)}{M_{ku}(v)} \right) \right|^2 dv \leq 0.$$

Therefore, if  $\mathbb{Q}(f; u) = 0$ , this implies that  $\frac{f(v)}{M_{ku}(v)}$  does not depend on  $v$ . The result follows easily.

To find the equations satisfied by  $\rho$  and  $u$ , it is necessary to remove the  $1/\varepsilon$  singularity in (2-15), i.e. to project the equation on the slow manifold. In gas dynamics, this is done



by using the conservations of mass, momentum and energy. Here, the model only enjoys conservation of mass, which is expressed by the fact that

$$(2-20) \quad \int Q(f) dv = 0, \quad \forall f.$$

Hence, integrating (2-15) with respect to  $v$  and using (2-20), we get that

$$(2-21) \quad \partial_t \rho_{f^\varepsilon} + \nabla_x \cdot j_{f^\varepsilon} = 0.$$

Letting  $\varepsilon \rightarrow 0$ , with (2-11), we get

$$(2-22) \quad \rho_{f^\varepsilon} \rightarrow \rho, \quad j_{f^\varepsilon} \rightarrow j_{f^0} = c_1 \rho u,$$

where  $c_1$  is the so called order-parameter and is given by

$$(2-23) \quad c_1 = c_1(\kappa) = \int M_{\kappa u}(v) (v \cdot u) dv..$$

This leads to (2-12).

We need another equation to find  $u$ . In gas dynamics, this is done by using momentum conservation, which in this context would be expressed by  $\int Q(f) v dv = 0$ . However, this equation is not true and the lack of momentum conservation relates to the particles being self-propelled and therefore, able to extract or release momentum from the underlying medium. However, in Degond and Motsch [2008a], I showed that weaker forms of conservations (named generalized collision invariants or GCI) hold and provide the missing equation.

More precisely, we define

**Definition 2.2.** For a given orientation  $u \in \mathbb{S}^{d-1}$ , we define a GCI associated with  $u$  as a function  $\psi(v)$  such that

$$(2-24) \quad \int Q(f; u)(v) \psi(v) dv = 0, \quad \forall f \text{ such that } P_{u^\perp} j_f = 0.$$

By restricting the set of  $f$  to which we request the conservations to apply, we enlarge the set of candidate GCI  $\psi$ . In Degond and Motsch [ibid.] (see also Frouvelle [2012]), we show that the following theorem:

**Theorem 2.3.** The set  $\mathcal{C}_u$  of GCI associated to a given orientation  $u$  is a linear vector space of dimension  $d$  expressed as follows:

$$(2-25) \quad \mathcal{C}_u = \{C + A \cdot P_{u^\perp} v h(u \cdot v) \mid C \in \mathbb{R}, A \in \{u\}^\perp\}.$$

Here, defining  $\theta$  by  $\cos \theta = u \cdot v$ ,  $h$  is given by

$$(2-26) \quad h(\cos \theta) = \frac{g(\theta)}{\sin \theta}, \quad \theta \in (0, \pi),$$

with  $g$  being the unique solution of the elliptic problem

$$(2-27) \quad -\frac{d}{d\theta} \left( \sin^{d-2} \theta e^{\kappa \cos \theta} \frac{dg}{d\theta} \right) + (d-2) \sin^{d-4} \theta e^{\kappa \cos \theta} g = \sin^{d-1} \theta e^{\kappa \cos \theta}$$

in the space

$$(2-28) \quad V = \{g \mid (d-2) \sin^{\frac{d}{2}-2} \theta g \in L^2(0, \pi), \quad \sin^{\frac{d}{2}-1} \theta g \in H_0^1(0, \pi)\}.$$

We recall that  $L^2(0, \pi)$  is the Lebesgue space of square-integrable functions on  $(0, \pi)$  and  $H_0^1(0, \pi)$  is the Sobolev space of functions which are in  $L^2(0, \pi)$  and whose first order derivative is in  $L^2(0, \pi)$  and which vanish at 0 and  $\pi$ .

The GCI have the remarkable property that

$$(2-29) \quad \int Q(f) P_{u_f^\perp} v h(u_f \cdot v) dv = 0, \quad \forall f.$$

Indeed,  $P_{u_f^\perp} v h(u_f \cdot v)$  is a GCI  $\psi$  associated with  $u_f$ . Thus, using (2-17), and the Equation (2-24) of GCI, we get

$$\int Q(f) \psi(v) dv = \int \mathbb{Q}(f, u_f) \psi(v) dv = 0,$$

as  $P_{u_f^\perp} j_f = |j_f| P_{u_f^\perp} u_f = 0$ . Multiplying (2-15) by  $P_{u_{f^\varepsilon}^\perp} v h(u_{f^\varepsilon} \cdot v)$ , applying (2-29) with  $f = f^\varepsilon$  to cancel the right-hand side of the resulting equation, letting  $\varepsilon \rightarrow 0$  and using (2-11), we get:

$$(2-30) \quad P_{u^\perp} \int (\partial_t + v \cdot \nabla_x)(\rho M_{ku}) h(u \cdot v) v dv = 0.$$

After some computations, this equation gives rise to (2-13), where the constant  $c_2$  depends on a suitable moment of the function  $h$ .

The GCI concept has provided a rigorous way to coarse-grain a large class of KM sharing similar structures Degond, Frouvelle, and Merino-Aceituno [2017], Degond, Manhart, and Yu [2017], and Degond and Motsch [2011]. As an example, we now consider the model of Degond, Frouvelle, and Merino-Aceituno [2017] and Degond, Frouvelle, Merino-Aceituno, and Trescases [2018] where self-propelled agents try to coordinate their full body attitude. This model is described in the next section.

**2.2 A new model of full body attitude alignment.** The microscopic model considers  $N$  agents with positions  $X_i(t) \in \mathbb{R}^3$  and associated rotation matrices  $A_i(t) \in \text{SO}(3)$  representing the rotation needed to map a fixed reference frame  $(e_1, e_2, e_3)$  to the local frame  $(A_i(t) e_1, A_i(t) e_2, A_i(t) e_3)$  attached to the body of agent  $i$  at time  $t$ . As the particles are self-propelled, agent  $i$  moves in the direction  $A_i(t) e_1$  with unit speed. Agents try to coordinate their body attitude with those of their neighbors. Following these principles, the particle model is written:

$$(2-31) \quad dX_i(t) = A_i(t) e_1 dt,$$

$$(2-32) \quad dA_i(t) = P_{T_{A_i(t)}} \circ (F_i(t) dt + 2\sqrt{\tau} dB_t^i), \quad F_i(t) = \nu \Lambda_i(t),$$

$$(2-33) \quad \Lambda_i(t) = \text{PD}(G_i(t)), \quad G_i(t) = \sum_{j \mid |X_j(t) - X_i(t)| \leq R} A_j(t).$$

Here,  $B_t^i$  are standard independent Brownian motions in the linear space of  $3 \times 3$  matrices (in which  $\text{SO}(3)$  is isometrically imbedded) describing idiosyncratic noise and  $2\sqrt{\tau}$  is the noise intensity.  $F_i$  is the force that aligns the body attitude of Agent  $i$  to the mean body attitude of the neighbors defined by  $\Lambda_i(t)$  with a force intensity  $\nu$ .  $\Lambda_i(t)$  is obtained by normalizing the matrix  $G_i(t)$  constructed as the sum of the rotation matrices of the neighbors in a ball of radius  $R$  centered at the position  $X_i(t)$  of Agent  $i$ . The normalization is obtained by using the polar decomposition of matrices. We suppose that  $G_i(t)$  is non-singular. Then there exists a unique rotation matrix  $\text{PD}(G_i(t))$  and a unique symmetric matrix  $S_i(t)$  such that  $G_i(t) = \text{PD}(G_i(t)) S_i(t)$ . The quantity  $P_{T_{A_i(t)}}$  denotes the orthogonal projection onto the tangent space  $T_{A_i(t)}$  to  $\text{SO}(3)$  at  $A_i(t)$  to guarantee that the dynamics maintains  $A_i(t)$  on  $\text{SO}(3)$ . The Stochastic Differential Equation (2-32) is again understood in the Stratonovich sense, using the symbol  $\circ$  to highlight this fact. As a consequence, the noise term provides a Brownian motion on  $\text{SO}(3)$  as shown in [Hsu \[2002\]](#). Note however that the noise intensity is  $2\sqrt{\tau}$  instead of  $\sqrt{2\tau}$  as before. This is because we endow  $\text{SO}(3)$  with the inner product  $A \cdot B = \frac{1}{2} \text{Tr}(A^T B)$ , where  $\text{Tr}$  stands for the trace and the exponent  $T$  for the matrix transpose, which corresponds to the standard metric on  $3 \times 3$  matrices divided by 2. With this convention, the noise  $2\sqrt{\tau}$  will exactly yields a diffusion coefficient equal to  $\tau$  in the mean-field limit.

The mean-field model now provides the evolution of the distribution function  $f = f(x, A, t)$  which depends on the position  $x \in \mathbb{R}^d$ , rotation matrix  $A \in \text{SO}(3)$  and time  $t > 0$ . It is written

$$(2-34) \quad \partial_t f + \nabla_x \cdot (A e_1 f) = \nabla_A \cdot (-(P_{T_A} F_f) f + \tau \nabla_A f),$$

$$(2-35) \quad F_f(x, t) = \nu \Lambda_f(x, t), \quad \Lambda_f(x, t) = \text{PD}(G_f(x, t)),$$

$$(2-36) \quad G_f(x, t) = \int_{|y-x| \leq R} \int_{\text{SO}(3)} f(y, B, t) B dB dy,$$

Here, as pointed out before,  $\nabla_A$  and  $\nabla_A \cdot$  stand for the gradient and divergence operators on  $SO(3)$  when endowed with the Riemannian structure induced by the euclidean norm  $\|A\| = \frac{1}{2} \text{Tr}(A^T A)$ . The measure on  $SO(3)$  is the Haar measure normalized to be a probability measure. The passage from (2-31)-(2-33) to (2-34)-(2-36) is open but in a variant where  $G_i$  is used in the expression of  $F_i$  instead of  $\Lambda_i$ , the proof of Bolley, Cañizo, and Carrillo [2012] is likely to extend rather straightforwardly. In the case presented here, the control of  $G_i(t)$  away from the set of singular matrices presents additional challenges. To the best of our knowledge, the mathematical theory of this model is nonexistent.

A similar rescaling as in the previous section leads to the following perturbation problem (dropping terms of order  $\varepsilon^2$ ):

$$(2-37) \quad \varepsilon(\partial_t f^\varepsilon + \nabla_x \cdot (A e_1 f^\varepsilon)) = \nabla_A \cdot (-(P_{T_A} F_{f^\varepsilon}) f^\varepsilon + \tau \nabla_A f^\varepsilon),$$

$$(2-38) \quad F_f(x, t) = \nu \lambda_f(x, t), \quad \lambda_f(x, t) = \text{PD}(g_f(x, t)),$$

$$(2-39) \quad g_f(x, t) = \int_{SO(3)} f(x, B, t) B \, dB,$$

where we have denoted by  $g_f$  the local modification of  $G_f$  (involving only values of  $f$  at location  $x$ ) and  $\lambda_f$  its associated polar decomposition. This model can be written:

$$(2-40) \quad \partial_t f^\varepsilon + \nabla_x \cdot (A e_1 f^\varepsilon) = \frac{1}{\varepsilon} Q(f^\varepsilon)$$

$$(2-41) \quad Q(f) = \nabla_A \cdot (-(P_{T_A} F_f) f + \tau \nabla_A f)$$

with  $F_f$  given by (2-38), (2-39). The von Mises distribution is now defined by

$$(2-42) \quad M_{\kappa\Lambda}(A) = \frac{1}{Z} \exp(\kappa \Lambda \cdot A),$$

where  $\Lambda \cdot A$  is the matrix inner product of  $\Lambda$  and  $A$  defined above,  $\kappa = \nu/\tau$  and  $Z$  is a normalization constant only depending on  $\kappa$ . Then,  $Q(f)$  can be written as

$$(2-43) \quad Q(f) = \mathbb{Q}(f; \lambda_f),$$

where  $\lambda_f$  is given by (2-38) and

$$(2-44) \quad \mathbb{Q}(f; \lambda)(A) = \tau \nabla_A \cdot \left( M_{\kappa\lambda}(A) \nabla_A \left( \frac{f(A)}{M_{\kappa\lambda}(A)} \right) \right).$$

In the same way as before, as  $\varepsilon \rightarrow 0$ ,  $f^\varepsilon \rightarrow f^0$ , where  $f^0$  is a solution of  $Q(f^0) = 0$ . This implies the existence of  $\rho = \rho(x, t) \in [0, \infty)$  and  $\lambda = \lambda(x, t) \in SO(3)$  such that

$$(2-45) \quad f^0(x, A, t) = \rho(x, t) M_{\kappa\lambda(x,t)}(A).$$

Now, we define the GCI as follows:

**Definition 2.4.** For a body orientation given by the rotation matrix  $\lambda \in SO(3)$ , we define a GCI associated with  $\lambda$  as a function  $\psi(A)$  such that

$$(2-46) \quad \int \mathbb{Q}(f; \lambda)(A) \psi(A) dA = 0, \quad \forall f \text{ such that } P_{T_A} g_f = 0.$$

Up to now, the above body attitude alignment model could have been written in any dimension, i.e. for  $A \in SO(d)$  for any dimension  $d$ . The following characterization of the set of GCI now requires the dimension  $d$  to be equal to 3. A characterization like this in the case of a general dimension  $d$  is still an open problem.

**Theorem 2.5.** The set  $\mathbb{C}_\lambda$  of GCI associated to the body orientation given by the rotation matrix  $\lambda \in SO(3)$  is a linear vector space of dimension 4 expressed as follows:

$$(2-47) \quad \mathbb{C}_\lambda = \{C + P \cdot (\lambda^T A) h(\lambda \cdot A) \mid C \in \mathbb{R}, P \in \mathbb{Q}\},$$

where  $\mathbb{Q}$  denotes the space of antisymmetric  $3 \times 3$  matrices and where  $h: (0, \pi) \rightarrow \mathbb{R}$  is the unique solution of

$$(2-48) \quad \begin{aligned} -\frac{d}{d\theta} \left( \sin^2(\theta/2) m(\theta) \frac{d}{d\theta} (\sin \theta h(\theta)) \right) + \frac{1}{2} \sin \theta m(\theta) h(\theta) \\ = -\sin^2(\theta/2) \sin \theta m(\theta), \end{aligned}$$

in the space

$$(2-49) \quad H = \{h : (0, \pi) \rightarrow \mathbb{R} \mid \sin \theta h \in L^2(0, \pi), \sin(\theta/2) \frac{d}{d\theta} (\sin \theta h) \in L^2(0, \pi)\}.$$

Here, we have denoted by

$$m(\theta) = \frac{1}{Z} \exp \left( \kappa \left( \frac{1}{2} + \cos \theta \right) \right),$$

where  $Z$  is the normalization constant involved in (2-42)

Using this expression of the GCI and the same methodology as in the previous section, in [Degond, Frouvelle, and Merino-Aceituno \[2017\]](#), we have proved the following:

**Theorem 2.6.** Suppose that the solution  $f^\varepsilon$  of (2-37), (2-38) has a limit  $f^0$  when  $\varepsilon \rightarrow 0$ . Then,  $f^0$  is given by (2-45) where  $\kappa = v/\tau$  and the pair  $(\rho, \lambda): (x, t) \in \mathbb{R}^3 \times [0, \infty) \mapsto$

$(\rho, \lambda)(x, t) \in [0, \infty) \times SO(3)$  satisfies the following “self-organized hydrodynamics for body attitude coordination” (SOHB) model:

$$(2-50) \partial_t \rho + c_1 \nabla_x \cdot (\rho \lambda e_1) = 0,$$

$$(2-51) \quad \rho(\partial_t \lambda + c_2(\lambda e_1 \cdot \nabla_x) \lambda) + \left[ (\lambda e_1) \times (c_3 \nabla_x \rho + c_4 \rho r_x(\lambda)) + c_4 \rho \delta_x(\lambda) \lambda e_1 \right]_{\times} \lambda = 0,$$

with the coefficients  $c_1$  to  $c_4$  depending on  $v$  and  $\tau$ . The quantities  $r_x(\lambda)$  and  $\delta_x(\lambda)$  are given by:

$$(2-52) \quad \delta_x(\lambda) = \text{Tr}\{\mathfrak{D}_x(\lambda)\}, \quad r_x(\lambda) = \mathfrak{D}_x(\lambda) - \mathfrak{D}_x(\lambda)^T,$$

where  $\mathfrak{D}_x(\lambda)$  is the matrix defined, for any vector  $w \in \mathbb{R}^3$ , as follows:

$$(2-53) \quad (w \cdot \nabla_x) \lambda = [\mathfrak{D}_x(\lambda) w]_{\times} \lambda.$$

Here and above, for a vector  $w \in \mathbb{R}^3$ , we denote by  $[w]_{\times}$  the antisymmetric matrix defined for any vector  $z \in \mathbb{R}^3$  by

$$(2-54) \quad [w]_{\times} z = w \times z,$$

where  $\times$  denote the cross product of two vectors.

We note that (2-53) makes sense as  $(w \cdot \nabla_x) \lambda$  belongs to the tangent space  $T_{\lambda}$  of  $SO(3)$  at  $\lambda$  and  $T_{\lambda} = \{P \lambda \mid P \in \mathfrak{Q}\}$ . So, there exists  $u \in \mathbb{R}^3$  such that  $(w \cdot \nabla_x) \lambda = [u]_{\times} \lambda$  and since  $u$  depends linearly on  $w$ , there exists a matrix  $\mathfrak{D}_x(\lambda)$  such that  $u = \mathfrak{D}_x(\lambda) w$ . The notation  $\mathfrak{D}_x(\lambda)$  recalls that the coefficients of this matrix are linear combinations of first order derivatives of  $\lambda$ . Using the exponential map, in the neighborhood of any point  $x_0$ , we can write (omitting the time-dependence)  $\lambda(x) = \exp([b(x)]_{\times}) \lambda(x_0)$  where  $b$  is a smooth function from a neighborhood of  $x_0$  into  $\mathbb{R}^3$ . It is shown in [Degond, Frouvelle, and Merino-Aceituno \[ibid.\]](#) that

$$\delta_x(\lambda)(x_0) = (\nabla_x \cdot b)(x_0), \quad r_x(\lambda)(x_0) = (\nabla_x \times b)(x_0),$$

and thus,  $\delta_x(\lambda)$  and  $r_x(\lambda)$  can be interpreted as local “divergence” and “curl” of the matrix field  $\lambda$ . We note that (2-51) equally makes sense. Indeed, the expression on the first line is a derivative of the rotation field  $\lambda$  and should consequently belong to  $T_{\lambda(x,t)}$ . But the second line has precisely the required structure as it is the product of an antisymmetric matrix with  $\lambda$ . Equation (2-50) is the continuity equation for the density of agents moving at bulk velocity  $c_1 \lambda e_1$  so that  $\lambda e_1$  describes the fluid direction of motion. Equation (2-51)

gives the evolution of  $\lambda$ . The first line describes transport at velocity  $c_2 \lambda e_1$  and since  $c_2 \neq c_1$ , the transport of  $\lambda$  occurs at a different speed from the transport of mass, as in the SOH model (2-12), (2-13). The second line describes how  $\lambda$  evolves during its transport. The first term (proportional to  $\nabla_x \rho$ ) is the action of the pressure gradient and has the effect of turning the direction of motion away from high density regions. The other two terms are specific to the body attitude alignment model and do not have their counterpart in the classical SOH model (2-12), (2-13). The expressions of the coefficients  $c_2$  to  $c_4$  involve moments of the function  $h$  intervening in the expression of the GCI. The mathematical theory of the SOHB model is entirely open. We note that the above theory can be recast in the unitary quaternion framework, as done in [Degond, Frouvelle, Merino-Aceituno, and Trescases \[2018\]](#).

### 3 Phase transitions

**3.1 A Vicsek model exhibiting multiple equilibria.** Now, we go back to the Vicsek model of [Section 2.1](#). More precisely, we consider the kinetic model (2-7)-(2-9) in the spatially homogeneous case (i.e. we drop all dependences and derivatives with respect to position  $x$ ) and with  $\varepsilon = 1$ . However, we are interested in the case where the coefficients  $\tau$  and  $v$  are functions of  $|j_f|$ . More precisely, we consider the system

$$(3-55) \quad \partial_t f(v, t) = Q(f)(v, t),$$

$$(3-56)$$

$$Q(f)(v, t) = \nabla_v \cdot \left( -v(|j_f(t)|) (P_{v^\perp} u_f(t)) f(v, t) + \tau(|j_f(t)|) \nabla_v f(v, t) \right),$$

$$(3-57) \quad u_f(t) = \frac{j_f(t)}{|j_f(t)|}, \quad j_f(t) = \int_{\mathbb{S}^{d-1}} f(w, t) w \, dw.$$

For future usage, we introduce the function  $k(|j|) = \frac{v(|j|)}{\tau(|j|)}$ , as well as  $\Phi$  the primitive of  $k$ :  $\Phi(r) = \int_0^r k(s) \, ds$ . Introducing the free energy

$$(3-58) \quad \mathfrak{F}(f) = \int_{\mathbb{S}^{d-1}} f(v) \log f(v) \, dv - \Phi(|j_f|),$$

we find the free energy dissipation inequality

$$(3-59)$$

$$\frac{d}{dt} \mathfrak{F}(f)(t) = -\mathfrak{D}(f)(t),$$

$$(3-60) \quad \mathfrak{D}(f)(t) = \tau(|j_f(t)|) \int_{\mathbb{S}^{d-1}} f(v, t) \left| \nabla_v (f(v, t) - k(|j_f(t)|) (v \cdot u_f(t))) \right|^2.$$

In Degond, Frouvelle, and Liu [2015] (see a special case in Degond, Frouvelle, and Liu [2013]), we first give the proof of the following

**Theorem 3.1.** *Given an initial finite nonnegative measure  $f_0$  in the Sobolev space  $H^s(\mathbb{S}^{d-1})$ , there exists a unique weak solution  $f$  of (3-55) such that  $f(0) = f_0$ . This solution is global in time. Moreover,  $f \in C^1(\mathbb{R}_+^*, C^\infty(\mathbb{S}^{d-1}))$ , with  $f(v, t) > 0$  for all positive  $t$ . Furthermore, we have the following instantaneous regularity and uniform boundedness estimates (for  $m \in \mathbb{N}$ , the constant  $C$  being independent of  $f_0$ ):*

$$\|f(t)\|_{H^{s+m}}^2 \leq C \left(1 + \frac{1}{t^m}\right) \|f_0\|_{H^s}^2.$$

For these solutions, the density  $\rho(t) = \int_{\mathbb{S}^{d-1}} f(v, t) dv$  is constant in time, i.e.  $\rho(t) = \rho$ , where  $\rho = \int_{\mathbb{S}^{d-1}} f_0(v) dv$ .

The equilibria, i.e. the solutions of  $Q(f) = 0$  are given by  $\rho M_{\kappa u}$  where  $\rho$  is the initial density as defined in Theorem 3.1 and  $M_{\kappa u}$  is still the von Mises Fisher distribution (2-10) with arbitrary value of  $u \in \mathbb{S}^{d-1}$ . However, now the value of  $\kappa$  is found by the resolution of a fixed-point equation (the consistency condition)

$$(3-61) \quad \kappa = k(|j_{\rho M_{\kappa u}}|).$$

This equation can be recast by noting that  $|j_{\rho M_{\kappa u}}| = \rho c_1(\kappa)$  where  $c_1(\kappa)$  is the order parameter (2-23). Assuming that the function  $k: |j| \in [0, \infty) \mapsto k(|j|) \in [0, \infty)$  is strictly increasing and surjective, we can define its inverse  $\iota: \kappa \in [0, \infty) \mapsto \iota(\kappa) \in [0, \infty)$ . This assumption may be seen as restrictive, but it is easy to remove it at the expense of more technicalities, which we want to avoid in this presentation. As by definition  $\iota(k(|j|)) = |j|$ , applying the function  $\iota$  to (3-61), we can recast it in

$$(3-62) \quad \text{either } \kappa = 0 \quad \text{or} \quad \frac{\iota(\kappa)}{c_1(\kappa)} = \rho.$$

Note that for  $\kappa = 0$ , the von Mises distribution is the uniform distribution on the sphere. We will call the corresponding equilibrium, “isotropic equilibrium”. Any von Mises distribution with  $\kappa > 0$  will be called a “non-isotropic equilibrium”. For a given  $\kappa > 0$ , the von Mises equilibria  $\rho M_{\kappa u}$  form a manifold diffeomorphically parametrized by  $u \in \mathbb{S}^{d-1}$ . Both  $\iota$  and  $c_1$  are increasing functions of  $\kappa$  so the ratio  $\frac{\iota(\kappa)}{c_1(\kappa)}$  has no defined monotonicity a priori. For a given  $\rho$  the number of solutions  $\kappa$  of (3-62) depends on the particular choice of the function  $k$ . However, we can state the following proposition:

**Proposition 3.2.** *Let  $\rho > 0$ . We define*

$$(3-63) \quad \rho_c = \lim_{\kappa \rightarrow 0} \frac{\iota(\kappa)}{c_1(\kappa)}, \quad \rho_* = \inf_{\kappa \in (0, \infty)} \frac{\iota(\kappa)}{c_1(\kappa)},$$



where  $\rho_c > 0$  may be equal to  $+\infty$ . Then we have  $\rho_c \geq \rho_*$ , and

- (i) If  $\rho < \rho_*$ , the only solution to (3-62) is  $\kappa = 0$  and the only equilibrium with total mass  $\rho$  is the uniform distribution  $f = \rho$ .
- (ii) If  $\rho > \rho_*$ , there exists at least one positive solution  $\kappa > 0$  to (3-62). It corresponds to a family  $\{\rho M_{\kappa u}, u \in \mathbb{S}^{d-1}\}$  of non-isotropic von Mises equilibria.
- (iii) The number of families of nonisotropic equilibria changes as  $\rho$  crosses the threshold  $\rho_c$ . Under regularity and non-degeneracy hypotheses, in a neighborhood of  $\rho_c$ , this number is even when  $\rho < \rho_c$  and odd when  $\rho > \rho_c$ .

Now, the key question is the stability of these equilibria. A first general result can be established thanks to the La Salle principle:

**Proposition 3.3.** *Let  $f_0$  be a positive measure on the sphere  $\mathbb{S}^{d-1}$ , with mass  $\rho$ , and  $f(t)$  the associated solution to (3-55). If no open interval is included in the set  $\{\kappa \in [0, \infty) \mid \rho c(\kappa) = \iota(\kappa)\}$ , then there exists a solution  $\kappa_\infty$  to (3-62) such that:*

$$(3-64) \quad \lim_{t \rightarrow \infty} |j_f(t)| = \rho c(\kappa_\infty)$$

and

$$(3-65) \quad \forall s \in \mathbb{R}, \lim_{t \rightarrow \infty} \|f(t) - \rho M_{\kappa_\infty u_{f(t)}}\|_{H^s} = 0.$$

In other words, under these conditions, the family of equilibria  $\{\rho M_{\kappa_\infty u} \mid u \in \mathbb{S}^{d-1}\}$  is an  $\omega$ -limit set of the trajectories of (3-55). Now, we study separately the stability of the isotropic and non-isotropic equilibria.

**3.2 Stability of the isotropic equilibria.** For the isotropic equilibria, we have the following two propositions:

**Proposition 3.4.** *Let  $f(t)$  be the solution to (3-55) associated with initial condition  $f_0$  of mass  $\rho$ . If  $\rho > \rho_c$ , and if  $j_{f_0} \neq 0$ , then we cannot have  $\kappa_\infty = 0$  in Proposition 3.3.*

**Proposition 3.5.** *Suppose that  $\rho < \rho_c$ . We define*

$$(3-66) \quad \lambda = (n-1)\tau_0\left(1 - \frac{\rho}{\rho_c}\right) > 0.$$

*Let  $f_0$  be an initial condition with mass  $\rho$ , and  $f$  the corresponding solution to (3-55). There exists  $\delta > 0$  independent of  $f_0$  such that if  $\|f_0 - \rho\|_{H^s} < \delta$ , then for all  $t \geq 0$*

$$\|f(t) - \rho\|_{H^s} \leq \frac{\|f_0 - \rho\|_{H^s}}{1 - \frac{1}{\delta}\|f_0 - \rho\|_{H^s}} e^{-\lambda t}.$$

Proposition 3.4 implies the instability of the uniform equilibria for  $\rho > \rho_c$  (provided the initial current  $j_{f_0}$  does not vanish) as the  $\omega$ -limit set of the trajectories consists of non-isotropic equilibria. Proposition 3.5 shows the stability of the uniform equilibria for  $\rho < \rho_c$  in any  $H^s$  norm with exponential decay rate given by (3-66). We stress that these are fully nonlinear stability/instability results.

**3.3 Stability of the non-isotropic equilibria.** Let  $\kappa > 0$  and  $\rho > 0$  be such that  $\kappa$  is a solution to (3-62). In addition to the hypotheses made so far on  $k$ , we assume that  $k$  is differentiable, with its derivative  $k'$  being itself Lipschitz. The following result shows that the stability or instability of the non-isotropic equilibria is determined by whether the function  $\kappa \mapsto \frac{\iota(\kappa)}{c_1(\kappa)}$  is strictly increasing or decreasing.

**Proposition 3.6.** *Let  $\kappa > 0$  and  $\rho = \frac{\iota(\kappa)}{c_1(\kappa)}$ . We denote by  $\mathfrak{F}_\kappa$  the value of  $\mathfrak{F}(\rho M_{\kappa u})$  (independent of  $u \in \mathbb{S}^{d-1}$ ).*

- (i) *Suppose  $(\frac{\iota}{c_1})'(\kappa) < 0$ . Then any equilibrium of the form  $\rho M_{\kappa u}$  is unstable, in the following sense: in any neighborhood of  $\rho M_{\kappa u}$ , there exists an initial condition  $f_0$  such that  $\mathfrak{F}(f_0) < \mathfrak{F}_\kappa$ . Consequently, in that case, we cannot have  $\kappa_\infty = \kappa$  in Proposition 3.3.*
- (ii) *Suppose  $(\frac{\iota}{c_1})'(\kappa) > 0$ . Then the family of equilibria  $\{\rho M_{\kappa u}, u \in \mathbb{S}^{d-1}\}$  is stable, in the following sense: for all  $K > 0$  and  $s > \frac{n-1}{2}$ , there exists  $\delta > 0$  and  $C$  such that for all  $f_0$  with mass  $\rho$  and with  $\|f_0\|_{H^s} \leq K$ , if  $\|f_0 - \rho M_{\kappa u}\|_{L^2} \leq \delta$  for some  $u \in \mathbb{S}^{d-1}$ , then for all  $t \geq 0$ , we have*

$$\begin{aligned} \mathfrak{F}(f) &\geq \mathfrak{F}_\kappa, \\ \|f - \rho M_{\kappa u_f}\|_{L^2} &\leq C \|f_0 - \rho M_{\kappa u_{f_0}}\|_{L^2}. \end{aligned}$$

Note that the marginal case  $(\frac{\iota}{c_1})'(\kappa) = 0$  is not covered by the above theorem and is still an open problem. In the stable case, the following proposition provides the rate of decay to an element of the same family of equilibria:

**Theorem 3.7.** *Suppose  $(\frac{\iota}{c_1})'(\kappa) > 0$ . Then, for all  $s > \frac{n-1}{2}$ , there exist constants  $\delta > 0$  and  $C > 0$  such that for any  $f_0$  with mass  $\rho$  satisfying  $\|f_0 - \rho M_{\kappa u}\|_{H^s} < \delta$  for some  $u \in \mathbb{S}^{d-1}$ , there exists  $u_\infty \in \mathbb{S}^{d-1}$  such that*

$$\|f - \rho M_{\kappa u_\infty}\|_{H^s} \leq C \|f_0 - \rho M_{\kappa u}\|_{H^s} e^{-\lambda t},$$

where the rate  $\lambda$  is given by

$$(3-67) \quad \lambda = \frac{c_1(\kappa) \tau(\iota(\kappa))}{\iota'(\kappa)} \Lambda_\kappa \left( \frac{\iota}{c_1} \right)'(\kappa).$$

The constant  $\Lambda_\kappa$  is the best constant for the following weighted Poincaré inequality (see the appendix of [Degond, Frouvelle, and Liu \[2013\]](#)):

$$(3-68) \quad \langle |\nabla_\omega g|^2 \rangle_M \geq \Lambda_\kappa \langle (g - \langle g \rangle_M)^2 \rangle_M,$$

where we have written  $\langle g \rangle_M$  for  $\int_{\mathbb{S}} g(v) M_{\kappa u}(v) dv$ .

## 4 Conclusion

In this short overview, we have surveyed some of the mathematical questions posed by collective dynamics and self-organization. We have particularly focused on two specific problems: the derivation of macroscopic models and the study of phase transitions. There are of course many other fascinating challenges posed by self-organized systems. These have shown to be an inexhaustible source of problems for mathematicians and a drive for the invention of new mathematical concepts.

## References

- I. Aoki (1982). “A simulation study on the schooling mechanism in fish.” *Bulletin of the Japan Society of Scientific Fisheries* 48, pp. 1081–1088 (cit. on p. [3943](#)).
- Per Bak, Chao Tang, and Kurt Wiesenfeld (1987). “Self-organized criticality: an explanation of  $1/f$  noise”. *Phys. Rev. A* (3) 59, pp. 381–384 (cit. on p. [3944](#)).
- Alethea B. T. Barbaro and Pierre Degond (2014). “[Phase transition and diffusion among socially interacting self-propelled agents](#)”. *Discrete Contin. Dyn. Syst. Ser. B* 19.5, pp. 1249–1278. MR: [3199779](#) (cit. on p. [3947](#)).
- Sepideh Bazazi, Jerome Buhl, Joseph J Hale, Michael L Anstey, Gregory A Sword, Stephen J Simpson, and Iain D Couzin (2008). “Collective motion and cannibalism in locust migratory bands”. *Current Biology* 18, pp. 735–739 (cit. on p. [3943](#)).
- F. Berthelin, P. Degond, M. Delitala, and M. Rascle (2008). “[A model for the formation and evolution of traffic jams](#)”. *Arch. Ration. Mech. Anal.* 187.2, pp. 185–220. MR: [2366138](#) (cit. on p. [3947](#)).
- Eric Bertin, Michel Droz, and Guillaume Grégoire (2009). “Hydrodynamic equations for self-propelled particles: microscopic derivation and stability analysis”. *Journal of Physics A: Mathematical and Theoretical* 42, p. 445001 (cit. on p. [3948](#)).
- Emmanuel Boissard, Pierre Degond, and Sebastien Motsch (2013). “[Trail formation based on directed pheromone deposition](#)”. *J. Math. Biol.* 66.6, pp. 1267–1301. MR: [3040976](#) (cit. on p. [3947](#)).

- François Bolley, José A. Cañizo, and José A. Carrillo (2012). “Mean-field limit for the stochastic Vicsek model”. *Appl. Math. Lett.* 25.3, pp. 339–343. MR: [2855983](#) (cit. on pp. [3949](#), [3955](#)).
- Eric Carlen, Pierre Degond, and Bernt Wennberg (2013). “Kinetic limits for pair-interaction driven master equations and biological swarm models”. *Math. Models Methods Appl. Sci.* 23.7, pp. 1339–1376. MR: [3042918](#) (cit. on p. [3946](#)).
- J. A. Carrillo, M. Fornasier, J. Rosado, and G. Toscani (2010). “Asymptotic flocking dynamics for the kinetic Cucker-Smale model”. *SIAM J. Math. Anal.* 42.1, pp. 218–236. MR: [2596552](#) (cit. on p. [3947](#)).
- Yao-li Chuang, Maria R. D’Orsogna, Daniel Marthaler, Andrea L. Bertozzi, and Lincoln S. Chayes (2007). “State transitions and the continuum limit for a 2D interacting, self-propelled particle system”. *Phys. D* 232.1, pp. 33–47. MR: [2369988](#) (cit. on p. [3947](#)).
- Iain D. Couzin, Jens Krause, Richard James, Graeme D. Ruxton, and Nigel R. Franks (2002). “Collective memory and spatial sorting in animal groups”. *J. Theoret. Biol.* 218.1, pp. 1–11. MR: [2027139](#) (cit. on p. [3947](#)).
- Adama Creppy, Franck Plouraboué, Olivier Praud, Xavier Druart, Sébastien Cazin, Hui Yu, and Pierre Degond (2016). “Symmetry-breaking phase transitions in highly concentrated semen”. *Journal of The Royal Society Interface* 13.123, p. 20160575 (cit. on p. [3947](#)).
- Felipe Cucker and Steve Smale (2007). “Emergent behavior in flocks”. *IEEE Trans. Automat. Control* 52.5, pp. 852–862. MR: [2324245](#) (cit. on p. [3947](#)).
- András Czirók, Eshel Ben-Jacob, Inon Cohen, and Tamás Vicsek (1996). “Formation of complex bacterial colonies via self-generated vortices”. *Physical Review E* 54.2, pp. 1791–18091 (cit. on p. [3944](#)).
- Pierre Degond, Amic Frouvelle, and Jian-Guo Liu (2013). “Macroscopic limits and phase transition in a system of self-propelled particles”. *J. Nonlinear Sci.* 23.3, pp. 427–456. MR: [3067586](#) (cit. on pp. [3947](#), [3959](#), [3962](#)).
- (2015). “Phase transitions, hysteresis, and hyperbolicity for self-organized alignment dynamics”. *Arch. Ration. Mech. Anal.* 216.1, pp. 63–115. MR: [3305654](#) (cit. on pp. [3947](#), [3959](#)).
- Pierre Degond, Amic Frouvelle, and Sara Merino-Aceituno (2017). “A new flocking model through body attitude coordination”. *Math. Models Methods Appl. Sci.* 27.6, pp. 1005–1049. MR: [3659045](#) (cit. on pp. [3953](#), [3956](#), [3957](#)).
- Pierre Degond, Amic Frouvelle, Sara Merino-Aceituno, and Ariane Trescases (2018). “Quaternions in collective dynamics”. *Multiscale Model. Simul.* 16.1, pp. 28–77. MR: [3743738](#) (cit. on pp. [3953](#), [3958](#)).
- Pierre Degond and Jiale Hua (2013). “Self-organized hydrodynamics with congestion and path formation in crowds”. *J. Comput. Phys.* 237, pp. 299–319. MR: [3020033](#) (cit. on p. [3947](#)).

- Pierre Degond, Jiale Hua, and Laurent Navoret (2011). “[Numerical simulations of the Euler system with congestion constraint](#)”. *J. Comput. Phys.* 230.22, pp. 8057–8088. MR: [2835410](#) (cit. on p. [3947](#)).
- Pierre Degond, Jian-Guo Liu, Sebastien Motsch, and Vladislav Panferov (2013). “[Hydrodynamic models of self-organized dynamics: derivation and existence theory](#)”. *Methods Appl. Anal.* 20.2, pp. 89–114. MR: [3119732](#) (cit. on p. [3951](#)).
- Pierre Degond, Angelika Manhart, and Hui Yu (2017). “[A continuum model for nematic alignment of self-propelled particles](#)”. *Discrete Contin. Dyn. Syst. Ser. B* 22.4, pp. 1295–1327. MR: [3639166](#) (cit. on pp. [3948](#), [3953](#)).
- Pierre Degond and Sébastien Motsch (2008a). “[Continuum limit of self-driven particles with orientation interaction](#)”. *Math. Models Methods Appl. Sci.* 18.suppl. Pp. 1193–1215. MR: [2438213](#) (cit. on pp. [3946](#), [3950](#), [3952](#)).
- (2008b). “[Large scale dynamics of the persistent turning walker model of fish behavior](#)”. *J. Stat. Phys.* 131.6, pp. 989–1021. MR: [2407377](#) (cit. on p. [3943](#)).
- (2011). “[A macroscopic model for a system of swarming agents using curvature control](#)”. *J. Stat. Phys.* 143.4, pp. 685–714. MR: [2800660](#) (cit. on pp. [3943](#), [3953](#)).
- Pierre Degond, Laurent Navoret, Richard Bon, and David Sanchez (2010). “Congestion in a macroscopic model of self-driven particles modeling gregariousness”. *Journal of Statistical Physics* 138.1-3, pp. 85–125 (cit. on p. [3947](#)).
- Michael L Domeier and Patrick L Colin (1997). “Tropical reef fish spawning aggregations: defined and reviewed”. *Bulletin of Marine Science* 60.3, pp. 698–726 (cit. on p. [3943](#)).
- Alessio Figalli, Moon-Jin Kang, and Javier Morales (2018). “[Global Well-posedness of the Spatially Homogeneous Kolmogorov–Vicsek Model as a Gradient Flow](#)”. *Arch. Ration. Mech. Anal.* 227.3, pp. 869–896. MR: [3744377](#) (cit. on p. [3949](#)).
- Amic Frouvelle (2012). “[A continuum model for alignment of self-propelled particles with anisotropy and density-dependent parameters](#)”. *Math. Models Methods Appl. Sci.* 22.7, pp. 1250011, 40. MR: [2924786](#) (cit. on p. [3952](#)).
- Amic Frouvelle and Jian-Guo Liu (2012). “[Dynamics in a kinetic model of oriented particles with phase transition](#)”. *SIAM J. Math. Anal.* 44.2, pp. 791–826. MR: [2914250](#) (cit. on p. [3947](#)).
- Isabelle Gallagher, Laure Saint-Raymond, and Benjamin Texier (2013). *From Newton to Boltzmann: hard spheres and short-range potentials*. Zurich Lectures in Advanced Mathematics. European Mathematical Society (EMS), Zürich, pp. xii+137. MR: [3157048](#) (cit. on p. [3945](#)).
- Irene M. Gamba and Moon-Jin Kang (2016). “[Global weak solutions for Kolmogorov–Vicsek type equations with orientational interactions](#)”. *Arch. Ration. Mech. Anal.* 222.1, pp. 317–342. MR: [3519972](#) (cit. on p. [3949](#)).

- Jacques Gautrais, Francesco Ginelli, Richard Fournier, Stéphane Blanco, Marc Soria, Hugues Chaté, and Guy Theraulaz (2012). “[Deciphering interactions in moving animal groups](#)”. *PLoS Comput. Biol.* 8.9, e1002678, 11. MR: [2993806](#) (cit. on p. [3943](#)).
- Jacques Gautrais, Christian Jost, Marc Soria, Alexandre Campo, Sébastien Motsch, Richard Fournier, Stéphane Blanco, and Guy Theraulaz (2009). “[Analyzing fish movement as a persistent turning walker](#)”. *J. Math. Biol.* 58.3, pp. 429–445. MR: [2470196](#) (cit. on p. [3943](#)).
- Francesco Ginelli, Fernando Peruani, Markus Bär, and Hugues Chaté (2010). “Large-scale collective properties of self-propelled rods”. *Physical review letters* 104.18, p. 184502 (cit. on p. [3948](#)).
- Seung-Yeal Ha and Jian-Guo Liu (2009). “[A simple proof of the Cucker-Smale flocking dynamics and mean-field limit](#)”. *Commun. Math. Sci.* 7.2, pp. 297–325. MR: [2536440](#) (cit. on p. [3947](#)).
- Seung-Yeal Ha and Eitan Tadmor (2008). “[From particle to kinetic and hydrodynamic descriptions of flocking](#)”. *Kinet. Relat. Models* 1.3, pp. 415–435. MR: [2425606](#) (cit. on p. [3947](#)).
- Jan Haskovec, Peter Markowich, Benoît Perthame, and Matthias Schlottbom (2016). “[Notes on a PDE system for biological network formation](#)”. *Nonlinear Anal.* 138, pp. 127–155. MR: [3485142](#) (cit. on p. [3947](#)).
- Dirk Helbing, Illés J Farkas, and Tamás Vicsek (2000). “Freezing by heating in a driven mesoscopic system”. *Physical review letters* 84.6, p. 1240 (cit. on p. [3944](#)).
- Elton P. Hsu (2002). *Stochastic analysis on manifolds*. Vol. 38. Graduate Studies in Mathematics. American Mathematical Society, Providence, RI, pp. xiv+281. MR: [1882015](#) (cit. on pp. [3948](#), [3954](#)).
- Ning Jiang, Linjie Xiong, and Teng-Fei Zhang (2016). “[Hydrodynamic limits of the kinetic self-organized models](#)”. *SIAM J. Math. Anal.* 48.5, pp. 3383–3411. MR: [3549879](#) (cit. on p. [3951](#)).
- Anaïs Khuong, Guy Theraulaz, Christian Jost, Andrea Perna, and Jacques Gautrais (2011). “A computational model of ant nest morphogenesis.” In: *ECAL*. MIT press, pp. 404–411 (cit. on p. [3943](#)).
- Oscar E. Lanford III (1976). “On a derivation of the Boltzmann equation”, 117–137. *Astérisque*, No. 40. MR: [0459449](#) (cit. on p. [3946](#)).
- Mathieu Leroy-Lerêtre, Giacomo Dimarco, Martine Cazales, Marie-Laure Boizeau, Bernard Ducommun, Valérie Lobjois, and Pierre Degond (2017). “[Are tumor cell lineages solely shaped by mechanical forces?](#)” *Bull. Math. Biol.* 79.10, pp. 2356–2393. MR: [3697232](#) (cit. on p. [3947](#)).
- Ryan Lukeman, Yue-Xian Li, and Leah Edelstein-Keshet (2010). “Inferring individual rules from collective behavior”. *Proceedings of the National Academy of Sciences* 107.28, pp. 12576–12580 (cit. on p. [3943](#)).

- Stéphane Mischler and Clément Mouhot (2013). “Kac’s program in kinetic theory”. *Invent. Math.* 193.1, pp. 1–147. MR: [3069113](#) (cit. on p. [3945](#)).
- Sebastien Motsch and Eitan Tadmor (2011). “A new model for self-organized dynamics and its flocking behavior”. *J. Stat. Phys.* 144.5, pp. 923–947. MR: [2836613](#) (cit. on p. [3947](#)).
- Mehdi Moussaid et al. (2012). “Traffic instabilities in self-organized pedestrian crowds”. *PLoS computational biology* 8.3, e1002442 (cit. on p. [3943](#)).
- Benoît Perthame, Fernando Quirós, and Juan Luis Vázquez (2014). “The Hele-Shaw asymptotics for mechanical models of tumor growth”. *Arch. Ration. Mech. Anal.* 212.1, pp. 93–127. MR: [3162474](#).
- Diane Peurichard, Fanny Delebecque, Anne Lorisgnol, Corinne Barreau, Jacques Rouquette, Xavier Descombes, Louis Casteilla, and Pierre Degond (2017). “Simple mechanical cues could explain adipose tissue morphology”. *Journal of theoretical biology* 429, pp. 61–81 (cit. on p. [3947](#)).
- Mathieu Poujade, Erwan Grasland-Mongrain, A Hertzog, J Jouanneau, Philippe Chavier, Benoît Ladoux, Axel Buguin, and Pascal Silberzan (2007). “Collective migration of an epithelial monolayer in response to a model wound”. *Proceedings of the National Academy of Sciences* 104.41, pp. 15988–15993 (cit. on p. [3944](#)).
- Jackie Shen (2007/08). “Cucker-Smale flocking under hierarchical leadership”. *SIAM J. Appl. Math.* 68.3, pp. 694–719. MR: [2375291](#) (cit. on p. [3947](#)).
- Boris I Shraiman (2005). “Mechanical feedback as a possible regulator of tissue growth”. *Proceedings of the National Academy of Sciences of the United States of America* 102.9, pp. 3318–3323 (cit. on p. [3944](#)).
- John Toner and Yuhai Tu (1998). “Flocks, herds, and schools: a quantitative theory of flocking”. *Phys. Rev. E* (3) 58.4, pp. 4828–4858. MR: [1651324](#) (cit. on p. [3945](#)).
- Tamás Vicsek, András Czirók, Eshel Ben-Jacob, Inon Cohen, and Ofer Shochet (1995). “Novel type of phase transition in a system of self-driven particles”. *Phys. Rev. Lett.* 75.6, pp. 1226–1229. MR: [3363421](#) (cit. on pp. [3944](#), [3945](#), [3948](#)).
- Tamás Vicsek and Anna Zafeiris (2012). “Collective motion”. *Physics Reports* 517.3-4, pp. 71–140 (cit. on pp. [3944](#), [3946](#), [3948](#)).

Received 2017-10-25.

PIERRE DEGOND

DEPARTMENT OF MATHEMATICS, IMPERIAL COLLEGE LONDON, LONDON, SW7 2AZ, UK

[pdegond@imperial.ac.uk](mailto:pdegond@imperial.ac.uk)

# ALGORITHMS FOR MOTION OF NETWORKS BY WEIGHTED MEAN CURVATURE

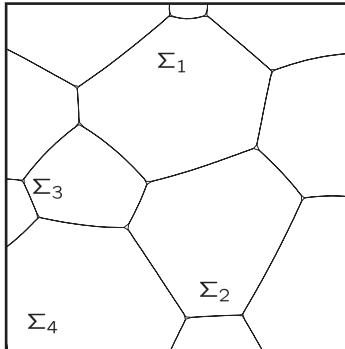
SELIM ESEDOĞLU

## Abstract

I will report on recent developments in a class of algorithms, known as threshold dynamics, for computing the motion of interfaces by mean curvature. These algorithms try to generate the desired interfacial motion just by alternating two very simple operations: Convolution, and thresholding. They can be extended to the multi-phase setting of networks of surfaces, and to motion by weighted (anisotropic) mean curvature, while maintaining the simplicity of the original version. These extensions are relevant in applications such as materials science, where they allow large scale simulation of models for microstructure evolution in polycrystals.

## 1 Introduction

We will discuss algorithms for simulating the motion of a network of intersecting interfaces in  $\mathbb{R}^d$ , with focus on  $d = 2$  or  $3$ . Mathematically, we describe such a network as the union of boundaries  $\cup_{i=1}^N \partial \Sigma_i$  of sets  $\Sigma = (\Sigma_1, \dots, \Sigma_N)$  (also called *phases*) that partition a domain  $D \subset \mathbb{R}^3$  (typically a periodic box) without overlaps or vacuum:



(1)

$$D = \bigcup_{j=1}^N \Sigma_j, \text{ and}$$

$$\Sigma_i \cap \Sigma_j = (\partial \Sigma_i) \cap (\partial \Sigma_j) \text{ for } i \neq j.$$



Many applications in science and engineering, ranging from models of microstructural evolution in polycrystalline materials to segmentation of images in computer vision, entail variational models with cost functions of the form

$$(2) \quad E(\Sigma) = \sum_{\substack{i,j=1 \\ i \neq j}}^N \int_{(\partial \Sigma_i) \cap (\partial \Sigma_j)} \sigma_{i,j}(n_{i,j}(x)) \, dS(x)$$

Here,  $dS$  is the length element in  $d = 2$  or surface area element in  $d = 3$ ,  $n_{i,j}(x)$  for  $x \in (\partial \Sigma_i) \cap (\partial \Sigma_j)$  denotes the unit normal from  $\Sigma_i$  to  $\Sigma_j$ , and the continuous, even functions  $\sigma_{i,j} : \mathbb{S}^{d-1} \rightarrow \mathbb{R}^+$  are the *surface tensions* associated with the interfaces.

In materials science, energy (2) and its  $L^2$  gradient descent dynamics we recall below, was proposed by Mullins [1956] as a continuum model for grain boundary motion in polycrystalline materials – a class that includes most metals and ceramics. In this context, the sets  $\Sigma_i$  in (1) represent the space occupied by single crystal pieces (grains) in the material that differ from one another only in their crystallographic orientations. When the material is heated, atoms may detach from one grain and attach to a neighbor, leading to the motion of the boundaries: Some grains grow at the expense of others, leading to many topological changes in the network as it coarsens; see Figures 1 and 2. There are models that describe how the surface tensions  $\sigma_{i,j}$  are to be determined from the orientations of any two grains  $\Sigma_i$  and  $\Sigma_j$ . The orientations may be chosen e.g. at random at the beginning of a simulation and are typically assumed to remain constant in time. The  $\sigma_{i,j}$  turn out to also depend on the normal to the interface between the two neighboring grains. For this application, it is therefore important to have numerical algorithms capable of treating the full generality of model (2).

When  $\sigma_{i,j}(x) = 1$  for all  $x$  and  $i \neq j$ , energy (2) becomes simply the sum of Euclidean surface areas of the interfaces in the network. In this form, it appears often as part of variational models in image segmentation, such as the Mumford-Shah model Mumford and Shah [1989] and its piecewise constant variants Chan and Vese [2001] and Vese and Chan [2002], where the sets  $\Sigma_i$  represent the space occupied by distinct objects in a scene. The goal of image segmentation is then to automatically discover these regions, which variational models such as Mumford and Shah [1989] exhibit as the minimizer of a cost function. Perimeter of the unknown sets is penalized in the cost function to control the level of detail in the segmentation obtained. While gradient flow for suitable approximations of (2) is certainly often employed to minimize the cost function, in this application the minimizer rather than the precise evolution required to reach it is of main interest, and there may be more effective ways than gradient flow to do so.

It is convenient to extend  $\sigma_{i,j}$  as one-homogeneous, continuous functions to all of  $\mathbb{R}^d$  as

$$(3) \quad \sigma(x) = |x| \sigma\left(\frac{x}{|x|}\right) \text{ for } x \neq 0$$

in which case well-posedness of model (2) requires them to be also convex. We will in fact assume that all  $\sigma_{i,j}$  have strongly convex and smooth unit balls in this discussion, in particular staying away from crystalline cases where the unit ball is a polytope.

For  $d = 2$  or  $3$ , we will study approximations for  $L^2$  gradient flow of energy (2), which is known as multiphase weighted mean curvature flow. In three dimensions, for an  $x \in (\partial\Sigma_i) \cap (\partial\Sigma_j)$  away from junctions (where three or more phases meet), normal speed under this flow is given by

$$(4) \quad v_\perp(x) = \mu_{i,j}(n_{i,j}(x)) \left( (\partial_{s_1}^2 \sigma_{i,j}(n_{i,j}(x)) + \sigma_{i,j}(n_{i,j}(x))) \kappa_1(x) \right. \\ \left. + (\partial_{s_2}^2 \sigma_{i,j}(n_{i,j}(x)) + \sigma_{i,j}(n_{i,j}(x))) \kappa_2(x) \right)$$

where  $\kappa_1$  and  $\kappa_2$  are the two principal curvatures, and  $\partial_{s_i}$  denotes differentiation along the great circle on  $\mathbb{S}^2$  that passes through  $n(x)$  and has as its tangent the  $i$ -th principal curvature direction. The additional factor  $\mu_{i,j}$  is known as the *mobility* associated with the interface  $(\partial\Sigma_i) \cap (\partial\Sigma_j)$ , and may be anisotropic:  $\mu_{i,j} : \mathbb{S}^2 \rightarrow \mathbb{R}^+$ . We will assume that it is smooth and has a one-homogeneous extension to  $\mathbb{R}^3$  that is a norm. In two dimensions, (4) simplifies to

$$(5) \quad v_\perp(x) = \mu_{i,j}(n_{i,j}(x)) \left( \sigma''_{i,j}(n_{i,j}(x)) + \sigma_{i,j}(n_{i,j}(x)) \right) \kappa(x).$$

In addition to (4), a condition known as the *Herring angle condition* [Herring \[1951\]](#) holds along triple junctions: In three dimensions, at a junction formed by the meeting of the three phases  $\Sigma_i$ ,  $\Sigma_j$ , and  $\Sigma_k$ , this condition reads

$$(6) \quad (\ell \times n_{i,j}) \sigma_{i,j}(n_{i,j}) + (\ell \times n_{j,k}) \sigma_{j,k}(n_{j,k}) + (\ell \times n_{k,i}) \sigma_{k,i}(n_{k,i}) \\ + n_{j,i} \sigma'_{i,j}(n_{i,j}) + n_{k,j} \sigma'_{j,k}(n_{j,k}) + n_{i,k} \sigma'_{k,i}(n_{k,i}) = 0$$

where  $\ell = n_{j,k} \times n_{i,j}$  is a unit vector tangent to the triple junction, and  $\sigma'_{i,j}(n_{i,j})$  denotes derivative of  $\sigma_{i,j}$  taken on  $\mathbb{S}^2$  in the direction of the vector  $\ell \times n_{i,j}$ . In the isotropic setting, (6) simplifies to the following more familiar form, known as Young's law:

$$(7) \quad \sigma_{i,j} n_{i,j} + \sigma_{j,k} n_{j,k} + \sigma_{k,i} n_{k,i} = 0.$$

which can be rearranged to determine the angles between interfaces at a triple junction in terms of their surface tensions. For example, in the simplest case  $\sigma_{i,j} = 1$  for all  $i \neq j$ , (7) implies all three angles at a triple junction are  $120^\circ$ .

Finally, we note that well-posedness of the multiphase energy (2) in its full generality is complicated [Ambrosio and Braides \[1990\]](#). At the very least, in addition to convexity, the  $\sigma_{i,j}$  need to satisfy a pointwise triangle inequality

$$(8) \quad \sigma_{i,j}(n) + \sigma_{j,k}(n) \geq \sigma_{i,k}(n)$$

for all distinct  $i, j$ , and  $k$ , and all  $n \in \mathbb{S}^{d-1}$ . In case the  $\sigma_{i,j}$  are positive constants, (8) is known to be also sufficient for well-posedness of model (2).

## 2 Isotropic and Equal Surface Tensions

In [Merriman, Bence, and Osher \[1992, 1994\]](#), the authors proposed a remarkably elegant algorithm for dynamics (Equations (4) and (7)) in the special case that all surface tensions and mobilities satisfy  $\sigma_{i,j}(x) = \mu_{i,j}(x) = 1$  for all  $x$  and  $i \neq j$ . Called *threshold dynamics* (also *diffusion generated motion*), it generates a discrete in time approximation to the flow from an initial partition  $\Sigma^0 = (\Sigma_1^0, \dots, \Sigma_N^0)$  as follows:

**Algorithm:** (from [Merriman, Bence, and Osher \[1994\]](#)): Given a time step size  $\delta t > 0$ , alternate the following steps:

1. Convolution:

$$(9) \quad \psi_i^k = K_{\sqrt{\delta t}} * \mathbf{1}_{\Sigma_i^k}.$$

2. Redistribution:

$$(10) \quad \Sigma_i^{k+1} = \left\{ x : \psi_i^k(x) \geq \max_{j \neq i} \psi_j^k(x) \right\}.$$

where  $K$  is a convolution kernel that has the properties

$$(11) \quad K(x) \in L^1(\mathbb{R}^d), \quad xK(x) \in L^1(\mathbb{R}^d), \text{ and } K(x) = K(-x)$$

and the notation  $K_\varepsilon(x) = \varepsilon^{-d} K(x/\varepsilon)$  denotes its rescaled version. In the original papers [Merriman, Bence, and Osher \[1992, 1994\]](#), the convolution kernel  $K$  is chosen to be the Gaussian

$$(12) \quad G(x) = \frac{1}{(4\pi)^{\frac{d}{2}}} \exp\left(-\frac{|x|^2}{4}\right)$$

but the intriguing possibility of replacing it with other kernels that may not be radially symmetric is also suggested.

Benefits of the algorithm include

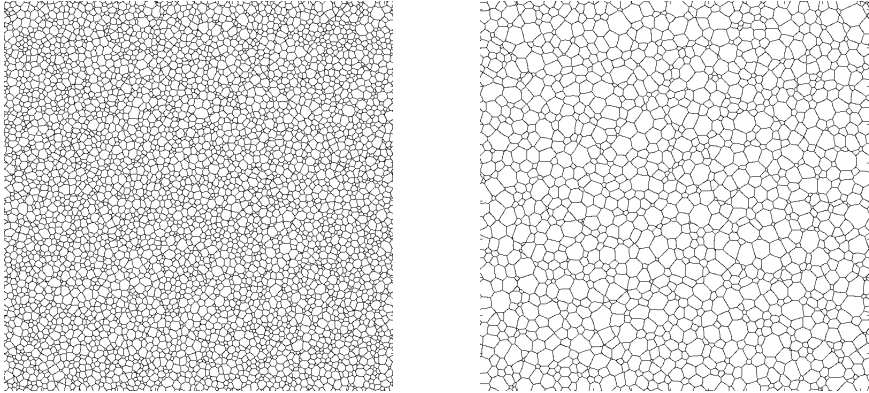


Figure 1: A large scale simulation in two dimensions at an earlier (left) and later (right) time using a variant [Elsey, Esedoğlu, and Smereka \[2009\]](#) of threshold dynamics ([Equations \(9\) and \(10\)](#)). This is the isotropic, equal surface tension and mobility case.

1. Unconditional stability: Time step size is restricted only by accuracy considerations.
2. Low per time step cost: Step (9) can be implemented on uniform grids via the fast Fourier transform at  $O(M \log M)$  cost, where  $M$  is the number of grid points. Step (10) is pointwise and costs even less.
3. All points  $x \in D$ , whether they are in the interior of a phase, on an interface  $(\partial\Sigma_i) \cap (\partial\Sigma_j)$ , or at a junction, are treated equally: No need to track or even detect surfaces or junctions. The correct Herring angle condition (all  $120^\circ$ ) is attained automatically at triple junctions.
4. As in phase field, level set, and other implicit interface methods, topological changes in the network occur with no need for intervention.

These benefits have made it possible to carry out very large scale simulations (hundreds of thousands of phases) of dynamics ([Equations \(4\) and \(6\)](#)) in both two and three dimensions – a capability desired by e.g. materials scientists interested in the statistics of shapes and sizes of grains during microstructural evolution. See [Figures 1 and 2](#) for sample computations from [Elsey, Esedoğlu, and Smereka \[2009, 2011\]](#) that used a variant of [Equations \(9\) and \(10\)](#), and [Barmak et al. \[2006\]](#) for examples of grain statistics of interest for materials scientists.

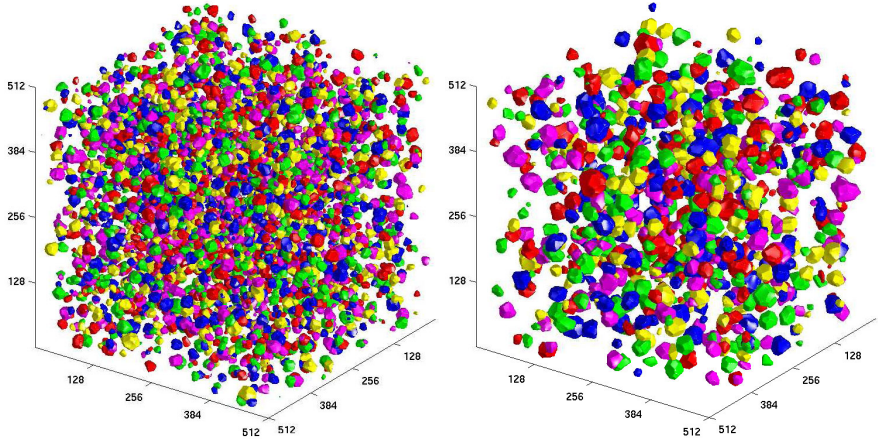


Figure 2: A large scale simulation in three dimensions using a variant [Elsley, Esedoğlu, and Smereka \[2009\]](#) of threshold dynamics ([Equations \(9\) and \(10\)](#)). Only some of the phases are shown to aid visualization. This is the isotropic, equal surface tension and mobility case. Taken from [Elsley, Esedoğlu, and Smereka \[2011\]](#).

In the two-phase setting, writing  $\Sigma = \Sigma_1$  so that  $\Sigma_2 = D \setminus \Sigma$  and taking  $K = G$ , steps ([Equations \(9\) and \(10\)](#)) can be combined to succinctly read

$$(13) \quad \Sigma^{k+1} = \left\{ x \in D : G_{\sqrt{\delta t}} * \mathbf{1}_{\Sigma^k}(x) \geq \frac{1}{2} \right\}$$

in the original form given in [Merriman, Bence, and Osher \[1992\]](#). The motivation behind [Equation \(13\)](#) is an older idea known as the phase field, or diffuse interface method: To approximate the motion by mean curvature of the boundary  $\partial\Sigma$  of a set  $\Sigma$ , one carries out gradient descent

$$(14) \quad u_t = \Delta u - \frac{1}{\varepsilon^2} W'(u)$$

for the energy

$$(15) \quad \int \frac{\varepsilon}{2} |\nabla u|^2 + \frac{1}{\varepsilon} W(u) dx$$

starting with the initial condition  $u(x, 0) = \mathbf{1}_{\Sigma}$ . Here,  $W$  is a double well potential with equal depth wells at 0 and 1, e.g.  $W(\xi) = \xi^2(1 - \xi)^2$ . The second term in [\(15\)](#) is thus a penalty term that forces  $u$  to approximate the characteristic function of a set as  $\varepsilon \rightarrow 0^+$ , while the Dirichlet energy term exacts a penalty on the rapid transition across the boundary

of the set. Following [Merriman, Bence, and Osher \[ibid.\]](#), time splitting [Equation \(14\)](#) leads to

$$(16) \quad \text{Step 1: } u_t = \Delta u, \text{ followed by Step 2: } u_t = -\frac{1}{\varepsilon^2} W'(u).$$

Step 1 explains the convolution with the Gaussian in [\(13\)](#), whereas Step 2 turns to thresholding in the limit  $\varepsilon \rightarrow 0$ : Gradient descent for the pointwise energy  $W(u)$ , represented by Step 2, ends in either one of the wells 0 or 1 of  $W$ , depending on whose basin of attraction  $u$  starts in. For  $W(\xi) = \xi^2(1 - \xi)^2$ , the basins of attraction are separated by  $\xi = \frac{1}{2}$ , which explains the threshold value of  $\frac{1}{2}$  in [\(13\)](#).

Unfortunately, this original motivation, based on time splitting the evolutionary PDE [\(14\)](#), turns out to be an inadequate explanation for even the consistency of threshold dynamics. Indeed, [\(15\)](#)  $\Gamma$ -converges to perimeter of sets [Modica and Mortola \[1977\]](#), and [\(14\)](#) approximates motion by mean curvature [Rubinstein, Sternberg, and Keller \[1989\]](#) and [Evans, Soner, and Souganidis \[1992\]](#), for e.g.  $W(\xi) = \xi^4(1 - \xi)^2$  also, the basins of attraction of which are separated by  $\frac{2}{3}$ . The naive time splitting idea then suggests [\(13\)](#) with threshold value  $\frac{1}{2}$  replaced by  $\frac{2}{3}$  as an algorithm for motion by mean curvature. However, a simple truncation error analysis (as in [Ruuth \[1996\]](#)) shows that in the limit  $\varepsilon \rightarrow 0^+$ , the resulting dynamics is *not* motion by mean curvature. We were lucky above in choosing a  $W$  that is symmetric about its local maximum. This also means that, in general, one cannot find extensions of threshold dynamics ([Equation \(13\)](#)) to more general flows simply by time splitting corresponding phase field models and sending  $\varepsilon \rightarrow 0^+$ .

However, consistency of the two-phase scheme [\(13\)](#) on smooth interfaces can be verified easily with a simple Taylor expansion. For example, in  $\mathbb{R}^2$  with  $K = G$ , take a point  $p \in \partial\Sigma$ . We may assume that  $p = 0$  and  $\partial\Sigma$  is given as the graph of a function  $f$  near 0 and is tangent to the  $x$ -axis there, as shown in [Figure 3](#). Then, according to [Mascarenhas \[1992\]](#) and [Ruuth \[1996, 1998b\]](#), Taylor expanding  $f$  at 0 in the convolution integral of [\(13\)](#) gives

$$(17) \quad (G_{\sqrt{8t}} * \mathbf{1}_\Sigma)(0, y) = \frac{1}{2} - \frac{1}{\sqrt{4\pi t}} y + \sqrt{\frac{t}{4\pi}} f''(0) + O(t)$$

as  $t \rightarrow 0$ , provided that  $y = O(t)$ . Setting [\(17\)](#) to  $\frac{1}{2}$  as scheme [\(13\)](#) prescribes, and solving for  $y$  (the new position of the interface along the normal direction at present) exhibits the curvature of the curve as the leading order contribution to normal speed.

The analogue of Taylor expansion [\(17\)](#) for general kernels  $K$  (that need not be radially symmetric) was given in [Ishii, Pires, and Souganidis \[1999\]](#) in any dimension  $d$ . When  $K \geq 0$ , the two-phase scheme [\(13\)](#) enjoys the following monotonicity property: If two different evolutions  $\Sigma^k$  and  $\Omega^k$  are generated by [Equation \(13\)](#) from the two different initial conditions  $\Sigma^0$  and  $\Omega^0$ , respectively, satisfying the ordering  $\Sigma^0 \subseteq \Omega^0$ , the same

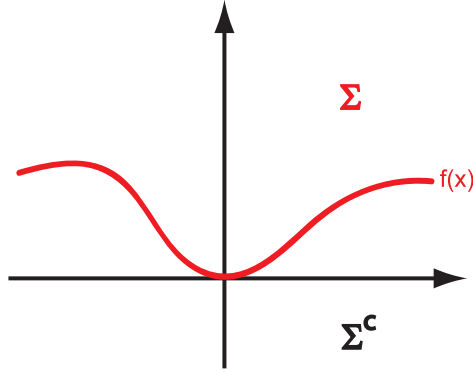


Figure 3: Consistency of the two-phase scheme (13) with curvature motion can be easily verified with a Taylor expansion.

order is preserved at later times by the algorithm:  $\Sigma^k \subseteq \Omega^k$  for all  $k$ . Combined with the consistency implied by (17), this comparison principle can be used to prove the convergence of scheme (13) to the viscosity solution of the level-set formulation of motion by mean curvature; see e.g. Evans [1993], Barles and Georgelin [1995], and Ishii, Pires, and Souganidis [1999] for the earliest rigorous convergence results for two-phase thresholding schemes. More recently, Swartz and Yip [2017] does not require the maximum principle and establishes convergence to the classical solution of two-phase mean curvature motion, with a rate.

In the multi-phase ( $N > 2$ ) setting, Ruuth [1996, 1998a] present a truncation error analysis similar to (17) in the vicinity of a triple junction in order to verify that Equations (9) and (10) imposes the correct (in this case symmetric,  $120^\circ$ ) Herring angle condition (7). This analysis also suggests an extension of Equations (9) and (10) to constant (isotropic) but unequal surface tensions. However, the resulting algorithm is considerably more complicated than the original, and contains some heuristic steps to handle multiple junctions. A natural, systematic extension of multi-phase threshold dynamics even to constant but unequal surface tensions (let alone anisotropic ones) was thus unavailable until recently.

### 3 New Algorithms: Arbitrary Surface Tensions

We will discuss the following questions:

1. What is the analogue of Equations (9) and (10) for:
  - Constant but possibly unequal surface tensions  $\sigma_{i,j} \in \mathbb{R}^+$ , and then

- The full generality of model (2), where  $\binom{n}{2}$  possibly distinct surface tension and mobility functions are specified?
2. Can we find a convolution kernel  $K$  for any given  $\sigma, \mu : \mathbb{S}^{d-1} \rightarrow \mathbb{R}^+$  pair? If so, can we ensure  $K \geq 0$  (hence two-phase monotonicity) or  $\widehat{K} \geq 0$ ?

Our starting point is a variational interpretation that was given in joint work [Esedoglu and Otto \[2015\]](#) with Felix Otto for the original threshold dynamics [Equations \(9\) and \(10\)](#). It turns out that there is a systematic way to derive elegant algorithms in the style of [Equations \(9\) and \(10\)](#) from certain non-local approximations to perimeter of sets. In the simplest two-phase setting of [\(13\)](#), consider the energy

$$(18) \quad E_\varepsilon(\Sigma) = \frac{1}{\varepsilon} \int_{\Sigma^c} K_\varepsilon * \mathbf{1}_\Sigma dx$$

Energies of this type and their limit as  $\varepsilon \rightarrow 0$  had been studied previously, e.g. in [Alberti and Bellettini \[1998\]](#), and with  $K$  the Gaussian in [Miranda, Pallara, Paronetto, and Preunkert \[2007\]](#). Called the “heat content” of the set  $\Sigma$  in [Miranda, Pallara, Paronetto, and Preunkert \[ibid.\]](#), energies  $E_\varepsilon$  converge to a multiple of the Euclidean perimeter of sets in the sense of  $\Gamma$ -convergence [Dal Maso \[1993\]](#).

As explained in [Esedoglu and Otto \[2015\]](#), [Equation \(13\)](#) can be recognized as the solution of the following optimization problem:

$$(19) \quad \Sigma^{k+1} = \arg \min_{\Sigma \subset D} E_{\sqrt{\delta t}}(\Sigma) + \frac{1}{\sqrt{\delta t}} \int (\mathbf{1}_\Sigma - \mathbf{1}_{\Sigma^k}) K_{\sqrt{\delta t}} * (\mathbf{1}_\Sigma - \mathbf{1}_{\Sigma^k}) dx$$

revealing a previously unknown connection between heat content and threshold dynamics. For any kernel  $K$  with positive Fourier transform  $\widehat{K} \geq 0$  (such as the Gaussian), the second term in [\(19\)](#) is easily seen to be positive; it also vanishes at  $\Sigma = \Sigma^k$ . It follows that

$$E_{\sqrt{\delta t}}(\Sigma^{k+1}) \leq E_{\sqrt{\delta t}}(\Sigma^k),$$

identifying [\(18\)](#) as a Lyapunov functional for scheme [\(13\)](#).

Moreover, [\(19\)](#) is reminiscent of the *minimizing movements* [De Giorgi \[1993\]](#) formulation of motion by mean curvature, due to [Almgren, Taylor, and Wang \[1993\]](#) and [Luckhaus and Sturzenhecker \[1995\]](#), the second term in [\(19\)](#) playing the role of the *movement limiter*. Indeed, it is easily verified on smooth interfaces that it measures (as  $\delta t \rightarrow 0$ ) the squared  $L^2$  norm of the normal vector field that is needed to perturb  $\Sigma^k$  to  $\Sigma$ . Along with the previously known  $\Gamma$ -convergence of energies [\(18\)](#), formulation [\(19\)](#) thus suggests very strongly that threshold dynamics [\(13\)](#) carries out gradient flow for approximately the right energy with respect to approximately the right metric.



One point is worth repeating: If all we want is a computational method to approximate the perimeter of a set, energy (18) would be a rather indirect and complicated way of doing it; certainly there are more practical and accurate methods. The reason for our interest is, as indicated above, these non-local approximations to perimeter turn out to offer a systematic way of deriving fast and elegant algorithms for curvature motion, such as (13).

**3.1 Arbitrary Isotropic Surface Tensions.** The following non-local energy is a natural candidate for approximating the surface area of  $(\partial\Sigma_i) \cap (\partial\Sigma_j)$  in the multiphase setting of (2), in the same nonlocal style as (18):

$$(20) \quad \frac{1}{\varepsilon} \int_{\Sigma_j} K_\varepsilon * \mathbf{1}_{\Sigma_i} dx$$

With  $K$  the Gaussian, for example, it measures the amount of heat that escapes from  $\Sigma_i$  to  $\Sigma_j$ , starting from the initial binary temperature distribution  $\mathbf{1}_{\Sigma_i}$ , which ought to be related to the size of the boundary between the two phases. This simple intuition leads us to the following non-local approximation for the multiphase model (2) in the isotropic case that all surface tensions  $\sigma_{i,j}$  are (possibly different) constants:

$$(21) \quad E_\varepsilon(\Sigma) = \frac{1}{\varepsilon} \sum_{\substack{i,j=1 \\ i \neq j}} \sigma_{i,j} \int_{\Sigma_j} K_\varepsilon * \mathbf{1}_{\Sigma_i} dx.$$

The analogue for (21) of the minimizing movements step (19) is

$$(22) \quad \Sigma^k = \arg \min_{\Sigma} E_{\sqrt{\delta t}}(\Sigma) - \frac{1}{\sqrt{\delta t}} \sum_{\substack{i,j=1 \\ i \neq j}}^N \sigma_{i,j} (\mathbf{1}_{\Sigma_j} - \mathbf{1}_{\Sigma_j^k}) K_{\sqrt{\delta t}} * (\mathbf{1}_{\Sigma_i} - \mathbf{1}_{\Sigma_i^k}) dx$$

the solution of which is given by the following algorithm from [Esedoğlu and Otto \[2015\]](#) which is the natural extension of the original threshold dynamics [Equations \(9\) and \(10\)](#) to isotropic, unequal surface tensions:

**Algorithm:** (from [Esedoglu and Otto \[ibid.\]](#)): Given a time step size  $\delta t > 0$ , alternate the following steps:

1. Convolution:

$$(23) \quad \psi_i^k = K_{\sqrt{\delta t}} * \sum_{j \neq i} \sigma_{i,j} \mathbf{1}_{\Sigma_j^k}.$$

2. Redistribution:

$$(24) \quad \Sigma_i^{k+1} = \left\{ x : \psi_i^k(x) \leq \min_{j \neq i} \psi_j^k(x) \right\}.$$

[Equations \(23\)](#) and [\(24\)](#) reduces to [Equations \(9\)](#) and [\(10\)](#) if  $\sigma_{i,j} = 1$  ( $i \neq j$ ). Two immediate questions concerning the new algorithm are:

1. Do the non-local energies  $E_\varepsilon$  in [\(21\)](#) approximate [\(2\)](#)?
2. Does [Equations \(23\)](#) and [\(24\)](#) decrease  $E_\varepsilon$ ?

**Theorem 1.** (from [Esedoglu and Otto \[ibid.\]](#)) Let  $K$  be the Gaussian, and let the surface tensions  $\sigma_{i,j}$  satisfy the triangle inequality [\(8\)](#). Then, as  $\varepsilon \rightarrow 0^+$ ,  $E_\varepsilon$   $\Gamma$ -converge to the appropriate formulation of [\(2\)](#) in terms of sets of finite perimeter.

Whether  $E_{\sqrt{\delta t}}(\Sigma)$  is a Lyapunov functional for scheme ([Equations \(23\)](#) and [\(24\)](#)) appears to depend (even when  $\widehat{K} \geq 0$ ) on whether the surface tension matrix  $\sigma_{i,j}$  is conditionally negative semi-definite, which is known [Schoenberg \[1938\]](#), [Avis and Deza \[1991\]](#), and [Deza and Laurent \[1997\]](#) to be related to isometric embedding of finite metric spaces in Euclidean spaces. Based on these references, we conclude [Equations \(23\)](#) and [\(24\)](#) dissipates energy [Equation \(21\)](#) if

1. There exist  $p_1, \dots, p_N \in \mathbb{R}^k$  for some  $k$  such that  $\sigma_{i,j} = |p_i - p_j|_1$ , or
2. There exist  $p_1, \dots, p_N \in \mathbb{R}^k$  for some  $k$  such that  $\sigma_{i,j} = |p_i - p_j|_2^2$ .

The latter is also necessary for the movement limiter in [\(22\)](#) to be positive.

As a more immediately applicable example of allowed surface tensions, let us consider models of grain boundary motion from materials science. In [Read and Shockley \[1950\]](#), Read and Shockley describe a well known surface tension model for a two dimensional material with a square lattice structure. Subsequently, extensions of this model to three dimensional crystallography were given, see e.g. [Holm, Hassold, and Miodownik \[2001\]](#). Let the orientations of the grains in the network be  $g_1, \dots, g_N \in SO(3)$ , describing the

rotations needed to map a reference cubic lattice to those of the grains. Then, the surface tension of the interface between grains  $\Sigma_i$  and  $\Sigma_j$  is given by

$$(25) \quad \sigma_{i,j} = \begin{cases} \frac{\theta}{\theta_*} \left(1 - \log\left(\frac{\theta}{\theta_*}\right)\right) & \text{if } \theta < \theta_* \\ \theta_* & \text{if } \theta \geq \theta_*. \end{cases} \quad \text{with } \theta = \min_{r \in \mathcal{O}} \arccos\left(\frac{\text{tr}(r g_j g_i^{-1}) - 1}{2}\right)$$

where  $\mathcal{O}$  is the octahedral group of symmetries of the cube in  $d = 3$ , and  $\theta_*$  is a cut-off value. The angle  $\theta$  represents the minimum angle needed to rotate the lattice of grain  $\Sigma_i$  to that of grain  $\Sigma_j$ ; Read and Shockley tell us that  $\sigma_{i,j}$  is the specific function shown of this angle. We have

**Theorem 2.** (from [Esedoglu and Otto \[2015\]](#)) *Let the surface tensions  $\sigma_{i,j}$  be determined from orientations  $g_i \in SO(3)$  of the grains by the Read and Shockley model (25). Then, movement limiter in (22) is positive, so that energy (21) is a Lyapunov function for [Equations \(23\) and \(24\)](#).*

Theorems 1 and 2 do not establish convergence of the *evolution* generated by [Equations \(9\) and \(10\)](#) or [\(23\) and \(24\)](#) to their intended limits. This was recently achieved by Laux and Otto in [Laux and Otto \[2016\]](#) and [Laux and Otto \[2017\]](#). In a culmination of the minimizing movements formulation (22) of threshold dynamics, they obtain the first convergence result for the *multi-phase* dynamics generated by thresholding algorithms ([Equations \(9\) and \(10\)](#)) and [Equations \(23\) and \(24\)](#). Roughly speaking, their result says

**Theorem 3.** (from [Laux and Otto \[2016\]](#)) *Given a sequence of  $\delta t \rightarrow 0$ , the piecewise constant in time extensions of the discrete in time approximations generated by [Equations \(23\) and \(24\)](#) have a subsequence that converges. If the time integral of their energies converge to that of the limit, then the limit solves the multi-phase version of the weak formulation of mean curvature motion given in [Luckhaus and Sturzenhecker \[1995\]](#).*

More recently, [Laux and Otto \[2017\]](#) establishes that this limit is a solution of motion by mean curvature also in Brakke's sense [Brakke \[1978\]](#). These are the first rigorous convergence results on practical numerical algorithms for multi-phase motion by mean curvature, persisting through possible topological changes.

**3.2 Anisotropic Surface Tensions and Mobilities.** Generalizations of Merriman, Bence, and Osher's [Equations \(9\) and \(10\)](#) to anisotropic surface energies had been considered in a number of works previously, though always in the two-phase setting.

One of the first contributions to the study of [Equation \(13\)](#) with general convolution kernels  $K$  (replacing  $G$ ) is by [Ishii, Pires, and Souganidis \[1999\]](#), who establish the convergence of the algorithm to the viscosity solution of the equation  $u_t = F(D^2u, Du)$

where

$$(26) \quad F(M, p) = \left( \int_{p^\perp} K(x) dS(x) \right)^{-1} \left( \frac{1}{2} \int_{p^\perp} \langle Mx, x \rangle K(x) dS(x) \right)$$

for  $p \in \mathbb{R}^d$  and  $M$  a  $d \times d$  symmetric matrix, provided that  $K$  is a *positive* convolution kernel with certain additional decay and continuity properties. Positivity of the kernel is required for the scheme to preserve the comparison principle that applies to the underlying interfacial motion, and is essential for the viscosity solutions approach taken in [Ishii, Pires, and Souganidis \[ibid.\]](#) (but the consistency calculation given in the paper applies to more general – i.e. sign changing – kernels).

Yet [Ishii, Pires, and Souganidis \[ibid.\]](#) does not address the inverse problem of constructing a convolution kernel for a given surface tension – mobility pair, which is perhaps the more practical question. The first contribution in this direction is by [Ruuth and Merriman \[2000\]](#), who propose a construction in  $\mathbb{R}^2$ . They show how to construct a kernel (characteristic function of a judiciously chosen star shaped domain) that, when used in (13), would generate a normal speed of the form

$$(27) \quad v_\perp(x) = (f''(\theta(x)) + f(\theta(x))) \kappa(x)$$

for a desired  $f : [0, 2\pi] \rightarrow \mathbb{R}^+$ , where  $\theta(x)$  is the angle that the normal at  $x \in \partial\Sigma$  makes with the positive  $x$ -axis. However, there are infinitely many surface tension and mobility pairs  $(\sigma, \mu)$  that correspond to the same  $f$  and hence the same normal speed in (27); the discussion in [Ruuth and Merriman \[ibid.\]](#) does not elucidate what the two factors  $\sigma$  and  $\mu$  are for their kernel construction. This is particularly important in the multi-phase setting, since surface tensions determine the angles (6) at junctions.

More recently, Bonnetier et. al. [Bonnetier, Bretin, and A. Chambolle \[2012\]](#) have proposed a construction that works in both  $\mathbb{R}^2$  and  $\mathbb{R}^3$ . The Fourier transform of their kernels is explicit in terms of the surface tension:

$$(28) \quad \widehat{K}(\xi) = \exp(-\sigma^2(\xi)).$$

It turns out that the corresponding mobility satisfies  $\mu = \sigma$ , an important but very special case. This construction often yields sign changing kernels, even in two dimensions, preventing the authors from giving a rigorous proof of convergence. Moreover, as soon as the anisotropy  $\sigma$  does not have an ellipsoid as its unit ball, (28) has a singularity at the origin, leading to slow decay of  $K$ .

The variational formulation (19) of threshold dynamics and its multiphase extension (22) prove particularly helpful with questions of anisotropy. For example, simply by evaluating the limit as  $\varepsilon \rightarrow 0^+$  of energy (18) on a set  $\Sigma$  with smooth boundary, we are led to

the following natural candidate for the surface tension associated with a given kernel  $K$ :

$$(29) \quad \sigma_K(n) := \frac{1}{2} \int_{\mathbb{R}^d} |n \cdot x| K(x) dx.$$

Likewise, evaluating the *movement limiter* in the minimizing movements formulation (19) on smooth interfaces yields the following natural candidate for the mobility associated with  $K$ :

$$(30) \quad \frac{1}{\mu_K(n)} := \int_{n^\perp} K(x) dS(x).$$

It can be verified [Elsey and Esedoğlu \[2017\]](#) on smooth interfaces that threshold dynamics (13) is consistent with the normal speed (4) where  $\sigma$  and  $\mu$  in (4) are given by (29) and (30). Formula (29) can be expressed in terms of the cosine transform  $\mathcal{T}$  for even functions on  $\mathbb{S}^d$ :

$$(31) \quad \sigma_K(n) = \mathcal{T} \omega_K := \int_{\mathbb{S}^{d-1}} \omega_K(x) |n \cdot x| dS(x)$$

where, according to (29),  $\omega_K$  is given by

$$(32) \quad \omega_K(x) = \frac{1}{2} \int_0^\infty K(rn) r^d dr$$

$\omega_K$  is known as the generating function of the anisotropy  $\sigma_K$ . Formula (30) can alternatively be written using the spherical Radon transform  $\mathfrak{g}_s$ :

$$(33) \quad \frac{1}{\mu_K} = \mathfrak{g}_s \int_0^\infty K(rn) r^{d-2} dr.$$

Also helpful are the following expressions of (29) and (30) in terms of the Fourier transform of the kernel  $K$ :

$$(34) \quad \begin{aligned} \sigma_K(n) &= -\frac{1}{2\pi} \text{F. P.} \int_{\mathbb{R}} \frac{\widehat{K}(n\xi)}{\xi^2} d\xi, \text{ and} \\ \mu_K(n) &= 2\pi \left( \int_{\mathbb{R}} \widehat{K}(n\xi) d\xi \right)^{-1}. \end{aligned}$$

These formulas allow us to draw upon existing results concerning the positivity of inverse cosine and inverse spherical Radon transforms in the convex geometry literature. For example, it is known that the generating function  $\mathcal{T}^{-1}\sigma$  of an anisotropy  $\sigma$  is always positive in  $\mathbb{R}^2$ , but may be negative for certain anisotropies in  $\mathbb{R}^3$  [Goodey and Weil \[1992\]](#) and

[Bolker \[1969\]](#). Those  $\sigma$  for which  $\mathcal{T}^{-1}\sigma$  is positive have a nice geometric characterization: The unit ball of the dual norm, known as the Wulff shape  $W_\sigma$  of the anisotropy  $\sigma$ , is a *zonoid*. Zonoids are convex bodies that are the limits with respect to the Hausdorff distance of centrally symmetric polytopes each face of which are also centrally symmetric [Goodey and Weil \[1992\]](#). For example, in  $\mathbb{R}^3$ , there is a neighborhood of the octahedron that contains no zonoids. On the other hand, (32) tells us that the corresponding anisotropy  $\sigma_K$  of any *positive* convolution kernel  $K$  in threshold dynamics [Equation \(13\)](#) must have a positive generating function  $\omega_K$  and hence has to be zonoidal.

Moreover, it turns out there are restrictions on the attainable mobilities with positive kernels as well. Via (33), this matter is clearly related to positivity properties of the inverse spherical Radon transform  $\mathfrak{g}_s^{-1}$ , which also appears prominently in the convex geometry literature, especially in the context of the Busemann-Petty problem [Busemann and Petty \[1956\]](#) and [Gardner \[1994b\]](#). Indeed, results given in [Gardner \[1994a\]](#) on  $\mathfrak{g}_s^{-1}$ , together with the foregoing discussion, yields the following limitation of threshold dynamics schemes (under assumptions on  $\sigma$  and  $\mu$  from the Introduction):

For certain surface tensions  $\sigma$  in  $\mathbb{R}^3$ , it is not possible to design a threshold dynamics [Equation \(13\)](#) that preserves the two-phase comparison principle. In particular, unless the Wulff shape of the anisotropy  $\sigma$  is the dilation of a zonoid by a sphere, there is no monotone threshold dynamics scheme for it.

It is interesting to compare with an alternative approach due to Chambolle and Novaga [M. Chambolle A. N. \[2006\]](#) for generating weighted motion by mean curvature that was inspired by threshold dynamics. Their idea is to replace the convolution (9) in the original [Equation \(13\)](#) with the solution of a nonlinear parabolic PDE:

<u>Threshold Dynamics</u>	<u>Nonlinear Threshold Dynamics</u>
Step 1: Convolution:	Step 1: Nonlinear Diffusion
$\psi^k = K_{\sqrt{\delta t}} * \mathbf{1}_{\Sigma^k}$	$\begin{cases} \partial_t \psi^k = \nabla \cdot (\sigma(\nabla \psi^k) \nabla \sigma(\nabla \psi^k)) \\ \psi^k(x, 0) = \mathbf{1}_{\Sigma^k}(x). \end{cases}$
Step 2: Thresholding:	Step 2: Thresholding:
$\Sigma^{k+1} = \left\{ x : \psi^k(x) \geq \frac{1}{2} \int_{\mathbb{R}^d} K(x) dx \right\}$	$\Sigma^{k+1} = \left\{ x : \psi^k(x, \delta t) \geq \frac{1}{2} \right\}$

The Chambolle and Novaga scheme preserves the comparison principle for any anisotropy  $\sigma$ , since the nonlinear diffusion equation in Step 1 of their algorithm enjoys this principle. Their Step 1, however, is more costly than the simple convolution involved in the

corresponding step of the original threshold dynamics scheme. It thus appears that the variational formulation (19) along with formulas (Equations (29) and (30)) suggest the following guideline in searching for diffusion generated motion algorithms with various desirable properties:

If we want to get away with just convolutions and avoid the costly solution of nonlinear PDE, we have to give up *something*: Namely, the two-phase comparison principle, for certain anisotropies in three dimensions.

Formulas (29) and (30), (32) and (33), and (34) also help explore how to *construct* a convolution kernel  $K$  to be used in threshold dynamics Equation (13) to generate motion (4) with a given desired surface tension and mobility pair  $\sigma, \mu : \mathbb{S}^{d-1} \rightarrow \mathbb{R}^+$ . For example, we can look for a kernel that in polar coordinates has the form

$$(35) \quad K(r, \theta) = \alpha(\theta, \phi) \eta(\beta(\theta, \phi) r)$$

where  $\eta : \mathbb{R} \rightarrow \mathbb{R}$  is any smooth, positive, non-zero function supported in  $[1, 2]$ . Substituting (35) in equations (32) and (33) yields simple, pointwise equations for  $\alpha(\theta, \phi)$  and  $\beta(\theta, \phi)$  in terms of  $\mathcal{T}^{-1}\sigma$  and  $\mathfrak{J}_s^{-1}(\frac{1}{\mu})$  which, in case  $d = 2$ , can always be solved with  $\alpha \geq 0$ :

**Theorem 4.** (from *Esedoglu, Jacobs, and Zhang [2017]*) *In  $\mathbb{R}^2$ , for any desired surface tension  $\sigma : \mathbb{S}^1 \rightarrow \mathbb{R}^+$  and any desired mobility  $\mu : \mathbb{S}^1 \rightarrow \mathbb{R}^+$ , there exists a smooth, positive, compactly supported convolution kernel  $K$  such that  $\sigma_K$  and  $\mu_K$  given by (29) and (30) satisfy  $\sigma_K = \sigma$  and  $\mu_K = \mu$ , so that threshold dynamics Equation (13) generates the corresponding motion by weighted curvature (4) and satisfies the comparison principle. This is also possible in  $\mathbb{R}^3$ , provided that the Wulff shape of  $\sigma$  is the dilation of a zonoid by a sphere.*

If we give up the two-phase comparison principle (as we must in general in  $\mathbb{R}^3$ ), we can turn to (34) and look for e.g. kernels of the form

$$(36) \quad \widehat{K}(\xi) = \exp\left(-\zeta(\alpha(\xi))\right) + \exp\left(-\zeta(\beta(\xi))\right).$$

where  $\zeta : \mathbb{R} \rightarrow \mathbb{R}^+$  is smooth, even, with  $\zeta(x) = 0$  for  $|x| \leq 1$  and  $\zeta(x) = x^2$  for  $|x| \geq 2$ . Once again, one gets simple pointwise equations for  $\alpha$  and  $\beta$  that can always be solved *Esedoglu, Jacobs, and Zhang [ibid.]*, yielding a version of (28) that allows baking the mobility as well as the surface tension into a kernel with positive Fourier transform.

The immediate analogue of our non-local energies (21) in the multi-phase anisotropic setting is

$$(37) \quad E_\varepsilon(\Sigma) = \frac{1}{\varepsilon} \sum_{\substack{i,j=1 \\ i \neq j}} \int_{\Sigma_j} (K_{i,j})_\varepsilon * \mathbf{1}_{\Sigma_i} dx$$

along with the following analogue of [Equations \(23\) and \(24\)](#):

**Algorithm:** (from [Elsey and Esedoğlu \[2017\]](#)) Alternate the following steps:

1. Convolution:

$$(38) \quad \psi_i^k = \sum_{j \neq i} (K_{i,j})_{\sqrt{\delta t}} * \mathbf{1}_{\Sigma_j^k}.$$

2. Thresholding:

$$(39) \quad \Sigma_i^{k+1} = \left\{ x : \psi_i^k(x) \leq \min_{j \neq i} \psi_j^k(x) \right\}.$$

The  $\binom{n}{2}$  surface tensions and mobilities can be baked into the kernels  $K_{i,j}$  by one of the new constructions [\(35\)](#) or [\(36\)](#). We repeat that in  $\mathbb{R}^2$ , there are infinitely many surface tension - mobility pairs  $(\sigma_{i,j}, \mu_{i,j})$  corresponding to the same normal speed [\(5\)](#) for each interface in the network. However, if kernels  $K$  are more stringently chosen by specifying their surface tensions and mobilities separately via [\(29\)](#) and [\(30\)](#) by e.g. the new kernel construction [\(35\)](#), numerical experiments show that in addition to achieving the correct normal speed along interfaces, threshold dynamics attains the correct angle conditions [\(6\)](#) at junctions, as the variational formulation [\(22\)](#) suggests. [Figure 4](#) shows a three-phase simulation using [\(38\)](#) and [\(39\)](#) from [Esedoğlu, Jacobs, and Zhang \[2017\]](#) where the kernels are constructed via [\(35\)](#) corresponding to the following surface tension-mobility pairs:

$$(40) \quad \sigma_{1,2}(x_1, x_2) = \sqrt{x_1^2 + x_2^2} \quad \mu_{1,2}(x_1, x_2) = 1,$$

$$(41) \quad \sigma_{1,3}(x_1, x_2) = \sqrt{\frac{1}{4}x_1^2 + x_2^2} + \sqrt{x_1^2 + \frac{1}{4}x_2^2} \quad \mu_{1,3}(x_1, x_2) = \frac{2x_1^2 + 3x_2^2}{4\sqrt{x_1^2 + x_2^2}}$$

$$(42) \quad \sigma_{2,3}(x_1, x_2) = \sqrt{x_1^2 + \frac{25}{16}x_2^2} \quad \mu_{2,3}(x_1, x_2) = 1.$$

Although [Equations \(38\) and \(39\)](#) thus appears to work as expected, we do not know sufficiently general conditions to be of interest on the surface tensions  $\sigma_{i,j}$  that would ensure an analogue of [Theorem 2](#), guaranteeing dissipation of energy [\(37\)](#). However, we can come up with slightly slower versions of [Equations \(38\) and \(39\)](#) for which an analogue of [Theorem 2](#) can be easily shown to hold. The idea is to refresh convolutions more frequently during the course of a single time step. In the interest of brevity, let us consider the two-phase setting as an example. The following analogue of the original threshold dynamics scheme [\(13\)](#) requires two convolutions per time step vs. one:



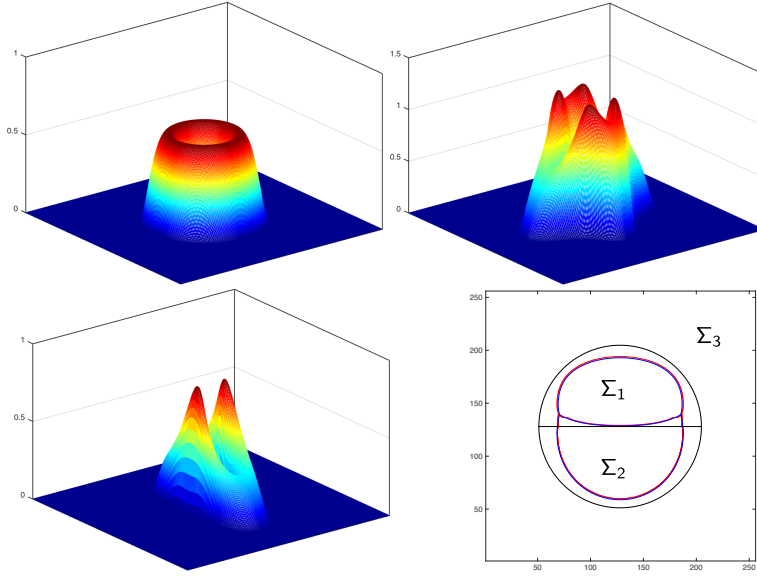


Figure 4: Kernels constructed using formulas (29) and (30) for the prescribed  $(\sigma_{i,j}, \mu_{i,j})$  pairs (40), (41), (42), and a sample three-phase simulation: The black curves are the initial condition, the red curves are by Equations (38) and (39) using the kernels shown, and the blue curves are a benchmark result using *front tracking* Bronsard and Wetton [1995] and Kinderlehrer, Livshitz, and Taasan [2006] – a very accurate method that can have difficulties with topological changes. Being able to *bake in* a target surface tension *and* mobility of an interface into the convolution kernel of threshold dynamics is a new capability elucidated by the variational formulation (19) and (22). Taken from Eshedoglu, Jacobs, and Zhang [2017].

**Algorithm:** (from Eshedoglu and Jacobs [2017]) Alternate the following steps:

1. 1st Convolution:

$$(43) \quad \psi^{k+\frac{1}{2}} = K_{\sqrt{\delta t}} * \mathbf{1}_{\Sigma^k}$$

2. 1st Thresholding:

$$(44) \quad \Sigma^{k+\frac{1}{2}} = \Sigma^k \cup \left\{ x : \psi^{k+\frac{1}{2}}(x) \geq \frac{1}{2} \right\}.$$

3. 2nd Convolution:

$$(45) \quad \psi^{k+1} = K_{\sqrt{\delta t}} * \mathbf{1}_{\Sigma^{k+\frac{1}{2}}}$$

Unlike (13), this slightly more costly version dissipates energy (18) for a much wider class of convolution kernels, e.g. any kernel  $K$  of the form  $K = f + g$  where  $f \geq 0$  and  $\widehat{g} \geq 0$ . This additional ease in establishing stability extends to multiple phases, so that similarly slowed down (but still unconditionally gradient stable) versions of Equations (38) and (39) are given in Esedoğlu and Otto [2015], Esedoğlu and Jacobs [2017], and Esedoğlu, Jacobs, and Zhang [2017] under a variety of assumptions on the convolution kernels that include the new constructions (35) and (36) that allow baking anisotropic surface tensions and mobilities simultaneously into convolution kernels.

## References

- G. Alberti and G. Bellettini (1998). “A non-local anisotropic model for phase transitions: asymptotic behavior of rescaled energies.” *European J. Appl. Math.* 9, pp. 261–284 (cit. on p. 3973).
- F. Almgren, J. E. Taylor, and L.-H. Wang (1993). “Curvature-driven flows: a variational approach”. *SIAM Journal on Control and Optimization* 31.2, pp. 387–438 (cit. on p. 3973).
- L. Ambrosio and A. Braides (1990). “Functionals defined on partitions in sets of finite perimeter II: semicontinuity, relaxation, and homogenization.” *J. Math. Pures et Appl.* 69, pp. 307–333. (Cit. on p. 3968).
- D. Avis and M. Deza (1991). “The cut cone,  $L^1$  embeddability, complexity, and multi-commodity flows.” *Networks* 6, pp. 595–617 (cit. on p. 3975).
- G. Barles and C. Georgelin (1995). “A simple proof of convergence for an approximation scheme for computing motions by mean curvature”. *SIAM J. Numer. Anal.* 32, pp. 484–500 (cit. on p. 3972).
- K. Barmak et al. (2006). “Grain boundary energy and grain growth in Al films: Comparison of experiments and simulations”. *Scripta Materialia* 54, pp. 1059–1063 (cit. on p. 3969).
- E. D. Bolker (1969). “A class of convex bodies”. *Transactions of the American Mathematical Society* 145, pp. 323–345 (cit. on p. 3978).
- E. Bonnetier, E. Bretin, and A. Chambolle (2012). “Consistency result for a non-monotone scheme for anisotropic mean curvature flow”. *Interfaces and Free Boundaries* 14.1, pp. 1–35 (cit. on p. 3977).
- K. A. Brakke (1978). *The motion of a surface by its mean curvature*. Vol. 20. Princeton University Press (cit. on p. 3976).
- L. Bronsard and B. Wetton (1995). “A numerical method for tracking curve networks moving with curvature motion”. *Journal of Computational Physics* 120.1, pp. 66–87 (cit. on p. 3982).

- H. Busemann and C. M. Petty (1956). “Problems on convex bodies”. *Mathematica Scandinavica* 4, pp. 88–94 (cit. on p. 3979).
- M. Chambolle A.; Novaga (2006). “Convergence of an algorithm for the anisotropic and crystalline mean curvature flow”. *SIAM Journal on Mathematical Analysis* 37.6, pp. 1978–1987 (cit. on p. 3979).
- T. F. Chan and L. Vese (Feb. 2001). “Active contours without edges”. *IEEE Transactions on Image Processing* 10.2, pp. 266–277 (cit. on p. 3966).
- G. Dal Maso (1993). *An Introduction to Gamma Convergence*. Progress in Nonlinear Differential Equations and their Applications 8. Birkhauser (cit. on p. 3973).
- E. De Giorgi (1993). “New problems on minimizing movements”. In: *Boundary value problems for PDE and applications*. Ed. by C. Baiocchi and J. L. Lions. Masson, pp. 81–98 (cit. on p. 3973).
- M. Deza and M. Laurent (1997). *Geometry of cuts and metrics*. Springer (cit. on p. 3975).
- M. Elsey and S. Esedoğlu (2017). “Threshold dynamics for anisotropic surface energies”. *Mathematics of Computation* In press. (Cit. on pp. 3978, 3981).
- M. Elsey, S. Esedoğlu, and P. Smereka (2009). “Diffusion generated motion for grain growth in two and three dimensions”. *Journal of Computational Physics* 228:21, pp. 8015–8033 (cit. on pp. 3969, 3970).
- (2011). “Large scale simulations of normal grain growth via diffusion generated motion”. *Proceedings of the Royal Society A: Mathematical, Physical, and Engineering Sciences* 467:2126, pp. 381–401 (cit. on pp. 3969, 3970).
- S. Esedoğlu and M. Jacobs (Jan. 2017). “Convolution kernels and stability of threshold dynamics methods”. *SIAM Journal on Numerical Analysis* 55, pp. 2123–2150 (cit. on pp. 3982, 3983).
- S. Esedoğlu, M. Jacobs, and P. Zhang (2017). “Kernels with prescribed surface tension & mobility for threshold dynamics schemes”. *Journal of Computational Physics* 337, pp. 62–83 (cit. on pp. 3980–3983).
- S. Esedoğlu and F. Otto (2015). “Threshold dynamics for networks with arbitrary surface tensions”. *Communications on Pure and Applied Mathematics* 68.5, pp. 808–864 (cit. on pp. 3973–3976, 3983).
- L. C. Evans (1993). “Convergence of an algorithm for mean curvature motion.” *Indiana University Mathematics Journal* 42, pp. 553–557 (cit. on p. 3972).
- L. C. Evans, H. M. Soner, and P. E. Souganidis (1992). “Phase transitions and generalized motion by mean curvature”. *Communications on Pure and Applied Mathematics* 45, pp. 1097–1123 (cit. on p. 3971).
- R. J. Gardner (1994a). “A positive answer to the Busemann-Petty problem in three dimensions”. *Annals of Mathematics* 140, pp. 435–447 (cit. on p. 3979).
- (1994b). “Intersection bodeis and the Busemann-Petty problem”. *Transactions of the American Mathematical Society* 342.1, pp. 435–445 (cit. on p. 3979).

- Paul Goodey and Wolfgang Weil (1992). “Centrally symmetric convex bodies and the spherical Radon transform”. *Journal of Differential Geometry* 35, pp. 675–688 (cit. on pp. [3978](#), [3979](#)).
- C. Herring (1951). “The Physics of Powder Metallurgy”. In: ed. by W. Kingston. McGraw Hill. Chap. Surface tension as a motivation for sintering, pp. 143–179 (cit. on p. [3967](#)).
- E. A. Holm, G. N. Hassold, and M. A. Miodownik (2001). “On misorientation distribution evolution during anisotropic grain growth”. *Acta Materialia* 49, pp. 2981–2991 (cit. on p. [3975](#)).
- H. Ishii, G. E. Pires, and P. E. Souganidis (1999). “Threshold dynamics type approximation schemes for propagating fronts”. *Journal of the Mathematical Society of Japan* 51, pp. 267–308 (cit. on pp. [3971](#), [3972](#), [3976](#), [3977](#)).
- D. Kinderlehrer, I. Livshitz, and S. Taasan (2006). “A variational approach to modeling and simulation of grain growth”. *SIAM Journal on Scientific Computing* 28.5, pp. 1694–1715 (cit. on p. [3982](#)).
- T. Laux and F. Otto (2016). “Convergence of the thresholding scheme for multi-phase mean-curvature flow”. *Calculus of Variations and Partial Differential Equations* 55.5, pp. 1–74 (cit. on p. [3976](#)).
- (2017). “[Brakke’s inequality for the thresholding scheme](#)”. arXiv: [1708.03071](#) (cit. on p. [3976](#)).
- S. Luckhaus and T. Sturzenhecker (1995). “Implicit time discretization for the mean curvature flow equation”. *Calculus of Variations and Partial Differential Equations* 3(2), pp. 253–271 (cit. on pp. [3973](#), [3976](#)).
- P. Mascarenhas (July 1992). *Diffusion generated motion by mean curvature*. CAM Report 92-33. (URL = <http://www.math.ucla.edu/applied/cam/index.html>). UCLA (cit. on p. [3971](#)).
- B. Merriman, J. K. Bence, and S. J. Osher (1992). “Diffusion generated motion by mean curvature”. In: *Proceedings of the Computational Crystal Growers Workshop*. Ed. by J. Taylor. AMS, pp. 73–83 (cit. on pp. [3968](#), [3970](#), [3971](#)).
- (1994). “Motion of multiple junctions: a level set approach”. *Journal of Computational Physics* 112.2, pp. 334–363 (cit. on p. [3968](#)).
- M. Miranda, D. Pallara, F. Paronetto, and M. Preunkert (2007). “Short-time heat flow and functions of bounded variation in  $\mathbb{R}^N$ ”. *Ann. Fac. Sci. Toulouse, Mathématiques* 16.1, pp. 125–145 (cit. on p. [3973](#)).
- L. Modica and S. Mortola (1977). “Un esempio di Gamma-convergenza”. *Boll. Un. Mat. Ital. B (5)* 14.1, pp. 285–299 (cit. on p. [3971](#)).
- W. W. Mullins (1956). “Two dimensional motion of idealized grain boundaries”. *J. Appl. Phys.* 27, pp. 900–904 (cit. on p. [3966](#)).

- D. Mumford and J. Shah (1989). “Optimal approximations by piecewise smooth functions and associated variational problems”. *Communications on Pure and Applied Mathematics* 42, pp. 577–685 (cit. on p. 3966).
- W. T. Read and W. Shockley (1950). “Dislocation models of crystal grain boundaries”. *Physical Review* 78:3, pp. 275–289 (cit. on p. 3975).
- J. Rubinstein, P. Sternberg, and J. B. Keller (1989). “Fast reaction, slow diffusion, and curve shortening”. *SIAM Journal on Applied Mathematics* 49.1, pp. 116–133 (cit. on p. 3971).
- S. J. Ruuth (1996). “Efficient algorithms for diffusion-generated motion by mean curvature”. PhD thesis. The University of British Columbia (cit. on pp. 3971, 3972).
- (1998a). “A diffusion generated approach to multiphase motion”. *Journal of Computational Physics* 145, pp. 166–192 (cit. on p. 3972).
- (1998b). “Efficient algorithms for diffusion-generated motion by mean curvature”. *Journal of Computational Physics* 144, pp. 603–625 (cit. on p. 3971).
- S. J. Ruuth and B. Merriman (2000). “Convolution generated motion and generalized Huygens’ principles for interface motion”. *SIAM Journal on Applied Mathematics* 60, pp. 868–890 (cit. on p. 3977).
- I. J. Schoenberg (1938). “Metric spaces and positive definite functions”. *Transactions of the American Mathematical Society* 44.3, pp. 522–536 (cit. on p. 3975).
- D. Swartz and N. K. Yip (2017). “Convergence of diffusion generated motion to motion by mean curvature”. *Communications in Partial Differential Equations* 42.10, pp. 1598–1643 (cit. on p. 3972).
- L. Vese and T. F. Chan (2002). “A multiphase level set framework for image segmentation using the Mumford and Shah model”. *International Journal of Computer Vision* 50.3, pp. 271–293 (cit. on p. 3966).

Received 2017-11-29.

SELIM ESEDOĞLU  
[esedoglu@umich.edu](mailto:esedoglu@umich.edu)

# SYMMETRY, INVARIANCE AND THE STRUCTURE OF MATTER

RICHARD D. JAMES

## Abstract

We present a mathematical view of the structure of matter based on the invariance of the classical equations of physics.

## Contents

<b>1</b>	<b>Symmetry and invariance</b>	<b>3986</b>
<b>2</b>	<b>The Periodic Table</b>	<b>3986</b>
<b>3</b>	<b>Objective structures</b>	<b>3988</b>
<b>4</b>	<b>An invariant manifold of molecular dynamics</b>	<b>3992</b>
<b>5</b>	<b>Continuum and structural mechanics</b>	<b>3997</b>
<b>6</b>	<b>Boltzmann equation</b>	<b>4002</b>
<b>7</b>	<b>Maxwell's equations</b>	<b>4004</b>
<b>8</b>	<b>Perspective</b>	<b>4009</b>

---

This work was supported by AFOSR (FA9550-15-1-0207), NSF (DMREF-1629026), ONR (N00014-14-1-0714), and the MURI program (FA9550-12-1-0458, FA9550-16-1-0566).

*MSC2010:* primary 82-02; secondary 20H15, 35Q60, 51F25, 70H33, 78A45.

*Keywords:* Atomic structure, symmetry, objective structures, isometry groups, molecular dynamics, invariant manifold, Boltzmann equation, homoenergetic solutions, Maxwell's equations, non-equilibrium statistical mechanics, X-ray methods.

## 1 Symmetry and invariance

Physicists and mathematicians have long tried to understand the structure of matter from a deductive viewpoint. Early examples are Hooke’s *Micrographia* [Hooke \[1665\]](#) and, inspired in part by microscopic observations, Euler’s “Physical investigations on the nature of the smallest parts of matter” [Euler \[1745\]](#). As the incredible difficulty of achieving rigorous results in this direction became better appreciated, the problem was narrowed to the “crystallization problem”: that is, prove for the simplest models of atomic forces that the Face-Centered Cubic lattice (FCC, defined below) minimizes the potential energy. Inspired by the seminal work of [Gardner and Radin \[1979\]](#) and also relying on recent advances in the calculus of variations, research on the crystallization problem has achieved significant advances [Friesecke and Theil \[2002\]](#), [Theil \[2006\]](#), and [Flatley and Theil \[2015\]](#). In these works the symmetry of the FCC lattice and the invariance of the underlying equations play a dominant role.

Our purpose is not to survey these advances, but rather to broaden the discussion by collecting a list of examples in which structure and invariance are intimately related. There are three benefits: 1) a treasure trove of interesting mathematical problems is revealed, 2) modern research on nanoscience is given a mathematical perspective, and 2) one realizes that the subject is more about invariance than structure.

## 2 The Periodic Table

We start at the most basic level: the Periodic Table of the elements. Most people think of the crystal structures of the elements in terms of Bravais lattices, and the standard databases are organized on this basis. A Bravais lattice is the infinite set of points  $\mathcal{L}(e_1, e_2, e_3) = \{v^1 e_1 + v^2 e_2 + v^3 e_3 : (v^1, v^2, v^3) \in \mathbb{Z}^3\}$ , where  $e_1, e_2, e_3$  are linearly independent vectors in  $\mathbb{R}^3$  called *lattice vectors*.

For example, consider lattice vectors  $e_1 = \alpha \hat{e}_1$ ,  $e_2 = \alpha \hat{e}_2$  and  $e_3 = \alpha(\hat{e}_1 + \hat{e}_2 + \gamma \hat{e}_3)/2$  where  $\hat{e}_1, \hat{e}_2, \hat{e}_3 = \hat{e}_1 \times \hat{e}_2$  are orthonormal and  $\alpha, \gamma > 0$ . The constants  $\alpha, \gamma$  that quantify the distances between atoms are called *lattice parameters*. The value  $\gamma = 1$  gives the Body-Centered Cubic (BCC) lattice. A famous observation of [Bain \[1924\]](#) is that there is exactly one other choice of  $\gamma > 0$  in which the associated Bravais lattice has cubic symmetry, that being  $\gamma = \sqrt{2}$ , which in fact gives the FCC lattice. About half of the Periodic Table consists of elements whose normal crystal structure at room temperature is either BCC or FCC. In fact, Bain theorized that best represented phase transformation in the Periodic Table, BCC→FCC, is achieved by passing  $\gamma$  from 1 to  $\sqrt{2}$ .

How about the other half? To discuss this more precisely, let us remove the last row of the Periodic Table, atomic numbers 87-118, which are typically radioactive and often highly unstable, and also number 85 (Astatine), for which there exists much less than 1

gram in the earth's crust at any one time and cannot be considered to have a bulk crystal structure. For definiteness, we take the accepted most common crystal structure at room temperature, unless the material is not solid at room temperature, in which case we take the accepted structure at zero temperature. Many (but not all) of the other half are 2-lattices, i.e., the union of two displaced Bravais lattices made with the same lattice vectors:

$$(1) \quad \{a + \mathcal{L}(e_1, e_2, e_3)\} \cup \{b + \mathcal{L}(e_1, e_2, e_3)\},$$

where  $a \neq b \in \mathbb{R}^3$ , or equivalently, the periodic extension of two atomic positions  $a, b$  using the periodicity  $e_1, e_2, e_3$ . For example, the third most prominent structure in the Periodic Table is the Hexagonal Close Packed (HCP) lattice for which we can choose  $e_1 = \sqrt{3}\alpha\hat{e}_1$ ,  $e_2 = \sqrt{3}\alpha((1/2)\hat{e}_1 + (\sqrt{3}/2)\hat{e}_2)$ ,  $e_3 = 2\alpha\sqrt{2}\hat{e}_3$  and, for example,  $a = 0$ ,  $b = \alpha\hat{e}_2 + \sqrt{2}\alpha\hat{e}_3$ . Clearly, HCP is not a Bravais lattice, since  $a + 2(b - a) = 2b$  does not belong to (1). HCP accounts for about 1/5 of the Periodic Table. Silicon and germanium (and carbon) adopt the diamond structure under ordinary conditions, which is also a 2-lattice. Many layered compounds such as the halogens, carbon (as graphite), oxygen and nitrogen are also 2-lattices, either as individual layers or as their accepted layered structures. Altogether, about 1/4 of the elements in the Periodic Table are 2-lattices. There are also examples that are not crystals at all under ordinary conditions, such as sulfur (a double ring) and boron (icosahedra, sometimes weakly bonded).

Even if they are not common, we also should mention the celebrated structures of nanotechnology: graphene, carbon nanotubes, the fullerenes, phosphorene, and the many other 2D materials now under study.

We will explore an alternative way of looking at the Periodic Table, and structure in general, via the concept of *objective structures* James [2006]. In fact the examples mentioned above have a common mathematical structure not based on Bravais lattices. An *objective atomic structure* (briefly, a 1-OS) has the defining property that each atom “sees the same environment”. Imagine Maxwell's demon, sitting on an atom, and looking at the environment (out to infinity). The demon hops to another atom, reorients itself in a certain way, and sees exactly the same environment. Mathematically, a set of points in  $\mathbb{R}^3$  is given,  $\mathcal{S} = \{x_1, x_2, \dots, x_N\} \in (R^3)^N$ ,  $N \leq \infty$  (most of the structures mentioned above are infinite).  $\mathcal{S}$  is a 1-OS if there are orthogonal transformations  $Q_1, \dots, Q_N$  such that

$$(2) \quad \{x_i + Q_i(x_j - x_1) : j = 1, \dots, N\} = \mathcal{S} \quad \text{for } i = 1, \dots, N.$$

Again, in words, the structure as viewed from atom 1,  $x_j - x_1$ , undergoes an orthogonal transformation  $Q_i(x_j - x_1)$  depending on  $i$ , is added back to atom  $i$ ,  $x_i + Q_i(x_j - x_1)$ , and the structure is restored. The surprising fact is that nearly all of the structures mentioned above, including the 2-lattices and those workhorse structures of nanotechnology, are examples of 1-OS.



It is indeed surprising. One would expect that identical environments would require some kind of isotropy of atomistic forces (for example, pair potentials, with force between a pair depending only on the distance). But some of the examples above are covalently bonded with complex electronic structure [Banerjee, Elliott, and James \[2015\]](#). Evidently, the property of identical environments is not coincidental. Unsurprisingly, many useful necessary conditions about equilibrium and stability follow from the definition [James \[2006\]](#). However, the basic reason why these structures are so common can be considered one of the fundamental open questions of atomic structure.

There is a glaring counter-example: atomic number 25, Manganese. In fact, this “most complex of all metallic elements” [Hobbs, Hafner, and Spišák \[2003\]](#) is (at room temperature and pressure) a 4-lattice. According to Hobbs et al. [Hobbs, Hafner, and Spišák \[ibid.\]](#), due to nearly degenerate spin configurations, the observed structure should be considered as containing four different magnetic atoms MnI, MnII, MnIII, MnIV. Briefly, Mn should be considered an alloy, rather than an element. There are a few other cases that could be considered equivocal: Is the structure of boron icosahedral (a 1-OS) or the weakly bonded lattice of icosahedra (not a 1-OS) that is sometimes given as its structure? But, overwhelmingly, the assertion made above about the prevalence of 1-OS on both the Periodic Table, and also for nanostructures made with one type of atom, is accurate.

Intuitively, one can easily imagine why such structures are interesting. If a property can be assigned to each atom, depending on its environment, it is *frame-indifferent* (independent of the  $Q_i$ ), and one can superpose it by summing over atoms, then an appreciable bulk property could result. This property, and fact that the patterns of bonding that in nanostructures differ appreciably those from bulk crystals, underlies significant research in nanoscience.

The idea of objective structures was articulated by [Crick and Watson \[1956\]](#), [Caspar and Klug \[1962\]](#) and in the less well-known work of [Crane \[1950\]](#). Caspar and Klug used the term *equivalence* to denote structures in which each subunit is “situated in the same environment”. The fundamental paper of [Dolbilin, Lagarias, and Senechal \[1998\]](#) proposed the concept of *regular point systems*, which adds to the idea of identical environments the hypotheses of uniform discreteness and relative denseness.

### 3 Objective structures

Structures containing only one element are interesting, but very special. There is a more general concept [James \[2006\]](#) applicable to the structures of many alloys and many molecular structures. Consider a structure consisting of  $N$  “molecules”, each consisting of  $M$  atoms. The terminology is for convenience – they may not be actual molecules. We say that a structure is an *objective molecular structure* (briefly, an M-OS) if one can

set up a one-to-one correspondence between atoms in each molecule such that equivalent atoms see that same environment. So, in this case we use a double-index notation  $\mathcal{S} = \{x_{i,k} \in \mathbb{R}^3 : i = 1, \dots, N, k = 1, \dots, M\}$ , where  $x_{i,k}$  is the position of atom  $k$  in molecule  $i$ . Here, with  $N \leq \infty$  and  $M < \infty$ .  $\mathcal{S}$  is an M-OS if  $x_{1,1}, \dots, x_{1,M}$  are distinct and there are  $NM$  orthogonal transformations  $Q_{i,k}$ ,  $i = 1, \dots, N$ ,  $k = 1, \dots, M$  such that

$$(3) \quad \{x_{i,k} + Q_{i,k}(x_{j,\ell} - x_{1,k}) : j = 1, \dots, N, \ell = 1, \dots, M\} = \mathcal{S}$$

for  $i = 1, \dots, N$ ,  $k = 1, \dots, M$ . Note that the reorientation  $Q_{i,k} \in O(3)$  is allowed to depend on both  $i$  and  $k$ . Briefly,  $x_{i,k}$  sees the same environment as  $x_{1,k}$ . This definition is the direct analog of *multiregular point systems* of [Dolbilin, Lagarias, and Senechal \[1998\]](#), but excluding the conditions of uniform discreteness and relative denseness. The author was led to it in a study with [Falk and James \[2006\]](#) of the helical tail sheath of Bacteriophage T-4, while writing a formula for the positions and orientations of its molecules consistent with measured electron density maps. An M-OS $_{M=1}$  is a 1-OS.

The definition of an M-OS can be written using a permutation  $\Pi$  on two indices  $(p, q) = \Pi(j, \ell)$ :

$$(4) \quad x_{i,k} + Q_{i,k}(x_{j,\ell} - x_{1,k}) = x_{\Pi(j,\ell)}.$$

It is not reflected by the notation here, but  $\Pi$  depends on the choice of  $(i, k)$ . We can also assign a species to each  $(j, \ell)$ . In most applications it would be required that atom  $(j, \ell)$  is the same species as atom  $(j', \ell)$ . Also, it would be required that  $\Pi$  preserve species, so that the environment of atom  $(i, k)$  matches the environment of atom  $(1, k)$  in both placement and species of atoms. The most interesting dimensions for the structure of matter are 3 and 2, but the definition is meaningful in any number of dimensions. Finally, in applications to atomic structure we are only interested in discrete M-OS. Of course, if one point of a 1-OS is an accumulation point, then every point is an accumulation point, since each point sees the same environment.

The assertions about 1-OS made in the preceding section are easily proved using the definitions above, but an even easier method is to note the following relation between objective structures and isometry groups. An isometry group is a group of elements of the form  $(Q|c)$ ,  $Q \in O(n)$ ,  $c \in \mathbb{R}^n$  based on the product  $(Q_1|c_1)(Q_2|c_2) = (Q_1Q_2|c_1 + Q_1c_2)$ , the identity  $(I|0)$ , and inverses  $(Q|c)^{-1} = (Q^T| -Q^Tc)$ . Isometries can act on  $\mathbb{R}^n$  in the obvious way:  $g(x) = Qx + c$  where  $g = (Q|c)$ . The product is designed to agree with composition of mappings:  $g_1g_2(x) = g_1(g_2(x))$ . As above, in view of the applications, we will put  $n = 3$ .

Let  $\mathcal{S} = \{x_{i,k} \in \mathbb{R}^3 : i = 1, \dots, N, k = 1, \dots, M\}$  be a discrete M-OS. Any such structure has an isometry group  $G$ :

$$(5) \quad G = \{(Q|c), Q \in O(3), c \in \mathbb{R}^3 : Qx_{i,k} + c = x_{\Pi(i,k)} \text{ for a permutation } \Pi\}.$$

Let  $\mathfrak{M}_1 = \{x_{1,k} : k = 1, \dots, M\}$  be “molecule 1”. We wish to show that  $\mathcal{S}$  is the orbit of molecule 1 under a discrete group of isometries. To see this, rearrange the definition of an M-OS as

$$(6) \quad R_{i,k}x_{j,\ell} + x_{i,k} - R_{i,k}x_{1,k} = x_{\Pi(j,\ell)}.$$

Hence,  $g_{i,k} := (R_{i,k} | x_{i,k} - R_{i,k}x_{1,k}) \in G$ . However, trivially,  $g_{i,k}(x_{1,k}) = x_{i,k}$ , and this holds for all  $i = 1, \dots, N$ ,  $k = 1, \dots, M$ . Hence,  $\mathcal{S}$  is contained in the orbit of  $\mathfrak{M}_1$  under  $G$ . Conversely, putting  $i = 1$  in (5) we have that the orbit of  $\mathfrak{M}_1$  under  $G$  is contained in  $\mathcal{S}$ .

This simple argument apparently has two flaws. First, the group  $G$  that one gets may not be discrete. That would be a serious flaw, since evidently we know very little about the nondiscrete groups of isometries, even in  $\mathbb{R}^3$ . (However, see remarks below for why these groups might be important to the structure of matter.) We should mention that discreteness is not merely a technical condition that rules out some special cases, but rather it plays a dominant role in the derivation of the groups, particularly in the subperiodic case appropriate to nanostructures. Second, in this argument there is nothing that prevents the images of  $\mathfrak{M}_1$  from overlapping. The latter is consistent with the definitions, and also advantageous from the physical viewpoint. That is, while we have imposed the condition that the points in  $\mathfrak{M}_1$  are distinct, the definition of M-OS allows  $x_{i,j} = x_{i',j}$  for  $i \neq i'$ . This is advantageous as it saves the result above. Also, it allows a structure such as ethane  $\text{C}_2\text{H}_6$  to be a 2-OS, which is certainly reasonable:  $\mathfrak{M}_1$  is C-H, each H sees the same environment, each C sees the same environment, and the image of C-H has overlapping Cs.

The geometric concept of identical environments allows  $Q_{i,k}$  to depend on both  $i, k$ . However, if  $\mathcal{S}$  is the orbit of  $x_{1,k}$ ,  $k = 1, \dots, M$ , under an isometry group  $g_1 = (Q_1|c_1), \dots, g_N = (Q_N|c_N)$ , i.e.,  $x_{i,k} = Q_i x_{1,k} + c_c$ , then  $Q_{i,k}$  in (3) can be chosen as  $Q_i$ , and thus is independent of  $k$ . This is seen by direct substitution of  $Q_{i,k} = Q_i$  into (3).

The nondiscreteness turns out not to be a problem. It is easily proved that if a nondiscrete group of isometries in 3-D generates a discrete structure when applied to a point  $x_1$ , it gives a single point, a 1-D Bravais lattice, or a 1-D 2-lattice.

Now we can revisit some of the assertions made in Section 2 concerning examples of 1-OS.

**Buckminsterfullerine ( $\text{C}_{60}$ ).** Let  $G = \{R_1, \dots, R_N\}$  be a finite subgroup of  $\text{O}(3)$  and  $x_1 \neq 0$ . (For  $\text{C}_{60}$  choose the icosahedral group,  $N = 60$ ) and let  $x_i = R_i x_1$ ,  $i = 1, \dots, N$ . One can also see directly that (2) is satisfied with  $Q_i = R_i$ . In case that  $x_1$  is fixed by some elements of  $R_1, \dots, R_N$ , then in this case one can replace  $G$  by  $G/G_{x_1}$  to obtain a free action (i.e., avoid duplication).

### Single-walled carbon nanotubes (of any chirality).

The formulas below can be found by rolling up a graphene sheet isometrically and seamlessly (see, e.g., [Dumitrica and James \[2007\]](#)) and then noticing the group structure. The positive integers  $(n, m)$  define the *chirality*. Letting  $\hat{e}_1, \hat{e}_2, \hat{e}_3$  be an orthonormal basis and  $R_\theta \in \text{SO}(3)$  a rotation with counterclockwise angle  $\theta$  and axis  $\hat{e}_3$ , carbon nanotubes are given by the formula

$$(7) \quad g_1^{v_1} g_2^{v_2} g_3^{v_3} (x_1), \quad v_1, v_2, v_3 \in \mathbb{Z},$$

with  $g_1 = (R_{\theta_1} | t_1)$ ,  $g_2 = (R_{\theta_2} | t_2)$  and  $g_3 = (-I + 2e \otimes e | 0)$ ,

$$(8) \quad t_1 = \tau_1 \hat{e}_3, \quad t_2 = \tau_2 \hat{e}_3, \quad e = \cos(\pi \xi) \hat{e}_1 + \sin(\pi \xi) \hat{e}_2, \quad x_1 = \rho \hat{e}_1 - \eta \hat{e}_3,$$

and

$$(9) \quad \begin{aligned} \theta_1 &= \frac{\pi(2n+m)}{n^2+m^2+nm}, \quad \theta_2 = \frac{\pi(2m+n)}{n^2+m^2+nm}, \quad \tau_1 = \frac{3m\ell_{C-C}}{2\sqrt{n^2+m^2+nm}}, \\ \tau_2 &= \frac{-3n\ell_{C-C}}{2\sqrt{n^2+m^2+nm}}, \quad \xi = \frac{(2n+m)+(2m+n)}{6(n^2+m^2+nm)}, \\ \rho &= \frac{\ell_{C-C}}{2\pi} \sqrt{3(n^2+m^2+nm)}, \quad \eta = \frac{\ell_{C-C}(m-n)}{4\sqrt{n^2+m^2+nm}}. \end{aligned}$$

The fixed integers  $n, m$  define the chirality of the nanotube and  $\ell_{C-C}$  is the carbon-carbon bond

length before rolling (To account for additional relaxation of the bond lengths after rolling one can simply omit the formula for the radius  $\rho$  and treat  $\rho$  as an independent parameter).

We see that  $g_1 g_2 = g_2 g_1$  and  $g_i g_3 = g_3 g_i^{-1}$ ,  $i = 1, 2$ , so  $g_1^{v_1} g_2^{v_2} g_3^{v_3}$ ,  $v_1, v_2, v_3 \in \mathbb{Z}$  is a (discrete) group. Therefore the orbit (7) describes a 1-OS. To obtain a free action, confine  $v_3 \in \{1, 2\}$ ,  $v_1 \in \mathbb{Z}$  and  $v_2 \in \{1, \dots, v_\star\}$ , where  $v_\star$  is the smallest positive integer such that  $g_1^v g_2^{v_\star} = id$  is solvable for  $v \in \mathbb{Z}$ .

**Any 2-lattice.** Of course any Bravais lattice is a 1-OS: use a suitable indexing in terms of triples of integers  $v = (v^1, v^2, v^3)$ , write  $x_v = v^1 e_1 + v^2 e_2 + v^3 e_3$  and choose  $Q_v = I$  in (2). As noted above and represented prominently in the Periodic Table, any 2-lattice is also a 1-OS. To see this, choose  $g_1 = (I | e_1)$ ,  $g_2 = (I | e_2)$ ,  $g_3 = (I | e_3)$  and  $g_4 = (-I | 0)$ . Then, for  $s = 1, 2$ ,

$$(10) \quad g_1^{v_1} g_2^{v_2} g_3^{v_3} g_4^s (x_1) = v^1 e_1 + v^2 e_2 + v^3 e_3 \pm x_1.$$

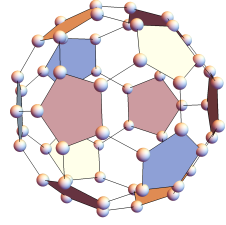


Figure 1: Buckminsterfullerene. Pentagons added for clarity.

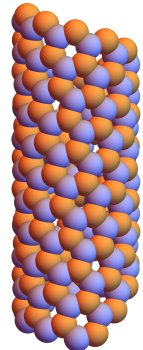


Figure 2: Carbon nanotube (a 1-OS) with chirality  $n = 3, m = 8$ ,

Referring to (1) we can choose  $x_1 = (b - a)/2$  and modify these isometries to translate the whole structure<sup>1</sup> by  $(a + b)/2$  to get exactly (1). Of course,  $\{g_1^{v_1} g_2^{v_2} g_3^{v_3} g_4^s : v_1, v_2, v_3 \in \mathbb{Z}^3, s = 1, 2\}$  is a (discrete) group, because  $g_1, g_2, g_3$  commute and  $g_i g_4 = g_4 g_i^{-1}$ ,  $i = 1, 2, 3$ .

This kind of argument works in any number of dimensions and therefore covers two-dimensional 2-lattices, such as graphene.

**HCP.** The hexagonal close packed lattice is a 2-lattice, as proved above, and therefore is a 1-OS by the result just above. However, it is useful to express it by a different group to expose an important issue. Beginning from the description HCP above, i.e.,

$$(11) \quad e_1 = \sqrt{3}\alpha\hat{e}_1, \quad e_2 = \sqrt{3}\alpha\left((1/2)\hat{e}_1 + (\sqrt{3}/2)\hat{e}_2\right), \quad e_3 = 2\alpha\sqrt{2}\hat{e}_3,$$

and with  $R_\theta \in \text{SO}(3)$  a counterclockwise rotation of  $\theta$  about  $\hat{e}_3$ , define

$$(12) \quad h = (R_{\pi/3} | (1/2)e_3), \quad t_1 = (I | e_1) \quad t_2 = (I | e_2).$$

The set  $\{h^i g_1^j g_2^k : i, j, k \in \mathbb{Z}\}$  is a group ( $t_1 t_2 = t_2 t_1$ ,  $t_2 h = h t_1$ ,  $t_1 h = h t_1 t_2^{-1}$ ), and the orbit of  $(2/3)e_2 - (1/3)e_1$  is HCP. This illustrates that we can have two groups, this one and the one of the preceding paragraph, not related by an affine transformation  $G \rightarrow aGa^{-1}$ ,  $a = (A|c)$ ,  $\det A \neq 0$ , that generate the same structure when the orbits of suitable points are taken.

For the purpose of this article we do not care about multiplication tables (we need the actual isometries with their parameter dependence), affine equivalence (example of HCP) or whether the closure of the fundamental domain is compact (not true for nanostructures). Embarrassingly, we do not even care much about symmetry. If we have a set of generators, depending smoothly on parameters, and the symmetry suddenly jumps up at values of the parameters – such as at  $\gamma = 1, \sqrt{2}$  in the example of Bain above – it makes no difference for any of the results given below. On the other hand, the analytical structure of the generators is critically important. For the purposes here it would be very useful to have a short lists of formulas for generators giving all objective structures, not further broken down according to their abstract groups.

## 4 An invariant manifold of molecular dynamics

Structure and invariance come together when we assign a set of differential equations having an invariance group that matches the group used to make the structure. Isometries

<sup>1</sup>Change each  $g_i$  to  $t g_i t^{-1}$  where  $t = (I|c)$ ,  $c = (a + b)/2$ .

are expected to play an important role because of the frame-indifference of atomic forces. Examples of exploiting symmetries in continuum theory [Ericksen \[1977\]](#) might suggest that a differential structure (i.e, a Lie group) is essential, but, in fact, the matching of discrete group and discrete structure is also possible.

Molecular dynamics is an interesting example. The basic invariance of the equations of molecular dynamics is frame-indifference and permutation invariance. Let us use the indexing of M-OS to describe these invariances, but without any assumptions about the structure. So we assume a collection of atomic positions  $\mathcal{S} = \{x_{i,k} \in \mathbb{R}^3 : i = 1, \dots, N, k = 1, \dots, M\}$  with  $N \leq \infty$  and  $M < \infty$ , and suppose that the force on atom  $(i, k)$  is given by

$$(13) \quad f_{i,k}(\dots, x_{j,1}, x_{j,2}, \dots, x_{j,M}, x_{j+1,1}, x_{j+1,2}, \dots, x_{j+1,M}, \dots).$$

As indicated, the force on atom  $(i, k)$  depends on the positions of all the atoms. We have  $NM$  such forces. They are subject to two fundamental invariances: frame-indifference and permutation invariance.

*Frame-indifference.* For  $Q \in O(3)$ ,  $c \in \mathbb{R}^3$ ,

$$(14) \quad \begin{aligned} \bar{f}_{i,k}(\dots, Qx_{j,1} + c, \dots, Qx_{j,M} + c, Qx_{j+1,1} + c, \dots, Qx_{j+1,M} + c, \dots) \\ = Qf_{i,k}(\dots, x_{j,1}, \dots, x_{j,M}, x_{j+1,1}, \dots, x_{j+1,M}, \dots), \end{aligned}$$

for all  $Q \in O(3)$ ,  $c \in \mathbb{R}^3$  and  $x_{j,\ell} \in (R^3)^{NM}$ .

*Permutation invariance.* For all permutations  $\Pi$  and  $x_{j,\ell} \in (R^3)^{NM}$ ,

$$(15) \quad \begin{aligned} \bar{f}_{i,k}(\dots, x_{\Pi(j,1)}, \dots, x_{\Pi(j,M)}, x_{\Pi(j+1,1)}, \dots, x_{\Pi(j+1,M)}, \dots) \\ = f_{\Pi(i,k)}(\dots, x_{j,1}, \dots, x_{j,M}, x_{j+1,1}, \dots, x_{j+1,M}, \dots), \end{aligned}$$

If we introduce species as described above, then  $\Pi$  is also required to preserve species: if atom  $i, k$  has species  $\mathcal{Q}$  and  $(p, q) = \Pi(i, k)$  then the species of atom  $p, q$  is  $\mathcal{Q}$ .

Typically,  $NM = \infty$ , in which case one cannot speak of a potential energy, but, in the finite case, if  $f_{i,k} = -\partial\varphi/\partial x_{i,k}$  then the conditions (14), (15) follow from the familiar invariances  $\varphi(\dots, Qx_{j,\ell} + c, \dots) = \varphi(\dots, x_{j,\ell}, \dots)$  and  $\varphi(\dots, x_{\Pi(j,\ell)}, \dots) = \varphi(\dots, x_{j,\ell}, \dots)$ , respectively.

As one can see from the examples (7)-(12), the formulas for objective structures contain lots of parameters. Eventually, we are going to solve the equations of molecular dynamics for functions depending on time,  $t > 0$ . To use the invariance as completely as possible without unduly restricting the number of atoms or their motions, we could allow these group parameters to depend on time. In general suppose we have an isometry group  $g_1 =$

$(Q_1(t)|c_1(t)), \dots, g_N = (Q_N(t)|c_N(t))$  smoothly depending on  $t$ . The property we will need is

$$(16) \quad \frac{d^2}{dt^2} g_i(y(t)) = \frac{d^2}{dt^2} (Q_i(t)y(t) + c_i(t)) = Q_i(t) \frac{d^2 y(t)}{dt^2},$$

but we will know very little *a priori* about  $y(t)$ ,  $t > 0$ , beyond some smoothness. Letting a superimposed dot indicate the time derivative and using  $\dot{Q}_i = Q_i W_i$ , where  $W_i^T = -W_i$ , the condition (16) is

$$(17) \quad \ddot{c}_i = -Q_i(W_i^2 y + \dot{W}_i y + 2W_i \dot{y}).$$

The only way this holds for any reasonable class<sup>2</sup> of smooth motions  $y(t)$  is

$$(18) \quad Q_i = \text{const.} \in O(3) \quad \text{and} \quad c_i = a_i t + b_i,$$

where  $a_i, b_i \in \mathbb{R}^3$ ,  $i = 1, \dots, N$ . While the latter may appear to be merely a Galilean transformation, the dependence on  $i$  gives many nontrivial cases. Of course, one has to check that the (18) is consistent with the group properties at each  $t > 0$ .

We say that the group  $G = \{g_1, \dots, g_N\}$ ,  $g_i = (Q_i | a_i t + b_i)$ , is a *time-dependent discrete group of isometries* if (18) is satisfied, and we use the notation  $g_i(y, t) = Q_i y + a_i t + b_i$ . We also assign a mass  $m_k > 0$  to each atom  $(1, 1), \dots, (1, M)$ , and we assume that atom  $i, k$  also has mass  $m_k$ , consistent with the remarks about species above. Now, instead of thinking of atoms  $(1, 1), \dots, (1, M)$  as molecule 1, we think in terms of a numerical method, and call atoms  $(1, 1), \dots, (1, M)$  the *simulated atoms*. In many cases they will not behave at all like a molecule. We will also have *nonsimulated atoms* and their positions will be given (as in an M-OS) by the group:

$$(19) \quad y_{i,k}(t) = g_i(y_{1,k}(t), t), \quad i = 1, \dots, N, \quad k = 1, \dots, M$$

Obviously we have assigned  $g_1 = id$ .

Let  $G = \{g_1, \dots, g_N\}$  be a time-dependent discrete group of isometries. Let initial positions  $y_k^\circ \in \mathbb{R}^3$  and initial velocities  $v_k^\circ$ ,  $k = 1, \dots, M$ , be given and suppose the simulated atoms  $y_{1,k}(t)$ ,  $t > 0$ , satisfy the equations of molecular dynamics for forces that are frame-indifferent and permutation invariant:

$$(20) \quad \begin{aligned} m_k \ddot{y}_{1,k} &= f_{1,k}(\dots, y_{j,1}, y_{j,2}, \dots, y_{j,M}, y_{j+1,1}, y_{j+1,2}, \dots, y_{j+1,M}, \dots) \\ &= f_{1,k}(\dots, g_j(y_{1,1}, t), \dots, g_j(y_{1,M}, t), g_{j+1}(y_{1,1}, t), \dots, g_{j+1}(y_{1,M}, t), \dots), \end{aligned}$$

subject to the initial conditions

$$(21) \quad y_{1,k}(0) = y_k^\circ, \quad \dot{y}_{1,k}(0) = v_k^\circ, \quad k = 1, \dots, M.$$

<sup>2</sup>It is not sufficient for our purposes to satisfy (17) in a statistical sense.

Then, the nonsimulated atoms also satisfy the equations of molecular dynamics:

$$(22) \quad m_k \ddot{y}_{i,k} = f_{1,k}(\dots, y_{j,1}, y_{j,2}, \dots, y_{j,M}, y_{j+1,1}, y_{j+1,2}, \dots, y_{j+1,M}, \dots).$$

Note that (20), (21) is a (nonautonomous) system of ODEs in standard form for the simulated atoms. We have not stated this as a theorem because we have not spelled out the (straightforward) conditions on  $f_{1,k}$  that would allow us to invoke one of the standard existence theorems of ODE theory. Another (also straightforward to handle) technical issue is that a standard atomic forces blow up repulsively when two atoms approach each other. Aside from these issues, the proof is a simple calculation that uses both frame-indifference and permutation invariance. To see this, fix  $i, k$  and suppose we want to prove that  $y_{i,k}(t)$  satisfies (22) as written. Write  $g_i = (Q_i | c_i)$ ,  $c_i = a_i t + b_i$ , so that  $g_i^{-1} = (Q_i^T | -Q_i^T c_i)$ . There is a permutation  $\Pi$  (depending on  $i$ ) such that  $y_{\Pi(j,\ell)}(t) = g_i^{-1}(y_{j,\ell}, t)$ . This permutation satisfies  $\Pi(i, k) = (1, k)$ . Now use (16), permutation invariance, and frame-indifference (in that order):

$$\begin{aligned} m_k \ddot{y}_{i,k} &= m_k Q_i \ddot{y}_{1,k} \\ &= Q_i f_{1,k}(\dots, y_{j,1}, \dots, y_{j,m}, y_{j+1,1}, \dots, y_{j+1,M}, \dots) \\ &= Q_i f_{\Pi(i,k)}(\dots, y_{j,1}, \dots, y_{j,m}, y_{j+1,1}, \dots, y_{j+1,M}, \dots) \\ &= Q_i f_{i,k}(\dots, y_{\Pi(j,1)}, \dots, y_{\Pi(j,m)}, y_{\Pi(j+1,1)}, \dots, y_{\Pi(j+1,M)}, \dots) \\ &= Q_i f_{i,k}(\dots, g_i^{-1}(y_{j,1}), \dots, g_i^{-1}(y_{j,m}), g_i^{-1}(y_{j+1,1}), \dots, g_i^{-1}(y_{j+1,M}), \dots) \\ &= Q_i f_{i,k}(\dots, Q_i^T(y_{j,1} - c_i), \dots, Q_i^T(y_{j,m} - c_i), Q_i^T(y_{j+1,1} - c_i), \\ &\quad \dots, Q_i^T(y_{j+1,M} - c_i), \dots) \\ (23) \quad &= f_{i,k}(\dots, y_{j,1}, \dots, y_{j,m}, y_{j+1,1}, \dots, y_{j+1,M}, \dots). \end{aligned}$$

This result can be rephrased as the existence of a (time-dependent) invariant manifold of molecular dynamics. Given the many isometry groups and their time dependences, this provides a multitude of mainly unstudied invariant manifolds of the equations of molecular dynamics. Their stability of course is also unknown.

We can describe these invariant manifolds in terms of the isometry groups. The conventional description is in phase space, using momenta  $p_{i,k} = m_k \dot{y}_{i,k}$  and positions  $q_{i,k} = y_{i,k}$ . Using our notation for time-dependent discrete isometry groups,  $G = \{g_1, \dots, g_N\}$ ,  $g_i = (Q_i | a_i t + b_i)$ , we observe that

$$(24) \quad p_{i,k} = Q_i p_{1,k} + m_k a_i, \quad q_{i,k} = Q_i q_{1,k} + a_i t + b_i,$$

which describes an affine manifold in phase space with a simple affine time dependence. Except for the trivial  $m_k$  dependence in the first of (24), this family of invariant manifolds



is independent of the material. That is, this large set of invariant manifolds is present whether one is simulating steel, water or air. (Of course, there is a large number, typically  $3NM = \infty$ , of dimensions too!). More importantly, this becomes a powerful simulation tool if the atomic forces have a cut-off. One can simulate certain large-scale flows, or the failure of nanostructures, by a small simulation [Dayal and James \[2010, 2012\]](#). One solves (20)–(21) merely for the simulated atoms, with forces given by all the atoms. Of course, this requires a good method for tracking some of the nonsimulated atoms, i.e., those within the cut-off. As a numerical method, this is called *objective molecular dynamics* [Dayal and James \[2010\]](#).

Let us take a simple example, the translation group. Using a convenient indexing in terms of triples of integers  $v = (v^1, v^2, v^3)$ , we write  $g_v = (I | a_v t + b_v)$ . We see that to satisfy closure with this time-dependence, we must have

$$(25) \quad a_v t + b_v = v^1(\hat{b}_1 + \hat{a}_1 t) + v^2(\hat{b}_1 + \hat{a}_1 t) + v^3(\hat{b}_1 + \hat{a}_1 t) = \sum_{\ell=1}^3 v^\ell (I + tA)e_\ell,$$

where  $e_\ell = \hat{b}_\ell$  and the  $3 \times 3$  matrix  $A$  is chosen so that  $\hat{a}_\ell = Ae_\ell$ . Tacitly, we have assumed that the  $e_\ell = \hat{b}_\ell$  are linearly independent, so that, initially, the atoms are not confined to a layer. The simulated atom positions are say  $y_1(t), \dots, y_M(t)$  and the nonsimulated atoms,  $y_{v,k}(t) = g_v(y_k(t)) = y_k(t) + v^\ell(I + tA)e_\ell$ ,  $v = (v^1, v^2, v^3) \in \mathbb{Z}^3, k = 1, \dots, M$ .

In this method atoms are moving around, filling space roughly uniformly. During computations, the simulated atoms quickly diffuse into the nonsimulated atoms. What is the macroscopic motion? We could spatially average the velocity, but that would be wrong: the velocity of continuum mechanics is not the average velocity of the particles! (For a simple explanation see [James \[2015\]](#).) To get the velocity of continuum mechanics we should average the momentum, and divide by the average density. Briefly, a suitable method in the present case is to prove that the center of mass of the simulated atoms moves with constant velocity which, by adding an overall Galilean transformation, we take to be zero<sup>3</sup>. Then the centers of mass of the images of the simulated atoms  $g_v(y_k(t))$  then lie on a grid deforming according to the motion<sup>4</sup>.

$$(26) \quad y(x, t) = (I + tA)x, \quad \text{or, in Eulerian form, } v(y, t) = A(I + tA)^{-1}y$$

Here,  $v(y, t)$  is the velocity field. Note that by looking at the motions of centers of mass, we are precisely doing a spatial average of the momentum and then dividing by the average

<sup>3</sup>This requires an additional assumption on the forces  $f_{i,k}$  that the resultant force on large volume, divided by the volume, tends to zero as the volume (at constant shape) goes to infinity [Dayal and James \[2012\]](#). This effectively rules out body forces, such as those due to gravity. It is easily proved directly for many accepted models of atomic forces.

<sup>4</sup>The Eulerian and Lagrangian forms are related by the parameterized ODE,  $\partial y / \partial t = v(y(x, t), t)$ ,  $y(0) = x \in \Omega \subset \mathbb{R}^3$ .

density. Given that we get to choose  $A$ , we get quite a few interesting motions. They can be very far-from-equilibrium, have nonzero vorticity (in fact, vortex stretching), and there are quite a few of both isochoric and non-isochoric examples.

Another interesting example is based on the (largest) Abelian helical group. In [Dayal and James \[2010\]](#) it was used to study the failure of carbon nanotubes when stretched at constant strain rate.

## 5 Continuum and structural mechanics

Let's take the translation group, leading to (26). We have a macroscopic velocity field  $v(y, t) = A(I + tA)^{-1}y$  arising from molecular dynamics simulation. For what choices of  $A$ , if any, does  $v(y, t)$  satisfy some accepted equations of continuum mechanics? We can try the Navier-Stokes equations in the incompressible case. First we check that there are choices of  $A$  such that  $\operatorname{div} v = 0$ . It is easily<sup>5</sup> seen that  $\operatorname{div} v = 0$  for  $t > 0$  if and only if  $\det A = \operatorname{tr} A = \operatorname{tr} A^2 = 0$ , which, in turn implies that there is an orthonormal basis in which  $A$  has the form

$$(27) \quad A = \begin{pmatrix} 0 & 0 & \kappa \\ \gamma_1 & 0 & \gamma_3 \\ 0 & 0 & 0 \end{pmatrix}.$$

(So, even for isochoric motions we can have a time-dependent vorticity,  $\operatorname{curl} v = (\gamma_3 - \kappa\gamma_1 t, -\kappa, \gamma_1)$  in this basis.) Now substitute  $v(y, t) = A(I + tA)^{-1}y$  into the Navier-Stokes equations

$$(28) \quad \rho \left( \frac{\partial v}{\partial t} + \nabla v v \right) = -\nabla p + \Delta v$$

i.e.,  $\rho(-A(I + tA)^{-1}A(I + tA)^{-1}y + A(I + tA)^{-1}A(I + tA)^{-1}y) = -\nabla p + 0$ , so, with  $p = \text{const.}$ , the Navier-Stokes equations are identically satisfied.

The key properties being exploited in this case is that the left hand side of the balance of linear momentum is identically zero, and, for the right hand side, the stress is only a function of time when evaluated for the velocity field  $v(y, t) = A(I + tA)^{-1}y$ . So its divergence is zero. In fact,  $v(y, t) = A(I + tA)^{-1}y$  identically satisfies the equations of all accepted models of fluid mechanics, including exotic models of non-Newtonian fluids and liquid crystals. The same is true of all accepted models of solid mechanics<sup>6</sup>. It is fascinating to observe that, despite the fact that molecular dynamics is time-reversible<sup>7</sup> and

<sup>5</sup>Even easier, use the equivalent  $\det(I + tA) = 1$  and write out the characteristic equation.

<sup>6</sup>It was advocated in [Dayal and James \[2010\]](#) that a fundamental requirement on models should be that  $v(y, t) = A(I + tA)^{-1}y$  is a solution, i.e., that all continuum models inherit the invariant manifold.

<sup>7</sup>The function  $y_{i,k}(-t)$  is a solution of (20) for initial conditions  $y_{i,k}^0, -v_{i,k}^0$ .

much of continuum mechanics is not time-reversible, this invariant manifold is inherited, in this sense, exactly.

Perhaps the most important feature of this family of solutions is that its form does not depend on the material. This is expected: as already noted, the invariant manifold of molecular dynamics (24) is also independent of the species of atoms being simulated. This feature strikes to the heart of experimental science, especially experimental mechanics. If you want to learn about a material by testing it, you should impose boundary conditions that are at least consistent with *a possible solution of the governing equations*. But one does not know the coefficients of the governing equations ahead-of-time<sup>8</sup>, because one has not yet measured the material properties! This fundamental dichotomy of experimental science is overcome by solutions of the type discussed here. Design the testing machine to produce boundary conditions consistent with a possible solution, and learn about the material by measuring the forces.

In fact, if one looks at the Couette viscometer in fluid mechanics or tension-torsion machine in solid mechanics, they do in fact, have a relation to these groups and their invariant manifolds. On the other hand these ideas could be more widely exploited in experimental science (e.g., Dayal and James [2012]).

These were all purely mechanical cases. What happens when one adds thermodynamics? Let's return to the invariant manifold (24). In atomistic theory temperature is usually interpreted as mean kinetic energy based on the velocity obtained, importantly, after subtracting off the mean velocity. The temperature is then the mean kinetic energy of the simulated atoms, assuming, as we have done above, that the center of mass of the simulated atoms moves with zero velocity. But, unlike the velocity, there is nothing about a simulation based on (20) that would determine this temperature, beyond the expectation that it depends on  $A$  and the initial conditions (21) and, of course, the atomic forces. In simulations it can be rapidly changing, and, in fact, it is expected in some situations to go to infinity in finite time<sup>9</sup>. In summary, the temperature  $\theta(t)$  is expected to be a function of time only, and not universal. This agrees with continuum theory, for which in most cases the energy equation reduces to an ODE for the temperature, when  $v(y, t) = A(I + tA)^{-1}y$ . In short, temperature is a function of time and its evolution is material dependent.

Experimental design is one of many subareas in continuum mechanics in which objective structures play an interesting role. Another is the blossoming area of structural mechanics called “origami structures”<sup>10</sup>. Fundamentally, frame-indifference is again being used: isometries take stress-free states to stress-free states.

<sup>8</sup>In cases on the cutting edge, one does not even know the form of the equations.

<sup>9</sup>See Section 6. Note that  $(I + tA)$  can lose invertibility in finite time.

<sup>10</sup>Already, the link between architecture and molecular structure was articulated by Caspar and Klug [1962]. See also Coxeter [1971]

Kawasaki's theorem in piecewise rigid origami concerns the  $2n$ -fold intersection. For example, in the ubiquitous case  $2n = 4$ , draw four lines on a piece of paper and fold along the lines<sup>11</sup>. This structure can be folded flat if and only if the sum of opposite angles is  $\pi$ . Without loss of generality  $\hat{e}_1, \hat{e}_2, \hat{e}_3$  are orthonormal, the paper is the  $\hat{e}_1, \hat{e}_2$ -plane and consider fold-lines coming out of the origin in directions

$$(29) \quad \begin{aligned} t_1 &= \hat{e}_1, & t_2 &= \cos \alpha \hat{e}_1 + \sin \alpha \hat{e}_2, & t_3 &= \cos(\alpha + \beta) \hat{e}_1 + \sin(\alpha + \beta) \hat{e}_2, \\ t_4 &= \cos(\pi + \beta) \hat{e}_1 + \sin(\pi + \beta) \hat{e}_2, \end{aligned}$$

corresponding to successive (sectors : angles)  $(\mathcal{S}_1 : \alpha), (\mathcal{S}_2 : \beta), (\mathcal{S}_3 : \pi - \alpha), (\mathcal{S}_4 : \pi - \beta)$  with  $0 < \alpha, \beta < \pi$ . Letting  $t_i^\perp = Q_3 t_i, i = 1, \dots, 4$ , where  $Q_3$  is a counter-clockwise rotation of  $\pi/2$  with axis  $\hat{e}_3$ , it is easy to write down the folding deformation  $y : \Omega \rightarrow \mathbb{R}^3$ , with  $0 \in \Omega \subset \mathbb{R}^2$ :

$$(30) \quad y(x) = \begin{cases} x, & x \cdot \hat{e}_3 = 0, x \cdot t_2^\perp < 0, x \cdot t_1^\perp \geq 0, \\ R_2(\eta)x, & x \cdot \hat{e}_3 = 0, x \cdot t_3^\perp < 0, x \cdot t_2^\perp \geq 0, \\ R_2(\eta)R_3(\xi)x, & x \cdot \hat{e}_3 = 0, x \cdot t_4^\perp < 0, x \cdot t_3^\perp \geq 0, \\ R_2(\eta)R_3(\xi)R_4(\omega)x, & x \cdot \hat{e}_3 = 0, x \cdot t_1^\perp < 0, x \cdot t_4^\perp \geq 0, \end{cases}$$

where  $\eta = \pm\omega$  and

$$(31) \quad \tan \xi = \begin{cases} \frac{(\cos \alpha - \cos \beta) \sin \omega}{\cos \omega - \cos \alpha \cos \beta \cos \omega + \sin \alpha \sin \beta}, & \eta = \omega, \\ \frac{(\cos \alpha + \cos \beta) \sin \omega}{\cos \omega + \cos \alpha \cos \beta \cos \omega - \sin \alpha \sin \beta}, & \eta = -\omega. \end{cases}$$

Here  $R_i(\theta) \in \text{SO}(3)$  has axis  $t_i$  and counter-clockwise angle  $\theta$ , and  $0 \leq \omega < \pi$  can be considered the homotopy parameter. We have fixed the overall rotation by putting  $y(x) = x$  in  $\mathcal{S}_1$ <sup>12</sup>

Now we make a special choice of  $\Omega$ : we choose it to be a general parallelogram, so that the fold-lines go from the origin to the corners. We have some freedom to assign angles and side lengths, as well as on the placement of the origin, but these restrictions can be easily organized. Now partially fold it, i.e., choose  $\pm$  and a value of  $0 < \omega < \pi$  in (30), (31). In the partly folded state let  $\ell_1, \ell_2, \ell_3, \ell_4$  be consecutive edges on the boundary of

<sup>11</sup>Or, simply crush a piece of paper and push down onto the table so it is flat. Upon unfolding, you will see numerous four-fold intersections with the sum of opposite angles equal to  $\pi$ . Even better, check many of the delightful folding arrangements discovered by Robert J. Lang and others [Miura, Kawasaki, Tachi, Uehara, Lang, and Wang-Iverson \[2015\]](#).

<sup>12</sup>In fact, this pair of homotopies, parameterized by  $0 \leq \omega < \pi$  and  $\pm$ , are the only piecewise rigid deformations of  $\Omega$  (with these fold lines) if  $y(x) = x$  in  $\mathcal{S}_1$  and  $\alpha \neq \beta, \alpha + \beta \neq \pi$ . If the latter holds there are some additional ones. The foldability of general arrays of 4-fold intersections, and a corresponding algorithm for folding them in terms of formulas like (30), is given in [Plucinsky, Feng, and James \[2017\]](#).

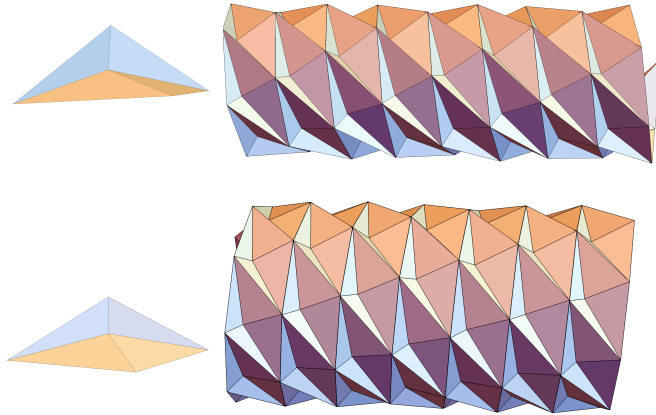


Figure 3: Helical origami structures generated by two commuting isometries whose powers give a discrete group. Bottom: the parallelogram is partly folded, as seen at left. Top: the same parallelogram is folded a little more. These solutions are isolated: for intermediate values of the homotopy parameter  $\omega$  the associated group is not discrete.

the deformed parallelogram, so that  $|\ell_1| = |\ell_3|$  and  $|\ell_2| = |\ell_4|$ . Choose two isometries  $g_1 = (Q_1|c_1)$ , and  $g_2 = (Q_2|c_2)$  out of the air, arrange that they commute, and arrange that  $g_1(\ell_1) = \ell_3$  and  $g_2(\ell_2) = \ell_4$ . Of course, the latter is possible because  $g_1, g_2$  are isometries, and there is obviously some freedom. This freedom is quantifiable without much difficulty. The underlying Abelian group is  $\{g_1^i g_2^j : i, j \in \mathbb{Z}\}$ .

Now we are done. The beauty of Abelian groups is that, not only does  $g_1^i(y(\Omega))$ ,  $i = 1, 2, \dots$  produce a perfectly fitting helical origami chain, and  $g_2^j(y(\Omega))$ ,  $j = 1, 2, \dots$  another such chain, but also  $g_1^i g_2^j(y(\Omega))$ ,  $i, j = 1, 2, \dots$  fills in the space between the chains perfectly with no gaps. See Figure 3.

However, Figure 3 is not the generic case. More typically, as  $i, j$  get large, the structure gets more and more complicated and begins to intersect itself. Of course, we knew that could happen because nothing above prevents self-intersections. But it is worse than that: there are accumulation points. The issue is: if we choose two commuting isometries “out of the air”, invariably we will get a non-discrete group. Discreteness is a highly restrictive condition for isometry groups, and is the main force behind the structure of the crystallographic groups of the International Tables. It is nevertheless worth illustrating the appearance of the structure one gets. This is done in Figure 4 with balls instead of origami, for clarity. If one cuts off the powers  $i, j$  early enough, one gets a perfectly nice

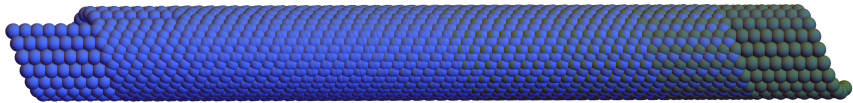


Figure 4: Orbit of a single blue ball under a subset of a nondiscrete Abelian group with two generators. The coloring is according to powers of one of these generators. Note that subsets of this structure coincide with structures that are locally 1-OS, 2-OS, etc.

structure that in fact, at least with two generators, exhibits locally identical environments for most of the atoms. For that reason, these non-discrete groups – or more accurately, their generators – about which evidently we know nothing, are in fact quite interesting.

Moreover, orbits of non-discrete groups (with restricted powers) are seen in biology. Perhaps the most obvious example is the biologically important *microtubule*. Its structure is closely given by the construction above with two generators and carefully restricted powers<sup>13</sup>. Another interesting example is from the work of Reidun Twarock and collaborators Keef, Wardman, Ranson, Stockley, and Twarock [2013] and Indelicato, Keef, Cermelli, Salthouse, Twarock, and Zanzotto [2012]. To understand the placement of receptors on the surface of a viral capsid, she takes the orbit of a certain non-discrete group with carefully restricted powers<sup>14</sup>. What are the nondiscrete isometry groups, and how do we restrict the powers of their generators in a rational way, perhaps guided by the concept of identical local environments?

In fact, one can satisfy all the matching conditions stated above using two commuting isometries that do generate a discrete group, and there are many choices. Figure 3 is an example. General theorems about these structures can be found in Feng, Plucinsky, and James [2017]. Beautiful origami structures that approximate an arbitrary Lipschitz map that shortens distances are given by Conti and Maggi [2008].

Before moving on, it is worth highlighting the fundamental problem of self-intersection in origami structures, since it so often prevents foldability and there are no good methods to decide this ahead-of-time<sup>15</sup>. For mappings  $y : \Omega \rightarrow \mathbb{R}^m$ ,  $\Omega \in \mathbb{R}^n$ , with  $n = m$  a lot is known that relates invertibility to invertibility on the boundary or to bounded measures of distortion Ball [1981], Ciarlet and Nečas [1987], and Iwaniec, Kovalev, and Onninen [2011]. The concept of global radius of curvature Carlen, Laurie, Maddocks, and Smutny

<sup>13</sup>Its seam can be considered a consequence of non-discreteness and carefully restricted powers.

<sup>14</sup>To see that the generated group is not discrete in their simplest example (Figure 2 of Keef, Wardman, Ranson, Stockley, and Twarock [2013]), let  $g_1 = (I | t)$ ,  $g_2 = (R | 0)$  be the generators considered, where  $R \in \text{SO}(2)$  is a rotation of  $\pi/5$  and  $0 \neq t \in \mathbb{R}^2$ . Then  $g_3 := g_2 g_1 g_2^{-2} g_1 g_2 = (I | (2 \cos(\pi/5)) t)$ . Thus  $g_1$  and  $g_3$  generate a nondiscrete subgroup because  $2 \cos(\pi/5) = (1/2)(1 + \sqrt{5})$  is irrational.

<sup>15</sup>Writing deformations in the form (30) – the continuum mechanics approach to origami structures – is a reasonable step 1.

[2005] has also been used for this purpose in knotted rods ( $n = 1, m = 3$ ). Both of these approaches seem relevant, but neither seems ideally suited.

## 6 Boltzmann equation

We return to the family of invariant manifolds of the equations of molecular dynamics, which was seen to be inherited in a perfect way by continuum mechanics. We now consider statistical theories intermediate between molecular dynamics and continuum mechanics. Of greatest interest, in view of its remarkable predictive power in the far-from-equilibrium case, is the Boltzmann equation.

The Boltzmann equation [Maxwell \[1867\]](#) and [Villani and Mouhot \[2015\]](#) is an evolution law for the molecular density function  $f(t, y, v)$ ,  $t > 0$ ,  $y \in \mathbb{R}^3$ ,  $v \in \mathbb{R}^3$ , the probability density of finding an atom at time  $t$ , in small neighborhood of position  $y$ , with velocity  $v$ . It satisfies the Boltzmann equation:

$$(32) \quad \frac{\partial f}{\partial t} + v \frac{\partial f}{\partial x} = \mathbb{C} f(v) := \int_{\mathbb{R}^3} \int_{S^2} B(n \cdot \omega, |v - v_*|) [f' f'_* - f_* f] d\omega dv_*,$$

where  $S^2$  is the unit sphere in  $\mathbb{R}^3$ ,  $n = n(v, v_*) = \frac{(v - v_*)}{|v - v_*|}$ ,  $(v, v_*)$  is a pair of velocities associated to the incoming collision of molecules and  $(v', v'_*)$  are outgoing velocities defined by collision rule

$$(33) \quad v' = v + ((v_* - v) \cdot \omega) \omega,$$

$$(34) \quad v'_* = v_* - ((v_* - v) \cdot \omega) \omega.$$

The form of the collision kernel  $B(n \cdot \omega, |v - v_*|)$  is obtained from the solution of the two-body problem of dynamics with the given force law between molecules. We use the conventional notation in kinetic theory,  $f = f(t, x, v)$ ,  $f_* = f(t, x, v_*)$ ,  $f' = f(t, x, v')$ ,  $f'_* = f(t, x, v'_*)$ .

Let  $A$  be any  $3 \times 3$  matrix. We consider the translation group and the time-dependent invariant manifold specified by (24) with isometry group (25), macroscopic velocity  $v(y, t) = A(I + tA)^{-1}y$ , and a corresponding molecular dynamics simulation with atom positions  $y_{v,k}(t)$ ,  $v \in \mathbb{Z}^3, k = 1, \dots, M$ . At time  $t$  consider a ball  $B_r(0)$  of any radius  $r > 0$  centered at the origin, and another ball  $B_r(y_v)$  of the same radius centered at  $y_v = (I + tA)x_v$ , where  $x_v = v^1 e_1 + v^2 e_2 + v^3 e_3$ . Both balls may contain some simulated atoms and some nonsimulated atoms. However, no matter how irregular the simulation, if I am given the velocities of atoms in  $B_r(0)$  at time  $t$ , then I immediately know the velocities of atoms in  $B_r(y_v)$  at time  $t$ . Specifically, if  $v_1, \dots, v_p$  are the velocities of atoms in  $B_r(0)$ , then  $v_1 + Ax_v, \dots, v_p + Ax_v$  are the velocities in  $B_r(y_v)$ . Or, in the Eulerian form appropriate

to the Boltzmann equation, the velocities in  $B_r(y_v)$  are

$$(35) \quad v_1 + A(I + tA)^{-1}y_v, \dots, v_p + A(I + tA)^{-1}y_v.$$

But  $f(t, y, v)$  is supposed to represent the probability density of finding a velocity  $v$  in a small neighborhood of  $y$ . Therefore, we expect that this simulation corresponds to a molecular density function satisfying

$$(36) \quad f(t, 0, v) = f(t, y, v + A(I + tA)^{-1}y),$$

or, rearranging,

$$(37) \quad f(t, y, v) = g(t, v - A(I + tA)^{-1}y).$$

Substitution of (37) into the Boltzmann equation formally gives an immediate reduction:  $g(t, w)$  satisfies

$$(38) \quad \frac{\partial g}{\partial t} - (A(I + tA)^{-1}w) \cdot \frac{\partial g}{\partial w} = \mathbb{C}g(w),$$

where the collision operator  $\mathbb{C}$  is defined as in (32). Once, again, despite the Boltzmann equation being time-irreversible, the invariant manifold of molecular dynamics is inherited in the most obvious way. Note that at the level of the Boltzmann equation, periodicity has disappeared.

Equation (38) was originally found without reference to molecular dynamics, but rather by noticing similarities between special solutions of equations of fluid mechanics and the moment equations<sup>16</sup> of the kinetic theory Galkin [1958] and Truesdell [1956]. Recently, an existence theorem for (38) has been given James, Nota, and Velázquez [2017], with surprising implications for the invariant manifold.

The most explicit results are for *Maxwellian molecules*. These are molecules that attract with a force proportional to the inverse 5<sup>th</sup> power of their separation. For the collision kernel  $B$  appropriate to these molecules the invariance of the left and right hand sides of (32) or (38) match. We focus on the entropy (minus the  $H$ -function) given by

$$(39) \quad \eta(t) = - \int_{\mathbb{R}^3} g(t, w) \log g(t, w) dw.$$

The asymptotic analysis of self-similar solutions James, Nota, and Velázquez [ibid.] gives, for a large class of choices of  $A$ ,

$$(40) \quad \eta(t) = \rho(t) \log \left( \frac{e(t)^{3/2}}{\rho(t)} \right) + C_g,$$

<sup>16</sup>Multiply (38) by polynomials in  $w$  and integrate over  $\mathbb{R}^3$ . The study of the solutions of the moment equations has an extensive history beginning with Galkin [1958] and Truesdell [1956] and reviewed in Truesdell and Muncaster [1980]



where the density  $\rho$  and temperature  $e$  are given by explicit formulas:

$$(41) \quad \rho(t) = \int_{\mathbb{R}^3} g(t, w) dw, \quad e(t) = \int_{\mathbb{R}^3} \frac{1}{2} w^2 g(w, t) dw,$$

and  $C_g$  is constant. In these far-from-equilibrium solutions the temperature and density can be rapidly changing, and the entropy rapidly increasing. Nevertheless, the relation (40) between entropy, density and temperature expressed by (40) *is the same as for the equilibrium Maxwellian distribution*. That's true except for one small but interesting point: the constant  $C_g$  is strictly less than that of the Maxwellian distribution. From an information theoretic viewpoint, the uncertainty of positions and velocities of atoms on the invariant manifold differs from those of an equilibrium state at the same temperature and density by a constant, even as the temperature evolves rapidly to infinity.

## 7 Maxwell's equations

Maxwell's equations have a bigger invariance group, the Lorentz group<sup>17</sup> of special relativity. It would be interesting to have a look at this full group, but we shall confine attention to its Euclidean subgroup of isometries. The solutions of Maxwell's equations do not describe matter itself, but they interact with matter. In fact, almost everything we know about the structure of matter comes by interpreting this interaction. This interpretation is not straightforward because, at the relevant wavelengths, we cannot measure the scattered electric or magnetic fields directly, but only the time average of the magnitude of their cross product<sup>18</sup>. Nevertheless, increasingly, such as in quasicrystals, the classification of atomic structures is *defined* in terms of this interaction.

Even in the case of the now accepted definition of quasicrystals, the incoming radiation is plane waves. That is, we assign electric and magnetic fields, respectively,

$$(42) \quad E(y, t) = n e^{i(k \cdot y - \omega t)} \quad \text{and} \quad B(y, t) = \frac{1}{\omega} (k \times n) e^{i(k \cdot y - \omega t)},$$

where  $n \in \mathbb{C}^3$ ,  $k \in \mathbb{R}^3$ ,  $k \cdot n = 0$ ,  $\omega = c|k|$  and  $c$  is the speed of light. These plane waves exert a force  $e(E + v \times B)$  on each electron of the structure, which vibrates with velocity  $v$ . Moving charges generate electromagnetic fields and so, the vibrating electrons, each with charge  $e$ , send out spherical waves which in the far-field are again approximately plane waves. The rigorous asymptotics of this process is delicate [Friescke, James, and Jüstel \[2017\]](#) and involves several small parameters in addition to the Fresnel number

<sup>17</sup>in fact, the conformal Lorentz group, which includes dilatations as well as Lorentz transformations [Bateman \[1910\]](#).

<sup>18</sup>that is, the time average of the Poynting vector (see [Friescke, James, and Jüstel \[2016\]](#), p. 1196).

$\text{dia}(\Omega)^2 |k|/d \ll 1$ . Here  $\Omega \subset \mathbb{R}^3$  is the illuminated region, and  $d$  is the distance to the detector. The results are formulas in terms of  $E_0(y) = n e^{ik \cdot y}$  for the electric and magnetic fields in the far-field:

$$(43) \quad \begin{aligned} E_{out}(y, t) &= -c_{el} \frac{e^{i(k'(y) \cdot y - \omega t)}}{|y - y_c|} \left( I - \frac{k'(y)}{|k'(y)|} \otimes \frac{k'(y)}{|k'(y)|} \right) \int_{\Omega} E_0(z) \rho(z) e^{-ik'(y) \cdot z} dz, \\ B_{out}(y, t) &= \frac{1}{\omega} k'(y) \times E_{out}(y, t), \end{aligned}$$

where  $\rho : \mathbb{R}^3 \rightarrow \mathbb{R}^{\geq}$  is the electronic density,  $x_c \in \Omega$  is a typical point of the illuminated region,  $c_{el}$  is a universal constant depending on the charge and mass of an electron, and

$$(44) \quad k'(y) = \frac{\omega}{c} \frac{y - y_c}{|y - y_c|}.$$

With a simple idealized example we can begin to understand plane wave X-ray methods. In the notation of [Section 2](#), we assume that the electronic density is a sum of Dirac masses at the points of a Bravais lattice generated by the linearly independing vectors  $e_1, e_2, e_3$ ,

$$(45) \quad \rho(y) = \sum_{z \in \mathcal{L}(e_1, e_2, e_3) \cap \Omega} \delta_z(y).$$

With this choice the integral in (43) is

$$(46) \quad \int_{\Omega} n e^{ik \cdot z} \rho(z) e^{-ik'(y) \cdot z} dz = \sum_{z \in \mathcal{L}(e_1, e_2, e_3) \cap \Omega} n e^{-i((k'(y) - k) \cdot z)}.$$

Therefore, if  $k'(y) - k$  belongs to the reciprocal lattice  $\mathcal{L}(e^1, e^2, e^3)$ ,  $e^i \cdot e_j = 2\pi \delta_j^i$ , then the exponential factor contributes 1 to the (complex<sup>19</sup>) magnitude of (46) for every lattice point: that is, constructive interference.

How much of this constructive interference is a consequence of choosing the electronic density to be a sum of Dirac masses? Almost nothing [Friesecke, James, and Jüstel \[2016\]](#) and [Friesecke \[2007\]](#): *Suppose instead we assume*

$$(47) \quad \rho_r(y) = \sum_{z \in \mathcal{L}(e_1, e_2, e_3), |z| < r} \varphi(y - z)$$

for a smooth function  $\varphi : \mathbb{R}^3 \rightarrow \mathbb{R}^{\geq}$  with compact support, or, more generally, in the Schwartz class  $\mathcal{S}(\mathbb{R}^3)$ . Here we have chosen  $\Omega = B_r(0)$  as the illuminated region. Then,

<sup>19</sup>Ibid. The time average of the Poynting vector for time-harmonic radiation is, up to a constant factor, the complex magnitude of the electric field in the time harmonic case.

the limit in the sense of distributions of the complex magnitude of the integral in the expression (43) for  $E_{out}$  is

$$(48) \quad \lim_{r \rightarrow \infty} \left| \int_{B_r(0)} E_0(z) \rho(z) e^{-ik'(y) \cdot z} dz \right| = \sum_{z' \in \mathfrak{L}(e^1, e^2, e^3)} |\hat{\varphi}(z')| \delta_{z'}(k'(y) - k).$$

From this result we not only see constructive interference but also strong destructive interference: the sum in (48) is zero when  $k'(y) - k$  does not belong to the reciprocal lattice. It is this result that underlies the 2 to 4 order-of-magnitude difference between peak heights and background, and the sharpness of the peaks, in X-ray methods. This in turn is what makes this method so accurate for structure determination. Discovery, improvement and application of the method has led to no less than 14 Nobel prizes.

All this works because of matching symmetries. In the calculation (46) it is the ability to combine the phase factors, or more precisely, that the translation group acting on plane waves gives a phase factor times the plane wave back again. For more general choices of  $\rho$ , we can use translation invariance (up to the multiplicative phase factor) of  $\rho(z)$  on the left hand side of (46) to condense the integral to a lattice sum of an integral over the unit cell, to see constructive interference. A more powerful method is the italicized theorem just above, which is proved by a direct application of the Poisson summation formula [Friesecke, James, and Jüstel \[2016\]](#). The property of plane waves being used is, for the translation  $g_c = (I \mid c)$ ,

$$(49) \quad g_c(n e^{ik \cdot y}) := n e^{ik \cdot (y+c)} = e^{ik \cdot c} (n e^{ik \cdot y}),$$

i.e., with the group action indicated on the left of (49), the plane wave  $n e^{ik \cdot y}$  is an eigenfunction of the translation group. The eigenvalues  $e^{ik \cdot c}$  are the characters of this Abelian group. The two key mathematical properties of plane waves are that, with the natural action (49), (i) they are eigenfunctions of the translation group and (ii) they are solutions of Maxwell's equations.

So much for plane waves. In principle, everything should work in the same way for any other Abelian isometry group  $G$ . As seen in [Section 2](#) and elsewhere, many of the most studied structures today are not crystals, and it would be good to have an accurate method of structure determination for them. Perhaps the most interesting mathematically are helical structures like single-walled carbon nanotubes<sup>20</sup>. For one, the helical groups do not fit the classification scheme of the International Tables of Crystallography – there are infinitely many helical groups according to that scheme. For another, helical (as well as many other) structures tend to resist crystallization. Third, even if helical structures can be crystallized, it is quite unclear that the structure will be close to the native structure.

---

<sup>20</sup>Due to the issues discussed here and the typical presence of mixed chiralities in samples, the lattice parameters of carbon nanotubes are not accurately known.

We shall consider time harmonic solutions of Maxwell's equations:

$$E(y, t) = E_0(y)e^{-i\omega t}, \quad B(y, t) = B_0(y)e^{-i\omega t}, \quad E_0 : \mathbb{R}^3 \rightarrow \mathbb{C}^3, \quad B_0 : \mathbb{R}^3 \rightarrow \mathbb{C}^3$$

In this case Maxwell's equations become

$$(50) \quad \Delta E_0 = -\frac{\omega^2}{c^2} E_0, \quad \operatorname{div} E_0 = 0, \quad B_0 = -\frac{i}{\omega} \operatorname{curl} E_0.$$

For time-harmonic radiation the electric and magnetic fields in the far-field are still given by (43), but now for a solution  $E_0(y)$  of (50).

A critical part is choosing the action so that (i) is nontrivial and (ii) exploits the invariance of Maxwell's equations. The right action is

$$(51) \quad \text{for } g = (Q | c) \in G, \quad g[E](y, t) = QE(g^{-1}(y), t) = QE(Q^T(y - c), t).$$

Here we use the bracket notation [...] to distinguish the action from that already introduced,  $g(y) = Qy + c$ . Summarizing, we have *design equations*:

$$(52) \quad (i) \text{ for all } g \in G, \quad g[E_0] = \chi_g E_0, \text{ and } (ii) \quad E_0 \text{ satisfies Maxwell's equations (50).}$$

Of course, plane waves satisfy the design equations.

The largest (discrete) Abelian helical group is

$$(53) \quad \{h^i g^j : i \in \mathbb{Z}, \quad j = 1, \dots, n\} \text{ where } h = (R_\theta | \tau e), \quad g = (R_{2\pi/n} | 0).$$

with  $R_\psi \in \operatorname{SO}(3)$  having angle  $\psi$  and axis through  $e$ ,  $|e| = 1$ ,  $0 < \theta < 2\pi$  and<sup>21</sup>  $n \in \mathbb{N}$ .

Exploiting (52) for the helical group (53) is quite easy if we begin with (i). First, the eigenvalue  $\chi_g$ ,  $g \in G$ , is seen to be a bounded continuous homomorphism from  $G$  to  $\mathbb{C} \setminus 0$  under multiplication in  $\mathbb{C}$ , and therefore a character of  $G$ . The characters are  $\chi_g = \chi(\theta, \tau) = e^{i(\alpha\theta + \beta\tau)}$ ,  $\alpha \in \mathbb{Z}$ ,  $\beta \in \mathbb{R}$ . Then, one can easily find the general form of  $E_0$  satisfying (i): in cylindrical coordinates  $(r, \varphi, z)$  this is  $E_0(r, \varphi, z) = e^{i(\alpha\varphi + \beta z)} R_\varphi E_0(r, 0, 0)$ . Finally, substitution of the latter into Maxwell's equations reduces them to a solvable system of ODEs. A general form of the result are *twisted waves*<sup>22</sup>:

$$(54) \quad E(y, t) = \frac{1}{2\pi} e^{-i\omega t} \int_{-\pi}^{\pi} e^{i\alpha\psi} R_\psi n e^{i y \cdot R_\psi k} d\psi, \quad k = (0, \gamma, \beta).$$

Here,  $n \in \mathbb{C}^3$  satisfies  $n \cdot k = 0$ . A picture of a twisted wave is shown in Figure 5.

<sup>21</sup>Strictly speaking, to be a helical group,  $\theta$  is an irrational multiple of  $2\pi$  but we will not need this restriction.

<sup>22</sup>For the form given here, see Jüstel [2014]; for alternative expressions see Jüstel, Friesecke, and James [2016].

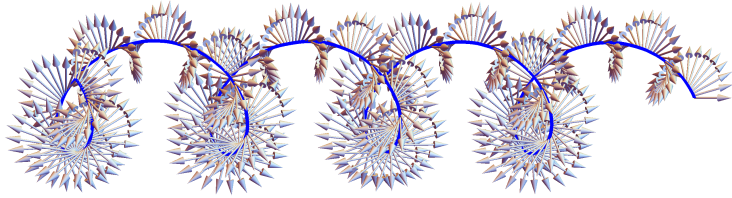


Figure 5: A twisted wave showing electric field vectors plotted along an integral curve (blue) of the Poynting vector.

Theoretically, twisted waves can be used for structure determination of helical structures similar to the way plane waves are used on periodic structures. A complete scheme for theoretical structure determination is proposed in [Friesecke, James, and Jüstel \[2016\]](#) and [Jüstel, Friesecke, and James \[2016\]](#). The key parameters that are varied are  $\alpha$  and  $\beta$ . We cannot describe this in detail here, but one can get a glimpse of the idea from [Figure 6](#). Suppose we have an helical objective structure as shown in [6a](#). We are looking down the axis. Each yellow atom sees the same environment; each red atom sees the same environment; each green atom sees the same environment. In [6b](#) we have superposed on this structure a twisted wave whose values of  $\alpha$ ,  $\beta$ ,  $n$  are tuned to give constructive interference, and we have plotted just the electric field vectors at the atoms. As one can see, all the red vectors are parallel, all the green vectors are parallel, and all the yellow vectors are

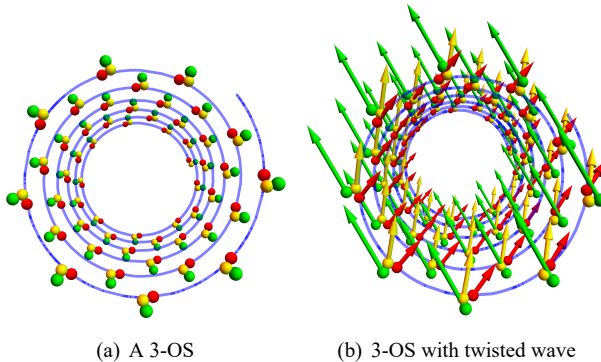


Figure 6: (a) A 3-OS with green, yellow and red atoms, viewed down the axis. (b) The same structure with a superimposed twisted wave highlighted at the atom positions.

parallel in this projection. One can imagine that, if the phases are properly tuned one can get constructive interference, measured at a detector on the axis. Moreover, the different length vectors should in fact give information about what is in the unit cell.

These pictures illustrate constructive interference. Destructive interference that occurs when the parameters or structure is tuned slightly off resonance relies on a far reaching generalization of the Poisson summation formula [Weil \[1964\]](#) which in turn requires that the group be extendable to a continuous symmetry group. In the case of the helical group this generalization can also be seen in a simpler way [Friezecke, James, and Jüstel \[2016\]](#).

## 8 Perspective

Clearly the subject of structure and invariance has a ragged boundary! Some questions that could have been considered a century ago seem not to have been asked, and simple questions posed then seem to be excruciatingly difficult. But the subject has a vibrant connection with materials science and technology today, with links to nanoscience, quasicrystals, origami, structure determination and multiscale mathematics. In this section we make a selection of what are (to the author) intriguing mathematical problems related to this line of thinking.

The most fundamental question seems to us to be: why do elements in the Periodic Table, and nanostructures made with one element, having widely differing atomic forces and bonding patterns, choose to crystallize as objective structures? Can that be proved in some framework, while initially avoiding the question of what is the detailed structure?

There is an intriguing link between subsets of nondiscrete groups of isometries, objective structures and quasicrystals. What are these groups, and how does one rationally choose the subsets. Does this lead to a more physically-based approach to quasicrystalline structures than the projection method?

Seeking some kind of general nonequilibrium statistical mechanics, that has something like the simplicity of equilibrium statistical mechanics, is in the author's view hopeless. After all, even the Boltzmann equation treats only the more rudimentary kind of material, and it has all the complexity of a general initial-value problem of a nonlinear integro-differential equation for a function of 7 variables. A classic approach is to try to simplify by looking near equilibrium. A fresh approach could be the following. We have followed a far-from-equilibrium invariant manifold from molecular dynamics to the Boltzmann equation to continuum mechanics. There seem to be many coincidences, such as an explicit relationship between density, temperature and entropy that holds far from equilibrium. These are highly suggestive that there may be a relatively simple statistical mechanics *on this manifold*. There, the only gradient is the velocity gradient. Of course, any such statistical mechanics cannot be based on  $\text{Hamiltonian} = \text{const.}$

The method of objective molecular dynamics presented in [Section 4](#) could be more widely used. There are also fundamental mathematical questions (stability) and subtle numerical issues (efficiency). Note that in a general continuum flow, a piecewise constant

spatial approximation of the Lagrangian velocity field nominally gives a set of elements each with a constant  $A$ . Is there a general multiscale method here?

What is the scope of the interaction light with matter? For example, even though the structure in [Figure 4](#) is not the orbit of a discrete group of isometries, it seems likely that, with the right radiation, we could get a pretty strong constructive interference from it.

And finally, like a lot of mathematics, the key to understanding origami seems to be rigidity. But, established lines of thinking about rigidity in differential geometry or elasticity seem not to be fruitful. On the other hand, the link between martensitic phase transformations and origami, already pioneered by [Conti and Maggi \[2008\]](#), seems to be highly suggestive.

**Acknowledgments.** The author thanks A. Banerjee, K. Dayal, R. Elliott, F. Feng, G. Friesecke, D. Jüstel, P. Plucinsky and R. Twarock for stimulating discussions about these topics.

## References

- Edgar C. Bain (1924). “The Nature of Martensite”. *Trans. AIME* 70.1, p. 25 (cit. on p. [3986](#)).
- John M. Ball (1981). “Global invertibility of Sobolev functions and the interpenetration of matter”. *Proceedings of the Royal Society of Edinburgh Section A: Mathematics* 88.3-4, pp. 315–328 (cit. on p. [4001](#)).
- Amartya S Banerjee, Ryan S Elliott, and Richard D James (2015). “A spectral scheme for Kohn–Sham density functional theory of clusters”. *Journal of Computational Physics* 287, pp. 226–253 (cit. on p. [3988](#)).
- Harry Bateman (1910). “The transformation of the electrodynamical equations”. *Proceedings of the London Mathematical Society* 2.1, pp. 223–264 (cit. on p. [4004](#)).
- Mathias Carlen, Ben Laurie, John H. Maddocks, and Jana Smutny (2005). “Biarcs, global radius of curvature, and the computation of ideal knot shapes”. *Physical and numerical models in knot theory* 36, pp. 75–108 (cit. on p. [4001](#)).
- Donald L. D. Caspar and Aaron Klug (1962). “Physical principles in the construction of regular viruses”. In: *Cold Spring Harbor symposia on quantitative biology*. Vol. 27. Cold Spring Harbor Laboratory Press, pp. 1–24 (cit. on pp. [3988](#), [3998](#)).
- Philippe G. Ciarlet and Jindřich Nečas (1987). “Injectivity and self-contact in nonlinear elasticity”. *Archive for Rational Mechanics and Analysis* 97.3, pp. 171–188 (cit. on p. [4001](#)).

- Sergio Conti and Francesco Maggi (2008). “Confining thin elastic sheets and folding paper”. *Archive for Rational Mechanics and Analysis* 187.1, pp. 1–48 (cit. on pp. 4001, 4010).
- H. S. M. Coxeter (1971). “Virus macromolecules and geodesic domes”. *A spectrum of mathematics*, pp. 98–107 (cit. on p. 3998).
- H. R. Crane (1950). “Principles and problems of biological growth”. *The Scientific Monthly* 70.6, pp. 376–389 (cit. on p. 3988).
- Francis H. C. Crick and James D. Watson (1956). “Structure of small viruses”. *Nature* 177.4506, pp. 473–475 (cit. on p. 3988).
- Kaushik Dayal and Richard D. James (2010). “Nonequilibrium molecular dynamics for bulk materials and nanostructures”. *Journal of the Mechanics and Physics of Solids* 58.2, pp. 145–163 (cit. on pp. 3996, 3997).
- (2012). “Design of viscometers corresponding to a universal molecular simulation method”. *Journal of Fluid Mechanics* 691, pp. 461–486 (cit. on pp. 3996, 3998).
- Nikolai P. Dolbilin, J. C. Lagarias, and Marjorie Senechal (1998). “Multiregular point systems”. *Discrete & Computational Geometry* 20.4, pp. 477–498 (cit. on pp. 3988, 3989).
- Traian Dumitrica and Richard D. James (2007). “Objective molecular dynamics”. *Journal of the Mechanics and Physics of Solids* 55.10, pp. 2206–2236 (cit. on p. 3991).
- J. L. Ericksen (1977). “Special Topics in Elastostatics”. In: *Advances in Applied Mechanics*. Ed. by C.-S. Yih. Vol. 17. Academic Press, pp. 189–244 (cit. on p. 3993).
- Leonhard Euler (1745). “Recherches physiques sur la nature des moindres parties de la matière”. *Berlin Academy Memoires* 1. (*Opera omnia*, III. 1: 6–15), pp. 28–32 (cit. on p. 3986).
- Wayne Falk and Richard D. James (2006). “Elasticity theory for self-assembled protein lattices with application to the martensitic phase transition in bacteriophage T4 tail sheath”. *Physical Review E* 73.1, p. 011917 (cit. on p. 3989).
- Fan Feng, Paul Plucinsky, and Richard D. James (2017). “Phase transformations in helical structures”. Preprint (cit. on p. 4001).
- L. C. Flatley and Florian Theil (2015). “Face-centered cubic crystallization of atomistic configurations”. *Archive for Rational Mechanics and Analysis* 218.1, pp. 363–416 (cit. on p. 3986).
- Gero Friesecke (2007). *Lectures on Fourier Analysis*. University of Warwick, Coventry, UK (cit. on p. 4005).
- Gero Friesecke, Richard D. James, and Dominik Jüstel (2016). “Twisted X-rays: incoming waveforms yielding discrete diffraction patterns for helical structures”. *SIAM Journal on Applied Mathematics* 76.3, pp. 1191–1218 (cit. on pp. 4004–4006, 4008, 4009).
- (2017). “The far-field intensity in kinematic X-ray diffraction for general incoming radiation”. Preprint (cit. on p. 4004).



- Gero Friesecke and Florian Theil (2002). “Validity and Failure of the Cauchy-Born Hypothesis in a Two-Dimensional Mass-Spring Lattice.” *Journal of Nonlinear Science* 12.5 (cit. on p. 3986).
- V. So. Galkin (1958). “On a class of solutions of Grad’s moment equations”. *Journal of Applied Mathematics and Mechanics* 22.3. (Russian version, *PMM* 20, 445–446, (1956)), pp. 532–536 (cit. on p. 4003).
- Clifford S. Gardner and Charles Radin (1979). “The infinite-volume ground state of the Lennard-Jones potential”. *Journal of Statistical Physics* 20.6, pp. 719–724 (cit. on p. 3986).
- D. Hobbs, J. Hafner, and D. Spišák (2003). “Understanding the complex metallic element Mn. I. Crystalline and noncollinear magnetic structure of  $\alpha$ -Mn”. *Physical Review B* 68.1, p. 014407 (cit. on p. 3988).
- Robert Hooke (1665). *Micrographia: or, Some physiological descriptions of minute bodies made by magnifying glasses*. First. London: J. Martyn and J. Allestry (cit. on p. 3986).
- Giuliana Indelicato, Tom Keef, Paolo Cermelli, David G. Salthouse, Reidun Twarock, and Giovanni Zanzotto (2012). “Structural transformations in quasicrystals induced by higher dimensional lattice transitions”. In: *Proc. R. Soc. A*. Vol. 468. The Royal Society, pp. 1452–1471 (cit. on p. 4001).
- Tadeusz Iwaniec, Leonid V. Kovalev, and Jani Onninen (2011). “Diffeomorphic approximation of Sobolev homeomorphisms”. *Archive for rational mechanics and analysis* 201.3, pp. 1047–1067 (cit. on p. 4001).
- Richard D. James (2015). “Continuum mechanics”. In: *The Princeton Companion to Applied Mathematics*. Ed. by Mark R. Dennis, Paul Glendinning, Paul A. Martin, Fadil Santosa, and Jared Tanner. Princeton University Press, pp. 446–458 (cit. on p. 3996).
- Richard D. James, Alessia Nota, and Juan J. L. Velázquez (2017). “Self-similar profiles for homoenergetic solutions of the Boltzmann equation: particle velocity distribution and entropy”. arXiv: 1710.03653 (cit. on p. 4003).
- Richard. D. James (2006). “Objective Structures”. *Journal of the Mechanics and Physics of Solids* 54.11, pp. 2354–2390 (cit. on pp. 3987, 3988).
- D. Jüstel (2014). “Radiation for the analysis of molecular structures with non-crystalline symmetry: modelling and representation theoretic design”. PhD thesis. Technical University of Munich (cit. on p. 4007).
- Dominik Jüstel, Gero Friesecke, and Richard D. James (2016). “Bragg–von Laue diffraction generalized to twisted X-rays”. *Acta Crystallographica Section A: Foundations and Advances* 72.2, pp. 190–196 (cit. on pp. 4007, 4008).
- Thomas Keef, Jessica P. Wardman, Neil A. Ranson, Peter G. Stockley, and Reidun Twarock (2013). “Structural constraints on the three-dimensional geometry of simple viruses: case studies of a new predictive tool”. *Acta Crystallographica Section A: Foundations of Crystallography* 69.2, pp. 140–150 (cit. on p. 4001).

- James Clerk Maxwell (1867). “On the dynamical theory of gases”. *Philosophical transactions of the Royal Society of London* 157, pp. 49–88 (cit. on p. 4002).
- Koryo Miura, Toshikazu Kawasaki, Tomohiro Tachi, Ryuhei Uehara, Robert J. Lang, and Patsy Wang-Iverson (2015). *Origami<sup>6</sup>: I. Mathematics*. American Mathematical Society (cit. on p. 3999).
- Paul Plucinsky, Fan Feng, and Richard D. James (2017). “An algorithm to generate all possible generalized Miura origami”. Preprint (cit. on p. 3999).
- Florian Theil (2006). “A proof of crystallization in two dimensions”. *Communications in Mathematical Physics* 262.1, pp. 209–236 (cit. on p. 3986).
- C. Truesdell (1956). “On the pressures and the flux of energy in a gas according to Maxwell’s kinetic theory, II”. *Journal of Rational Mechanics and Analysis* 5.1, pp. 55–128 (cit. on p. 4003).
- Clifford Truesdell and Robert G Muncaster (1980). *Fundamentals of Maxwell’s Kinetic Theory of a Simple Monatomic Gas*. Vol. 83. Academic Press (cit. on p. 4003).
- Cédric Villani and Clément Mouhot (2015). “Kinetic theory”. In: *The Princeton Companion to Applied Mathematics*. Ed. by Mark R. Dennis, Paul Glendinning, Paul A. Martin, Fadil Santosa, and Jared Tanner. Princeton University Press, pp. 446–458 (cit. on p. 4002).
- André Weil (1964). “Sur certains groupes d’opérateurs unitaires”. *Acta mathematica* 111.1, pp. 143–211 (cit. on p. 4009).

Received 2017-11-30.

RICHARD D. JAMES  
DEPARTMENT OF AEROSPACE ENGINEERING AND MECHANICS  
UNIVERSITY OF MINNESOTA  
[james@umn.edu](mailto:james@umn.edu)



# MATHEMATICS FOR CRYO-ELECTRON MICROSCOPY

AMIT SINGER

## Abstract

Single-particle cryo-electron microscopy (cryo-EM) has recently joined X-ray crystallography and NMR spectroscopy as a high-resolution structural method for biological macromolecules. Cryo-EM was selected by Nature Methods as Method of the Year 2015, large scale investments in cryo-EM facilities are being made all over the world, and the Nobel Prize in Chemistry 2017 was awarded to Jacques Dubochet, Joachim Frank and Richard Henderson “for developing cryo-electron microscopy for the high-resolution structure determination of biomolecules in solution”. This paper focuses on the mathematical principles underlying existing algorithms for structure determination using single particle cryo-EM.

## 1 Introduction

The field of structural biology is currently undergoing a transformative change [Kühlbrandt \[2014\]](#) and [Smith and Rubinstein \[2014\]](#). Structures of many biomolecular targets previously insurmountable by X-ray crystallography are now being obtained using single particle cryo-EM to resolutions beyond 4Å on a regular basis [Liao, Cao, Julius, and Cheng \[2013\]](#), [Amunts et al. \[2014\]](#), and [Bartesaghi, Merk, Banerjee, Matthies, X. Wu, Milne, and Subramaniam \[2015\]](#). This leap in cryo-EM technology, as recognized by the 2017 Nobel Prize in Chemistry, is mainly due to hardware advancements including the invention of the direct electron detector and the methodological development of algorithms for data processing. Cryo-EM is a very general and powerful technique because it does not require the formation of crystalline arrays of macromolecules. In addition, unlike X-ray crystallography and nuclear magnetic resonance (NMR) that measure ensembles of particles, single particle cryo-EM produces images of individual particles. Cryo-EM therefore

---

Partially supported by Award Number R01GM090200 from the NIGMS, FA9550-17-1-0291 from AFOSR, Simons Foundation Math+X Investigator Award, and the Moore Foundation Data-Driven Discovery Investigator Award.

*MSC2010:* primary 92C55; secondary 68U10, 44A12, 62H12, 33C55, 90C22.

has the potential to analyze conformational changes and energy landscapes associated with structures of complexes in different functional states.

As there exist many excellent review articles and textbooks on single particle cryo-EM [Frank \[2006\]](#), [van Heel, Gowen, et al. \[2000\]](#), [Nogales \[2016\]](#), [Glaeser \[2016\]](#), [Subramaniam, Kühlbrandt, and Henderson \[2016\]](#), [C. O. S. Sorzano and Carazo \[2017\]](#), and [F. J. Sigworth \[2016\]](#), we choose to solely focus here on the mathematical foundations of this technique. Topics of great importance to practitioners, such as the physics and optics of the electron microscope, sample preparation, and data acquisition are not treated here.

In cryo-EM, biological macromolecules are imaged in an electron microscope. The molecules are rapidly frozen in a thin layer of vitreous ice, trapping them in a nearly-physiological state. The molecules are randomly oriented and positioned within the ice layer. The electron microscope produces a two-dimensional tomographic projection image (called a micrograph) of the molecules embedded in the ice layer. More specifically, what is being measured by the detector is the integral in the direction of the beaming electrons of the electrostatic potential of the individual molecules.

Cryo-EM images, however, have very low contrast, due to the absence of heavy-metal stains or other contrast enhancements, and have very high noise due to the small electron doses that can be applied to the specimen without causing too much radiation damage. The first step in the computational pipeline is to select “particles” from the micrographs, that is, to crop from each micrograph several small size images each containing a single projection image, ideally centered. The molecule orientations associated with the particle images are unknown. In addition, particle images are not perfectly centered, but this would be of lesser concern to us for now.

The imaging modality is akin to the parallel beam model in Computerized Tomography (CT) of medical images, where a three-dimensional density map of an organ needs to be estimated from tomographic images. There are two aspects that make single particle reconstruction (SPR) from cryo-EM more challenging compared to classical CT. First, in medical imaging the patient avoids movement, hence viewing directions of individual projections are known to the scanning device, whereas in cryo-EM the viewing directions are unknown. Electron Tomography (ET) employs tilting and is often used for cellular imaging, providing reconstructions of lower resolution due to increased radiation damage for the entire tilt series. While it is possible to tilt the specimen and register relative viewing directions among images within a tilt series, radiation damage destroys high frequency content and it is much more difficult to obtain high resolution reconstructions using ET. In SPR, each particle image corresponds to a different molecule, ideally of the same structure, but at different and unknown orientation. Second, the signal-to-noise ratio (SNR) typical of cryo-EM images is smaller than one (more noise than signal). Thus, to obtain a reliable three-dimensional density map of a molecule, the information from many images of identical molecules must be combined.

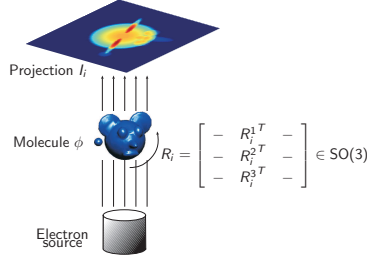


Figure 1: Schematic drawing of the imaging process: every projection image corresponds to some unknown rotation of the unknown molecule. The effect of the point spread function is not shown here.

## 2 Image formation model and inverse problems

The mathematical image formation model is as follows (Figure 1). Let  $\phi : \mathbb{R}^3 \rightarrow \mathbb{R}$  be the electrostatic potential of the molecule. Suppose that following the step of particle picking, the dataset contains  $n$  particle images, denoted  $I_1, \dots, I_n$ . The image  $I_i$  is formed by first rotating  $\phi$  by a rotation  $R_i$  in  $SO(3)$ , then projecting the rotated molecule in the  $z$ -direction, convolving it with a point spread function  $H_i$ , sampling on a Cartesian grid of pixels of size  $L \times L$ , and contaminating with noise:

$$(1) \quad I_i(x, y) = H_i \star \int_{-\infty}^{\infty} \phi(R_i^T r) dz + \text{“noise”}, \quad r = (x, y, z)^T.$$

The rotations  $R_1, \dots, R_n \in SO(3)$  are unknown. The Fourier transform of the point spread function is called the contrast transfer function (CTF), and it is typically known, or can be estimated from the data, at least approximately, although it may vary from one image to another. Equivalently, we may rewrite the forward model (Equation (1)) as

$$(2) \quad I_i = H_i \star PR \circ \phi + \text{“noise”},$$

where  $R \circ \phi(r) = \phi(R^T r)$  and  $P$  is the tomographic projection operator in the  $z$ -direction,  $Pf(x, y) = \int_{\mathbb{R}} f(x, y, z) dz$ . We write “noise” in Equations (1) and (2) as a full discussion of the noise statistics and its possible dependence on the structure itself (i.e., structural noise) are beyond the scope of this paper.

The basic cryo-EM inverse problem, called the cryo-EM reconstruction problem, is to estimate  $\phi$  given  $I_1, \dots, I_n$  and  $H_1, \dots, H_n$ , without knowing  $R_1, \dots, R_n$ . Notice that cryo-EM reconstruction is a non-linear inverse problem, because the rotations are unknown; if the rotations were known, then it would become a linear inverse problem, for

which there exist many classical solvers. Because images are finitely sampled,  $\phi$  cannot be estimated beyond the resolution of the input images.

An even more challenging inverse problem is the so-called heterogeneity cryo-EM problem. Here each image may originate from a different molecular structure corresponding to possible structural variations. That is, to each image  $I_i$  there may correspond a different molecular structure  $\phi_i$ . The goal is then to estimate  $\phi_1, \dots, \phi_n$  from  $I_1, \dots, I_n$ , again, without knowing the rotations  $R_1, \dots, R_n$ . Clearly, as stated, this is an ill-posed inverse problem, since we are required to estimate more output parameters (three-dimensional structures) than input data (two-dimensional images). In order to have any hope of making progress with this problem, we would need to make some restrictive assumptions about the potential functions  $\phi_1, \dots, \phi_n$ . For example, the assumption of discrete variability implies that there is only a finite number of distinct conformations from which the potential functions are sampled from. Then, the goal is to estimate the number of conformations, the conformations themselves, and their distribution. Another popular assumption is that of continuous variability with a small number of flexible motions, so that  $\phi_1, \dots, \phi_n$  are sampled from a low-dimensional manifold of conformations. Either way, the problem is potentially well-posed only by assuming an underlying low-dimensional structure on the distribution of possible conformations.

In order to make this exposition less technical, we are going to make an unrealistic assumption of ideally localized point spread functions, or equivalently, constant contrast transfer functions, so that  $H_1, \dots, H_n$  are eliminated from all further consideration here. All methods and analyses considered below can be generalized to include the effect of non-ideal CTFs, unless specifically mentioned otherwise.

### 3 Solving the basic cryo-EM inverse problem for clean images

Even with clean projection images, the reconstruction problem is not completely obvious (Figure 2). A key element to determining the rotations of the images is the Fourier projection slice theorem [Natterer \[1986\]](#) that states that the two-dimensional Fourier transform of a tomographic projection image is the restriction of the three-dimensional Fourier transform of  $\phi$  to a planar central slice perpendicular to the viewing direction:

$$(3) \quad \mathcal{F} PR \circ \phi = SR \circ \mathcal{F} \phi,$$

where  $\mathcal{F}$  denotes the Fourier transform (over  $\mathbb{R}^2$  on the left hand side of [Equation \(3\)](#), and over  $\mathbb{R}^3$  on the right hand side of [Equation \(3\)](#)), and  $S$  is the restriction operator to the  $xy$ -plane ( $z = 0$ ).

The Fourier slice theorem implies the common line property: the intersection of two (non-identical) central slices is a line. Therefore, for any pair of projection images, there

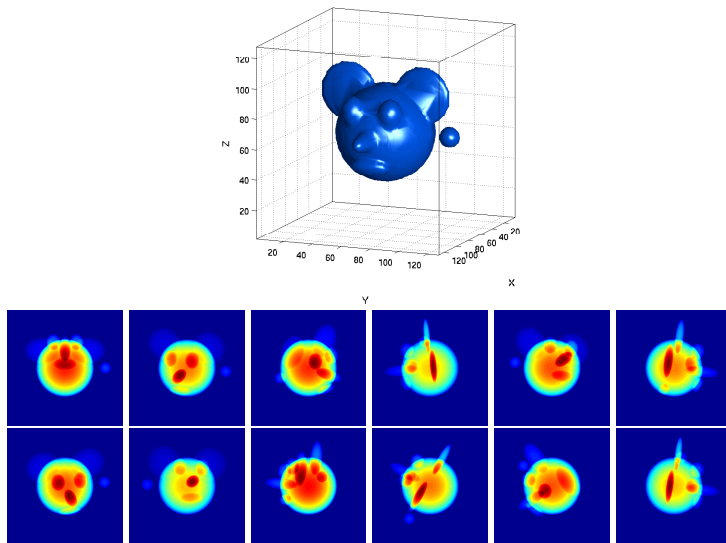


Figure 2: Illustration of the basic cryo-EM inverse problem for clean images. How to estimate the three-dimensional structure (top) from clean projection images taken at unknown viewing angles (bottom)?



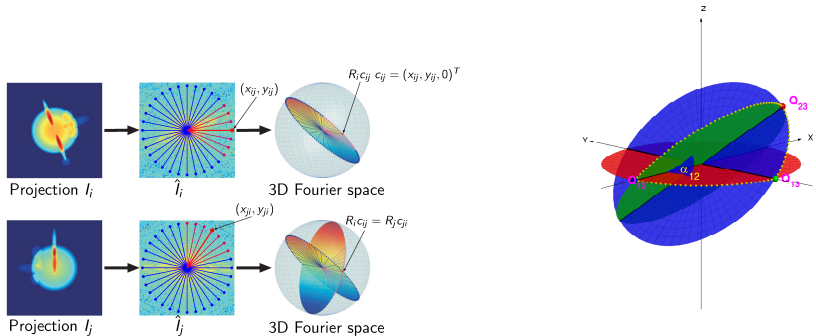


Figure 3: Left: Illustration of the Fourier slice theorem and the common line property. Right: Angular reconstitution

is a pair of central lines (one in each image) on which their Fourier transforms agree (Figure 3, left panel). For non-symmetric generic molecular structures it is possible to uniquely identify the common-line, for example, by cross-correlating all possible central lines in one image with all possible central lines in the other image, and choosing the pair of lines with maximum cross-correlation. The common line pins down two out of the three Euler angles associated with the relative rotation  $R_i^{-1}R_j$  between images  $I_i$  and  $I_j$ . The angle between the two central planes is not determined by the common line. In order to determine it, a third image is added, and the three common line pairs between the three images uniquely determine their relative rotations up to a global reflection (Figure 3, right panel). This procedure is known as “angular reconstitution”, and it was proposed independently by Vainshtein and A. Goncharov [1986] and van Heel [1987]. Notice that the handedness of the molecule cannot be determined by single particle cryo-EM, because the original three-dimensional object and its reflection give rise to identical sets of projection images with rotations related by the following conjugation,  $\tilde{R}_i = JR_iJ^{-1}$ , with  $J = J^{-1} = \text{diag}(1, 1, -1)$ . For molecules with non-trivial point group symmetry, e.g., cyclic symmetry, there are multiple common lines between pairs of images, and even self-common lines that enable rotation assignment from fewer images.

As a side comment, notice that for the analog problem in lower dimension of reconstructing a two-dimensional object from its one-dimensional tomographic projections taken at unknown directions, the Fourier slice theorem does not help in determining the viewing directions, because it only has a trivial geometric implication that the Fourier transform of the line projections intersect a point, the zero frequency, corresponding to the total mass of the density. Yet, it is possible to uniquely determine the viewing directions by relating the moments of the projections with those of the original object, as originally proposed by A.

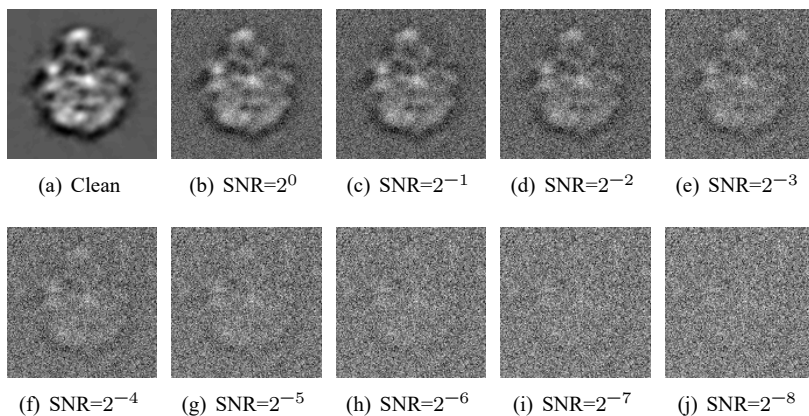


Figure 4: Simulated projections of size  $129 \times 129$  pixels at various levels of SNR.

[Goncharov \[1987\]](#) and further improved and analyzed by [Basu and Bresler \[2000a,b\]](#). An extension of the moment method to 3-D cryo-EM reconstruction was also proposed [A. B. Goncharov \[1988\]](#) and [A. Goncharov and Gelfand \[1988\]](#). As the moment based method is very sensitive to noise and cannot handle varying CTF in a straightforward manner, it mostly remained a theoretical curiosity.

## 4 Solving the basic cryo-EM inverse problem for noisy images

For noisy images it is more difficult to correctly identify the common lines. [Figure 4](#) shows a simulated clean projection image contaminated by white Gaussian noise at various levels of SNR, defined as the ratio between the signal variance to noise variance. [Section 4](#) specifies the fraction of correctly identified common lines as a function of the SNR for the simulated images, where a common line is considered to be correctly identified if both central lines deviate by no more than  $10^\circ$  from their true directions. The fraction of correctly identified common lines deteriorates quickly with the SNR. For SNR values typical of experimental images, the fraction of correctly identified common lines is around 0.1, and can be even lower for smaller molecules of lower SNR. As angular reconstitution requires three pairs of common lines to be correctly identified, its probability to succeed is only  $10^{-3}$ . Moreover, the procedure of estimating the rotations of additional images sequentially using their common lines with the previously rotationally assigned images quickly accumulates errors.

$\log_2(\text{SNR})$	$p$
20	0.997
0	0.980
-1	0.956
-2	0.890
-3	0.764
-4	0.575
-5	0.345
-6	0.157
-7	0.064
-8	0.028
-9	0.019

Table 1: Fraction  $p$  of correctly identified common lines as a function of the SNR.

The failure of angular reconstitution at low SNR, raises the question of how to solve the cryo-EM reconstruction problem at low SNR. One possibility is to use better common line approaches that instead of working their way sequentially like angular reconstitution use the entire information between all common lines at once, in an attempt to find a set of rotations for all images simultaneously. Another option is to first denoise the images in order to boost the SNR and improve the detection rate of common lines. Denoising can be achieved for example by a procedure called 2-D classification and averaging, in which images of presumably similar viewing directions are identified, rotationally aligned, and averaged, thus diminishing the noise while maintaining the common signal. While these techniques certainly help, and in many cases lead to successful *ab-initio* three-dimensional modeling (at least at low resolution), for small molecules with very low SNR they still fail.

The failure of these algorithms is not due to their lack of sophistication, but rather a fundamental one: It is impossible to accurately estimate the image rotations at very low SNR, regardless of the algorithmic procedure being used. To understand this inherent limitation, consider an oracle that knows the molecular structure  $\phi$ . Even the oracle would not be able to accurately estimate image rotations at very low SNR. In an attempt to estimate the rotations, the oracle would produce template projection images of the form  $PR \circ \phi$ , and for each noisy reference experimental image, the oracle would look for its best match among the template images, that is, the rotation  $R$  that minimizes the distance between the template  $PR \circ \phi$  and the reference image. At very low SNR, the random contribution of the noise dominates the distance, and the oracle would be often fooled to assign wrong rotations with large errors.

Since at very low SNR even an oracle cannot assign rotations reliably, we should give up on any hope for a sophisticated algorithm that would succeed in estimating the rotations at any SNR. Instead, we should mainly focus on algorithms that try to estimate the structure  $\phi$  without estimating rotations. This would be the topic of the next section. Still, because in practice algorithms for estimating rotations are quite useful for large size molecules, we would quickly survey those first.

**4.1 Common-line approaches.** There are several procedures that attempt to simultaneously estimate all rotations  $R_1, \dots, R_n$  from the common lines between all pairs of images at once [Singer, Coifman, F. J. Sigworth, Chester, and Shkolnisky \[2010\]](#), [Singer and Shkolnisky \[2011\]](#), [Shkolnisky and Singer \[2012\]](#), and [Wang, Singer, and Wen \[2013\]](#). Due to space limitations, we only briefly explain the semidefinite programming (SDP) relaxation approach [Singer and Shkolnisky \[2011\]](#). Let  $(x_{ij}, y_{ij})$  be a point on the unit circle indicating the location of the common line between images  $I_i$  and  $I_j$  in the local coordinate system of image  $I_i$  (see [Figure 3](#), left panel). Also, let  $c_{ij} = (x_{ij}, y_{ij}, 0)^T$ . Then, the common-line property implies that  $R_i c_{ij} = R_j c_{ji}$ . Such a linear equation can be written for every pair of images, resulting an overdetermined system, because the number of equations is  $O(n^2)$ , whereas the number of variables associated with the unknown rotations is only  $O(n)$ . The least squares estimator is the solution to minimization problem

$$(4) \quad \min_{R_1, R_2, \dots, R_n \in SO(3)} \sum_{i \neq j} \|R_i c_{ij} - R_j c_{ji}\|^2.$$

This is a non-convex optimization problem over an exponentially large search space. The SDP relaxation and its rounding procedure are similar in spirit to the Goemans-Williamson SDP approximation algorithm for Max-Cut [Goemans and Williamson \[1995\]](#). Specifically, it consists of optimizing over a set of positive definite matrices with entries related to the rotation ratios  $R_i^T R_j$  and satisfying the block diagonal constraints  $R_i^T R_i = I$ , while relaxing the rank-3 constraint. There is also a spectral relaxation variant, which is much more efficient to compute than SDP and its performance can be quantified using representation theory [Hadani and Singer \[2011\]](#), but requires the distribution of the viewing directions to be uniform.

A more recent procedure [A. S. Bandeira, Y. Chen, and Singer \[2015\]](#) attempts to solve an optimization problem of the form

$$(5) \quad \min_{R_1, R_2, \dots, R_n \in SO(3)} \sum_{i \neq j} f_{ij}(R_i^T R_j)$$

using an SDP relaxation that generalizes an SDP-based algorithm for unique games [Charikar, K. Makarychev, and Y. Makarychev \[2006\]](#) to  $SO(3)$  via classical representation theory.

The functions  $f_{ij}$  encode the cost for the common line implied by the rotation ratio  $R_i^T R_j$  for images  $I_i$  and  $I_j$ . The unique feature of this approach is that the common lines do not need to be identified, but rather all possibilities are taken into account and weighted according to the pre-computed functions  $f_{ij}$ .

**4.2 2-D classification and averaging.** If images corresponding to similar viewing direction can be identified, then they can be rotationally (and translationally) aligned and averaged to produce “2-D class averages” that enjoy a higher SNR. The 2-D class averages can be used as input to common-line based approaches for rotation assignment, as templates in semi-automatic procedures for particle picking, and to provide a quick assessment of the particles.

There are several computational challenges associated with the 2-D classification problem. First, due to the low SNR, it is difficult to detect neighboring images in terms of their viewing directions. It is also not obvious what metric should be used to compare images. Another difficulty is associated with the computational complexity of comparing all pairs of images and finding their optimal in-plane alignment, especially for large datasets consisting of hundreds of thousands of particle images.

Principal component analysis (PCA) of the images offers an efficient way to reduce the dimensionality of the images and is often used in 2-D classification procedures [Van Heel and Frank \[1981\]](#). Since particle images are just as likely to appear in any in-plane rotation (e.g., by rotating the detector), it makes sense to perform PCA for all images and their uniformly distributed in-plane rotations. The resulting covariance matrix commutes with the group action of in-plane rotation. Therefore, it is block-diagonal in any steerable basis of functions in the form of outer products of radial functions and Fourier angular modes. The resulting procedure, called steerable PCA is therefore more efficiently computed compared to standard PCA [Zhao, Shkolnisky, and Singer \[2016\]](#). In addition, the block diagonal structure implies a considerable reduction in dimensionality: for images of size  $L \times L$ , the largest block size is  $O(L \times L)$ , whereas the original covariance is of size  $L^2 \times L^2$ . Using results from the spiked covariance model in high dimensional PCA [Johnstone \[2001\]](#), this implies that the principal components and their eigenvalues are better estimated using steerable PCA, and modern eigenvalue shrinkage procedures can be applied with great success [Bhamre, Zhang, and Singer \[2016\]](#).

The steerable PCA framework also paves the way to a natural rotational invariant representation of the image [Zhao and Singer \[2014\]](#). Images can therefore be compared using their rotational invariant representation, saving the cost associated with rotational alignment. In addition, efficient algorithms for approximate nearest neighbors search can be applied for initial classification of the images. The classification can be further improved

by applying vector diffusion maps [Singer and H.-T. Wu \[2012\]](#) and [Singer, Zhao, Shkolnisky, and Hadani \[2011\]](#), a non-linear dimensionality reduction method that generalizes Laplacian eigenmaps [Belkin and Niyogi \[2002\]](#) and diffusion maps [Coifman and Lafon \[2006\]](#) by also exploiting the optimal in-plane transformation between neighboring images.

## 5 How to solve the cryo-EM problem at very low SNR?

The most popular approach for cryo-EM reconstruction is iterative refinement. Iterative refinement methods date back to (at least) [Harauz and Ottensmeyer \[1983, 1984\]](#) and are the cornerstone of modern software packages for single particle analysis [Shaikh, Gao, Baxter, Asturias, Boisset, Leith, and Frank \[2008\]](#), [van Heel, Harauz, Orlova, R. Schmidt, and Schatz \[1996\]](#), [C. Sorzano, Marabini, Velázquez-Muriel, Bilbao-Castro, S. H. Scheres, Carazo, and Pascual-Montano \[2004\]](#), [Tang, Peng, Baldwin, Mann, Jiang, Rees, and Ludtke \[2007\]](#), [Hohn et al. \[2007\]](#), [Grigorieff \[2007\]](#), [S. Scheres \[2012\]](#), and [Punjani, Rubinstein, Fleet, and Brubaker \[2017\]](#). Iterative refinement starts with some initial 3-D structure  $\phi_0$  and at each iteration project the current structure at many different viewing directions to produce template images, then match the noisy reference images with the template images in order to assign rotations to the noisy images, and finally perform a 3-D tomographic reconstruction using the noisy images and their assigned rotations. Instead of hard assignment of rotations, a soft assignment in which each rotation is assigned a distribution rather than just the best match, can be interpreted as an expectation-maximization procedure for maximum likelihood of the structure  $\phi$  while marginalizing over the rotations, which are treated as nuisance parameters. The maximum likelihood framework was introduced to the cryo-EM field by [F. Sigworth \[1998\]](#) and its implementation in the RELION software package [S. Scheres \[2012\]](#) is perhaps most widely used nowadays. Notice that a requirement for the maximum likelihood estimator (MLE) to be consistent is that the number of parameters to be estimated does not grow indefinitely with the number of samples (i.e., number of images in our case). The Neyman-Scott “paradox” [Neyman and Scott \[1948\]](#) is an example where maximum likelihood is inconsistent when the number of parameters grows with the sample size. The MLE of  $\phi$  and  $R_1, \dots, R_n$  is therefore not guaranteed to be consistent. On the other hand, the MLE of  $\phi$  when treating the rotations as hidden parameters is consistent.

The MLE approach has been proven very successful in practice. Yet, it suffers from several important shortcomings. First, expectation-maximization and other existing optimization procedures are only guaranteed to converge to a local optimum, not necessary the global one. Stochastic gradient descent [Punjani, Rubinstein, Fleet, and Brubaker \[2017\]](#)

and frequency marching [Barnett, Greengard, Pataki, and Spivak \[2017\]](#) attempt to mitigate that problem. MLE requires an initial starting model, and convergence may depend on that model, a phenomenon known as “model bias”. MLE can be quite slow to compute, as many iterations may be required for convergence, with each iteration performing a computationally expensive projection template-reference matching and tomographic reconstruction, although running times are significantly reduced in modern GPU implementations. From a mathematical standpoint, it is difficult to analyze the MLE. In particular, what is the sample complexity of the cryo-EM reconstruction problem? That is, how many noisy images are needed for successful reconstruction?

## 6 Kam’s autocorrelation analysis

About 40 years ago, [Kam \[1980\]](#) proposed a method for 3-D *ab-initio* reconstruction which is based on computing the autocorrelation and higher order correlation functions of the 3-D structure in Fourier space from the 2-D noisy projection images. Remarkably, it was recently shown in [A. S. Bandeira, Blum-Smith, Perry, Weed, and Wein \[2017\]](#) that these correlation functions determine the 3-D structure uniquely (or at least up to a finite number of possibilities). Kam’s method completely bypasses the estimation of particle rotations and estimates the 3-D structure directly. The most striking advantage of Kam’s method over iterative refinement methods is that it requires only one pass over the data for computing the correlation functions, and as a result it is extremely fast and can operate in a streaming mode in which data is processed on the fly while being acquired. Kam’s method can be regarded as a method of moments approach for estimating the structure  $\phi$ . The MLE is asymptotically efficient, therefore its mean squared error is typically smaller than that of the method of moments estimator. However, for the cryo-EM reconstruction problem the method of moments estimator of Kam is much faster to compute compared to the MLE. In addition, Kam’s method does not require a starting model. From a theoretical standpoint, Kam’s theory sheds light on the sample complexity of the problem as a function of the SNR. For example, using Kam’s method in conjunction with tools from algebraic geometry and information theory, it was shown that in the case of uniformly distributed rotations, the sample complexity scales as  $1/\text{SNR}^3$  in the low SNR regime [A. S. Bandeira, Blum-Smith, Perry, Weed, and Wein \[ibid.\]](#).

Interest in Kam’s theory has been recently revived due to its potential application to X-ray free electron lasers (XFEL) [Kam \[1977\]](#), [Liu, B. K. Poon, Saldin, Spence, and Zwart \[2013\]](#), [Starodub et al. \[2012\]](#), [Saldin, Shneerson, et al. \[2010\]](#), [Saldin, H.-C. Poon, Schwander, Uddin, and M. Schmidt \[2011\]](#), and [Kurta et al. \[2017\]](#). However, Kam’s method has so far received little attention in the EM community. It is an idea that was clearly ahead of its time: There was simply not enough data to accurately estimate second

and third order statistics from the small datasets that were available at the time (e.g. typically just dozens of particles). Moreover, accurate estimation of such statistics requires modern techniques from high dimensional statistical analysis such as eigenvalue shrinkage in the spiked covariance model that have only been introduced in the past two decades. Estimation is also challenging due to the varying CTF between micrographs and non-perfect centering of the images. Finally, Kam's method requires a uniform distribution of particle orientations in the sample, an assumption that usually does not hold in practice.

In [Bhamre, Zhang, and Singer \[2016\]](#), we have already addressed the challenge of varying CTF and also improved the accuracy and efficiency of estimating the covariance matrix from projection images by combining the steerable PCA framework [Zhao and Singer \[2013\]](#) and [Zhao, Shkolnisky, and Singer \[2016\]](#) with optimal eigenvalue shrinkage procedures [Johnstone \[2001\]](#), [Donoho, Gavish, and Johnstone \[2013\]](#), and [Gavish and Donoho \[2017\]](#). Despite this progress, the challenges of non-perfect centering of the images that limits the resolution and the stringent requirement for uniformly distributed viewing directions, still put severe limitations on the applicability of Kam's method in cryo-EM.

Here is a very brief account of Kam's theory. Kam showed that the Fourier projection slice theorem implies that if the viewing directions of the projection images are uniformly distributed, then the autocorrelation function of the 3-D volume with itself over the rotation group  $SO(3)$  can be directly computed from the covariance matrix of the 2-D images, i.e. through PCA. Specifically, consider the spherical harmonics expansion of the Fourier transform of  $\phi$

$$(6) \quad \mathfrak{F}\phi(k, \theta, \varphi) = \sum_{l=0}^{\infty} \sum_{m=-l}^l A_{lm}(k) Y_l^m(\theta, \varphi),$$

where  $Y_l^m$  are the spherical harmonics, and  $A_{lm}$  are functions of the radial frequency  $k$ . Kam showed that from the covariance matrix of the 2-D Fourier transform of the 2-D projection images it is possible to extract matrices  $C_l$  ( $l = 0, 1, 2, \dots$ ) that are related to the radial functions  $A_{lm}$  through

$$(7) \quad C_l(k_1, k_2) = \sum_{m=-l}^l A_{lm}(k_1) \overline{A_{lm}(k_2)}.$$

For images sampled on a Cartesian grid of pixels, each  $C_l$  is a matrix of size  $K_l \times K_l$ , where  $K_l$  is determined by a sampling criterion dating back to [Klug and Crowther \[1972\]](#) to avoid aliasing.  $K_l$  is a monotonic decreasing function of  $l$ , and we set  $L$  as the largest  $l$  in the spherical harmonics expansion for which  $K_l \geq l$ . In matrix notation, [Equation \(7\)](#) is equivalent to

$$(8) \quad C_l = A_l A_l^*,$$



where  $A_l$  is a matrix of size  $K_l \times (2l + 1)$  whose  $m$ 'th column is the vector  $A_{lm}$  and whose rows are indexed by the radial frequency  $k$ , and where  $A^*$  is the Hermitian conjugate of  $A$ . However, the factorization of  $C_l$  in Equation (8), also known as the Cholesky decomposition, is not unique: If  $A_l$  satisfies Equation (8), then for any  $(2l + 1) \times (2l + 1)$  unitary matrix  $U$  (i.e.,  $U$  satisfies  $UU^* = U^*U = I_{2l+1}$ ), also  $A_l U$  satisfies Equation (8). In fact, since the molecular density  $\phi$  is real-valued, its Fourier transform is conjugate-symmetric, and hence the matrices  $A_l$  are purely real for even  $l$ , and purely imaginary for odd  $l$ . Therefore, Equation (8) determines  $A_l$  uniquely up to an orthogonal matrix  $O_l$  of size  $(2l + 1) \times (2l + 1)$  (i.e.,  $O_l$  is a real valued matrix satisfying  $O_l O_l^T = O_l^T O_l = I_{2l+1}$ ). Formally, we take a Cholesky decomposition of the estimated  $C_l$  to obtain a  $K_l \times (2l + 1)$  matrix  $F_l$  satisfying  $C_l = F_l F_l^*$ . Accordingly,  $A_l = F_l O_l$  for some unknown orthogonal matrix  $O_l$ .

In other words, from the covariance matrix of the 2-D projection images we can retrieve, for each  $l$ , the radial functions  $A_{lm}$  ( $m = -l, \dots, l$ ) up to an orthogonal matrix  $O_l$ . This serves as a considerable reduction of the parameter space: Originally, a complete specification of the structure requires, for each  $l$ , a matrix  $A_l$  of size  $K_l \times (2l + 1)$ , but the additional knowledge of  $C_l$  reduces the parameter space to that of an orthogonal matrix of size  $(2l + 1) \times (2l + 1)$  which has only  $l(2l + 1)$  degrees of freedom, and typically  $K_l \gg l$ .

In Bhamre, Zhang, and Singer [2015] we showed that the missing orthogonal matrices  $O_1, O_2, \dots, O_L$  can be retrieved by “orthogonal extension”, a process that relies on the existence of a previously solved similar structure and in which the orthogonal matrices are grafted from the previously resolved similar structure to the unknown structure. However, the structure of a similar molecule is usually unavailable. We also offered another method for retrieving the orthogonal matrices using “orthogonal replacement”, inspired by molecular replacement in X-ray crystallography. While orthogonal replacement does not require any knowledge of a similar structure, it assumes knowledge of a structure that can bind to the molecule (e.g., an antibody fragment of known structure that binds to a protein).

An alternative approach for determining the orthogonal matrices was already proposed by Kam [1980] and Kam and Gafni [1985], who suggested using higher order correlations. Specifically, Kam proposed using triple products of the form  $\hat{I}^2(k_1)\hat{I}(k_2)$  and quadruple products of the form  $\hat{I}^2(k_1)\hat{I}^2(k_2)$ , where  $\hat{I}$  is the 2-D Fourier transform of image  $I$ . The main disadvantage of using higher order correlations is noise amplification: Methods based on triple correlations require number of images that scale as  $1/\text{SNR}^3$ , and even more badly as  $1/\text{SNR}^4$  in the case of quadruple correlation. The higher correlation terms that Kam proposed are not complete. In general, a triple product takes the form

$\hat{I}(k_1)\overline{\hat{I}(k_2)}\hat{I}(k_3)$ . Kam is using only a slice of the possible triple products (namely, setting  $k_3 = k_1$ ) due to the large number of coefficients it results in. This is closely related to restricting the bispectrum due to its high dimensionality [Marabini and Carazo \[1996\]](#). In that respect we note that a vast reduction in the dimensionality of the triple correlation (or bispectrum coefficients) can be achieved by only using triple products of the steerable PCA coefficients. The number of meaningful PCA expansion coefficient is typically of the order of a few hundreds (depending on the noise level), much smaller than the number of pixels in the images.

## 7 A mathematical toy model: multi-reference alignment

The problem of multi-reference alignment serves as a mathematical toy model for analyzing the cryo-EM reconstruction and heterogeneity problems. In the multi-reference alignment model, a signal is observed by the action of a random circular translation and the addition of Gaussian noise. The goal is to recover the signal's orbit by accessing multiple independent observations ([Figure 5](#)). Specifically, the measurement model is of the form

$$(9) \quad y_i = R_i x + \varepsilon_i, \quad x, y_i, \varepsilon_i \in \mathbb{R}^L, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2 I_{L \times L}), \quad i = 1, 2, \dots, n.$$

While pairwise alignment succeeds at high SNR, accurate estimation of rotations is impossible at low SNR, similar to the fundamental limitation in cryo-EM. Two natural questions arise: First, how to estimate the underlying signal at very low SNR, and how many measurements are required for accurate estimation.

Just like in the cryo-EM reconstruction problem, an expectation-maximization type algorithm can be used to compute the MLE of the signal  $x$ , treating the cyclic shifts as nuisance parameters. Alternatively, a method of moments approach would consist of estimating correlation functions that are invariant to the group action. Specifically, the following are invariant features (in Fourier / real space), and the number of observations needed for accurate estimation by the central limit theorem:

- Zero frequency / average pixel value:

$$(10) \quad \frac{1}{n} \sum_{i=1}^n \hat{y}_i(0) \rightarrow \hat{x}(0) \quad \text{as } n \rightarrow \infty. \quad \text{Need } n \gtrsim \sigma^2.$$

- Power spectrum / autocorrelation:

$$(11) \quad \frac{1}{n} \sum_{i=1}^n |\hat{y}_i(k)|^2 \rightarrow |\hat{x}(k)|^2 + \sigma^2 \quad \text{as } n \rightarrow \infty. \quad \text{Need } n \gtrsim \sigma^4.$$

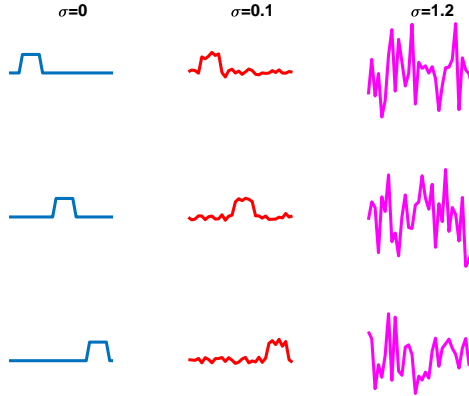


Figure 5: Multi-reference alignment of 1-D periodic signals, at different noise levels  $\sigma$ .

- Bispectrum / triple correlation [Tukey \[1953\]](#):

$$(12) \quad \frac{1}{n} \sum_{i=1}^n \hat{y}_i(k_1) \hat{y}_i(k_2) \hat{y}_i(-k_1 - k_2) \rightarrow \hat{x}(k_1) \hat{x}(k_2) \hat{x}(-k_1 - k_2)$$

as  $n \rightarrow \infty$  Need  $n \gtrsim \sigma^6$ .

The bispectrum  $B_X(k_1, k_2) = \hat{x}(k_1) \hat{x}(k_2) \hat{x}(-k_1 - k_2)$  contains phase information and is generically invertible (up to global shift) [Kakarala \[1993\]](#) and [Sadler and Giannakis \[1992\]](#). It is therefore possible to accurately reconstruct the signal from sufficiently many noisy shifted copies for arbitrarily low SNR without estimating the shifts and even when estimation of shifts is poor. Notice that if shifts are known, then  $n \gtrsim 1/\text{SNR}$  is sufficient for accurate estimation of the signal. However, not knowing the shifts make a big difference in terms of the sample complexity, and  $n \gtrsim 1/\text{SNR}^3$  for the shift-invariant method. In fact, no method can succeed with asymptotically fewer measurements (as a function of the SNR) in the case of uniform distribution of shifts [Perry, Weed, A. Bandeira, Rigollet, and Singer \[2017\]](#), [A. Bandeira, Rigollet, and Weed \[2017\]](#), and [Abbe, J. M. Pereira, and Singer \[2017\]](#). The computational complexity and stability of a variety of bispectrum inversion algorithms was studied in [Bendory, Boumal, Ma, Zhao, and Singer \[2017\]](#) and [H. Chen, Zehni, and Zhao \[2018\]](#). A somewhat surprising result is that multi-reference alignment with non-uniform (more precisely, non-periodic) distribution of shifts can be solved with just the first two moments and the sample complexity is proportional to  $1/\text{SNR}^2$  [Abbe, Bendory, Leeb, J. Pereira, Sharon, and Singer \[2017\]](#) and [Abbe, J. M. Pereira, and Singer \[2018\]](#). The method of moments can also be applied

to multi-reference alignment in the heterogeneous setup, and avoids both shift estimation and clustering of the measurements [Boumal, Bendory, Lederman, and Singer \[2017\]](#).

The analysis of the multi-reference alignment model provides key theoretical insights into Kam's method for cryo-EM reconstruction. In addition, the multi-reference alignment problem also offers a test bed for optimization algorithms and computational tools before their application to the more challenging problems of cryo-EM.

## 8 Summary

Computational tools are a vital component of the cryo-EM structure determination process that follows data collection. Still, there are many computational aspects that are either unresolved or that require further research and development. New computational challenges constantly emerge from attempts to further push cryo-EM technology towards higher resolution, higher throughput, smaller molecules, and highly flexible molecules. Important computational challenges include mapping conformational landscapes, structure validation, dealing with low SNR for small molecule reconstruction, motion correction and video processing, ab-initio modeling, and sub-tomogram averaging, among others. We emphasize that this paper is of limited scope, and therefore addressed only a few core elements of the reconstruction pipeline, mainly focusing on the cryo-EM reconstruction problem.

Moreover, the paper did not aim to present any new algorithms and techniques, but instead provide a review of some of the already existing methods and their analysis, with perhaps some new commentary. Although the heterogeneity problem is arguably one of the most important challenges in cryo-EM analysis nowadays, techniques for addressing this problem were not discussed here mainly for space limitations. Another reason to defer the review of methods for the heterogeneity problem is that techniques are still being developed, and that aspect of the cryo-EM analysis is less mature and not as well understood compared to the basic cryo-EM reconstruction problem.

To conclude, mathematics plays a significant role in the design and analysis of algorithms for cryo-EM. Different aspects of representation theory, tomography and integral geometry, high dimensional statistics, random matrix theory, information theory, algebraic geometry, signal and image processing, dimensionality reduction, manifold learning, numerical linear algebra, and fast algorithms, all come together in helping structural biologists discover new biology using cryo-electron microscopy.

## References

- Emmanuel Abbe, Tamir Bendory, William Leeb, João Pereira, Nir Sharon, and Amit Singer (2017). “[Multireference Alignment is Easier with an Aperiodic Translation Distribution](#)”. arXiv: [1710.02793](#) (cit. on p. [4028](#)).
- Emmanuel Abbe, João M Pereira, and Amit Singer (2017). “Sample complexity of the boolean multireference alignment problem”. In: *Information Theory (ISIT), 2017 IEEE International Symposium on*. IEEE, pp. 1316–1320 (cit. on p. [4028](#)).
- (2018). “[Estimation in the group action channel](#)”. arXiv: [1801.04366](#) (cit. on p. [4028](#)).
- Alexey Amunts et al. (2014). “Structure of the yeast mitochondrial large ribosomal sub-unit”. *Science* 343.6178, pp. 1485–1489 (cit. on p. [4013](#)).
- Afonso S Bandeira, Ben Blum-Smith, Amelia Perry, Jonathan Weed, and Alexander S Wein (2017). “[Estimation under group actions: recovering orbits from invariants](#)”. arXiv: [1712.10163](#) (cit. on p. [4024](#)).
- Afonso S Bandeira, Yutong Chen, and Amit Singer (2015). “[Non-unique games over compact groups and orientation estimation in cryo-em](#)”. arXiv: [1505.03840](#) (cit. on p. [4021](#)).
- Afonso Bandeira, Philippe Rigollet, and Jonathan Weed (2017). “[Optimal rates of estimation for multi-reference alignment](#)”. arXiv: [1702.08546](#) (cit. on p. [4028](#)).
- Alex Barnett, Leslie Greengard, Andras Pataki, and Marina Spivak (2017). “[Rapid Solution of the Cryo-EM Reconstruction Problem by Frequency Marching](#)”. *SIAM Journal on Imaging Sciences* 10.3, pp. 1170–1195 (cit. on p. [4024](#)).
- Alberto Bartesaghi, Alan Merk, Soojay Banerjee, Doreen Matthies, Xiongwu Wu, Jacqueline LS Milne, and Sriram Subramaniam (2015). “2.2Å resolution cryo-EM structure of  $\beta$ -galactosidase in complex with a cell-permeant inhibitor”. *Science* 348.6239, pp. 1147–1151 (cit. on p. [4013](#)).
- Samit Basu and Yoram Bresler (2000a). “Feasibility of tomography with unknown view angles”. *IEEE Transactions on Image Processing* 9.6, pp. 1107–1122 (cit. on p. [4019](#)).
- (2000b). “Uniqueness of tomography with unknown view angles”. *IEEE Transactions on Image Processing* 9.6, pp. 1094–1106 (cit. on p. [4019](#)).
- Mikhail Belkin and Partha Niyogi (2002). “Laplacian eigenmaps and spectral techniques for embedding and clustering”. In: *Advances in neural information processing systems*, pp. 585–591 (cit. on p. [4023](#)).
- Tamir Bendory, Nicolas Boumal, Chao Ma, Zhizhen Zhao, and Amit Singer (2017). “Bispectrum inversion with application to multireference alignment”. *IEEE Transactions on Signal Processing* 66.4, pp. 1037–1050 (cit. on p. [4028](#)).
- Tejal Bhamre, Teng Zhang, and Amit Singer (2015). “Orthogonal matrix retrieval in cryo-electron microscopy”. In: *Biomedical Imaging (ISBI), 2015 IEEE 12th International Symposium on*. IEEE, pp. 1048–1052 (cit. on p. [4026](#)).

- (2016). “Denoising and covariance estimation of single particle cryo-EM images”. *Journal of structural biology* 195.1, pp. 72–81 (cit. on p. [4022](#), [4025](#)).
- Nicolas Boumal, Tamir Bendory, Roy R Lederman, and Amit Singer (2017). “[Heterogeneous multireference alignment: a single pass approach](#)”. arXiv: [1710.02590](#) (cit. on p. [4029](#)).
- Moses Charikar, Konstantin Makarychev, and Yury Makarychev (2006). “Near-optimal algorithms for unique games”. In: *Proceedings of the thirty-eighth annual ACM symposium on Theory of computing*. ACM, pp. 205–214 (cit. on p. [4021](#)).
- Hua Chen, Mona Zehni, and Zhizhen Zhao (2018). “[A Spectral Method for Stable Bispectrum Inversion with Application to Multireference Alignment](#)”. arXiv: [1802.10493](#) (cit. on p. [4028](#)).
- Ronald R Coifman and Stéphane Lafon (2006). “Diffusion maps”. *Applied and computational harmonic analysis* 21.1, pp. 5–30 (cit. on p. [4023](#)).
- David L Donoho, Matan Gavish, and Iain M Johnstone (2013). “[Optimal shrinkage of eigenvalues in the spiked covariance model](#)”. arXiv: [1311.0851](#) (cit. on p. [4025](#)).
- J. Frank (2006). *Three-dimensional electron microscopy of macromolecular assemblies*. Academic Press (cit. on p. [4014](#)).
- Matan Gavish and David L Donoho (2017). “Optimal shrinkage of singular values”. *IEEE Transactions on Information Theory* 63.4, pp. 2137–2152 (cit. on p. [4025](#)).
- Robert M Glaeser (2016). “How good can cryo-EM become?” *Nature Methods* 13.1, pp. 28–32 (cit. on p. [4014](#)).
- Michel X Goemans and David P Williamson (1995). “Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming”. *Journal of the ACM (JACM)* 42.6, pp. 1115–1145 (cit. on p. [4021](#)).
- A. B. Goncharov (1988). “[Integral geometry and three-dimensional reconstruction of randomly oriented identical particles from their electron microphotos](#)”. *Acta Appl. Math.* 11.3, pp. 199–211 (cit. on p. [4019](#)).
- AB Goncharov (1987). “Methods of integral geometry and finding the relative orientation of identical particles arbitrarily arranged in a plane from their projections onto a straight line”. In: *Soviet Physics Doklady*. Vol. 32, p. 177 (cit. on p. [4018](#)).
- AB Goncharov and MS Gelfand (1988). “Determination of mutual orientation of identical particles from their projections by the moments method”. *Ultramicroscopy* 25.4, pp. 317–327 (cit. on p. [4019](#)).
- Nikolaus Grigorieff (2007). “FREALIGN: high-resolution refinement of single particle structures”. *Journal of structural biology* 157.1, pp. 117–125 (cit. on p. [4023](#)).
- Ronny Hadani and Amit Singer (2011). “Representation theoretic patterns in three dimensional Cryo-Electron Microscopy I: The intrinsic reconstitution algorithm”. *Annals of mathematics* 174.2, p. 1219 (cit. on p. [4021](#)).

- George Harauz and FP Ottensmeyer (1983). “Direct three-dimensional reconstruction for macromolecular complexes from electron micrographs”. *Ultramicroscopy* 12.4, pp. 309–319 (cit. on p. 4023).
- (1984). “Nucleosome reconstruction via phosphorous mapping”. *Science* 226, pp. 936–941 (cit. on p. 4023).
- Marin van Heel, Brent Gowen, et al. (2000). “Single-particle electron cryo-microscopy: towards atomic resolution”. *Quarterly reviews of biophysics* 33.4, pp. 307–369 (cit. on p. 4014).
- Marin van Heel, George Harauz, Elena V Orlova, Ralf Schmidt, and Michael Schatz (1996). “A new generation of the IMAGIC image processing system”. *Journal of structural biology* 116.1, pp. 17–24 (cit. on p. 4023).
- Michael Hohn et al. (2007). “SPARX, a new environment for Cryo-EM image processing”. *Journal of structural biology* 157.1, pp. 47–55 (cit. on p. 4023).
- Iain M Johnstone (2001). “On the distribution of the largest eigenvalue in principal components analysis”. *Annals of statistics*, pp. 295–327 (cit. on pp. 4022, 4025).
- Ramakrishna Kakarala (1993). “A group-theoretic approach to the triple correlation”. In: *Higher-Order Statistics, 1993., IEEE Signal Processing Workshop on*. IEEE, pp. 28–32 (cit. on p. 4028).
- Z. Kam (1980). “The reconstruction of structure from electron micrographs of randomly oriented particles”. *J. Theor. Biol.* 82.1, pp. 15–39 (cit. on pp. 4024, 4026).
- Z Kam and I Gafni (1985). “Three-dimensional reconstruction of the shape of human wart virus using spatial correlations”. *Ultramicroscopy* 17.3, pp. 251–262 (cit. on p. 4026).
- Zvi Kam (1977). “Determination of macromolecular structure in solution by spatial correlation of scattering fluctuations”. *Macromolecules* 10.5, pp. 927–934 (cit. on p. 4024).
- A Klug and R A Crowther (1972). “Three-dimensional image reconstruction from the viewpoint of information theory”. *Nature* 238.5365, pp. 435–440 (cit. on p. 4025).
- Werner Kühlbrandt (2014). “The Resolution Revolution”. *Science* 343.6178, pp. 1443–1444 (cit. on p. 4013).
- Ruslan P Kurta et al. (2017). “Correlations in scattered x-ray laser pulses reveal nanoscale structural features of viruses”. *Physical review letters* 119.15, p. 158102 (cit. on p. 4024).
- Maofu Liao, Erhu Cao, David Julius, and Yifan Cheng (2013). “Structure of the TRPV1 ion channel determined by electron cryo-microscopy”. *Nature* 504.7478, pp. 107–112 (cit. on p. 4013).
- Haiguang Liu, Billy K Poon, Dilano K Saldin, John CH Spence, and Peter H Zwart (2013). “Three-dimensional single-particle imaging using angular correlations from X-ray laser data”. *Acta Crystallographica Section A: Foundations of Crystallography* 69.4, pp. 365–373 (cit. on p. 4024).

- Roberto Marabini and José María Carazo (1996). “On a new computationally fast image invariant based on bispectral projections”. *Pattern recognition letters* 17.9, pp. 959–967 (cit. on p. [4027](#)).
- F. Natterer (1986). *The mathematics of computerized tomography*. Springer (cit. on p. [4016](#)).
- Jerzy Neyman and Elizabeth L Scott (1948). “Consistent estimates based on partially consistent observations”. *Econometrica: Journal of the Econometric Society*, pp. 1–32 (cit. on p. [4023](#)).
- Eva Nogales (2016). “The development of cryo-EM into a mainstream structural biology technique”. *Nature Methods* 13.1, pp. 24–27 (cit. on p. [4014](#)).
- Amelia Perry, Jonathan Weed, Afonso Bandeira, Philippe Rigollet, and Amit Singer (2017). “The sample complexity of multi-reference alignment”. arXiv: [1707.00943](#) (cit. on p. [4028](#)).
- Ali Punjani, John L Rubinstein, David J Fleet, and Marcus A Brubaker (2017). “cryo-SPARC: algorithms for rapid unsupervised cryo-EM structure determination”. *Nature Methods* 14.3, pp. 290–296 (cit. on p. [4023](#)).
- Brian M Sadler and Georgios B Giannakis (1992). “Shift-and rotation-invariant object reconstruction using the bispectrum”. *JOSA A* 9.1, pp. 57–69 (cit. on p. [4028](#)).
- D K Saldin, V L Shneerson, et al. (2010). “Structure of a single particle from scattering by many particles randomly oriented about an axis: toward structure solution without crystallization?”. *New J. Phys.* 12.3, p. 035014 (cit. on p. [4024](#)).
- D. K. Saldin, H.-C. Poon, P. Schwander, M. Uddin, and M. Schmidt (Aug. 2011). “Reconstructing an icosahedral virus from single-particle diffraction experiments”. *Opt. Express* 19.18, p. 17318 (cit. on p. [4024](#)).
- S. Scheres (2012). “RELION: implementation of a Bayesian approach to cryo-EM structure determination”. *Journal of structural biology* 180.3, pp. 519–530 (cit. on p. [4023](#)).
- Tanvir R Shaikh, Haixiao Gao, William T Baxter, Francisco J Asturias, Nicolas Boisset, Ardean Leith, and Joachim Frank (2008). “SPIDER image processing for single-particle reconstruction of biological macromolecules from electron micrographs”. *Nature protocols* 3.12, pp. 1941–1974 (cit. on p. [4023](#)).
- Yoel Shkolnisky and Amit Singer (2012). “Viewing direction estimation in cryo-EM using synchronization”. *SIAM journal on imaging sciences* 5.3, pp. 1088–1110 (cit. on p. [4021](#)).
- FJ Sigworth (1998). “A maximum-likelihood approach to single-particle image refinement”. *Journal of structural biology* 122.3, pp. 328–339 (cit. on p. [4023](#)).
- Fred J Sigworth (2016). “Principles of cryo-EM single-particle image processing”. *Micromscopy* 65.1, pp. 57–67 (cit. on p. [4014](#)).
- Amit Singer, Ronald R Coifman, Fred J Sigworth, David W Chester, and Yoel Shkolnisky (2010). “Detecting consistent common lines in cryo-EM by voting”. *Journal of structural biology* 169.3, pp. 312–322 (cit. on p. [4021](#)).



- Amit Singer and Yoel Shkolnisky (2011). “Three-dimensional structure determination from common lines in cryo-EM by eigenvectors and semidefinite programming”. *SIAM journal on imaging sciences* 4.2, pp. 543–572 (cit. on p. 4021).
- Amit Singer and H-T Wu (2012). “Vector diffusion maps and the connection Laplacian”. *Communications on pure and applied mathematics* 65.8, pp. 1067–1144 (cit. on p. 4023).
- Amit Singer, Zhizhen Zhao, Yoel Shkolnisky, and Ronny Hadani (2011). “Viewing angle classification of cryo-electron microscopy images using eigenvectors”. *SIAM Journal on Imaging Sciences* 4.2, pp. 723–759 (cit. on p. 4023).
- Martin TJ Smith and John L Rubinstein (2014). “Beyond blob-ology”. *Science* 345.6197, pp. 617–619 (cit. on p. 4013).
- Carlos Oscar S Sorzano and Jose Maria Carazo (2017). “Challenges ahead Electron Microscopy for Structural Biology from the Image Processing point of view”. arXiv: 1701.00326 (cit. on p. 4014).
- COS Sorzano, Roberto Marabini, Javier Velázquez-Muriel, José Román Bilbao-Castro, Sjors HW Scheres, José M Carazo, and Alberto Pascual-Montano (2004). “XMIPP: a new generation of an open-source image processing package for electron microscopy”. *Journal of structural biology* 148.2, pp. 194–204 (cit. on p. 4023).
- Dmitri Starodub et al. (2012). “Single-particle structure determination by correlations of snapshot X-ray diffraction patterns”. *Nature communications* 3, p. 1276 (cit. on p. 4024).
- Sriram Subramaniam, Werner Kühlbrandt, and Richard Henderson (2016). “CryoEM at IUCrJ: a new era”. *IUCrJ* 3.Pt 1, pp. 3–7 (cit. on p. 4014).
- Guang Tang, Liwei Peng, Philip R Baldwin, Deepinder S Mann, Wen Jiang, Ian Rees, and Steven J Ludtke (2007). “EMAN2: an extensible image processing suite for electron microscopy”. *Journal of structural biology* 157.1, pp. 38–46 (cit. on p. 4023).
- JW Tukey (1953). “The spectral representation and transformation properties of the higher moments of stationary time series”. *Reprinted in The Collected Works of John W. Tukey* 1, pp. 165–184 (cit. on p. 4028).
- B.K. Vainshtein and A.B. Goncharov (1986). “Determination of the spatial orientation of arbitrarily arranged identical particles of unknown structure from their projections”. *Soviet Physics Doklady* 31, p. 278 (cit. on p. 4018).
- M. van Heel (1987). “Angular reconstitution: A posteriori assignment of projection directions for 3D reconstruction”. *Ultramicroscopy* 21.2, pp. 111–123 (cit. on p. 4018).
- Marin Van Heel and Joachim Frank (1981). “Use of multivariate statistics in analysing the images of biological macromolecules”. *Ultramicroscopy* 6.2, pp. 187–194 (cit. on p. 4022).
- Lanhui Wang, Amit Singer, and Zaiwen Wen (2013). “Orientation determination of cryo-EM images using least unsquared deviations”. *SIAM journal on imaging sciences* 6.4, pp. 2450–2483 (cit. on p. 4021).

- Zhizhen Zhao, Yoel Shkolnisky, and Amit Singer (2016). “Fast steerable principal component analysis”. *IEEE transactions on computational imaging* 2.1, pp. 1–12 (cit. on pp. [4022](#), [4025](#)).
- Zhizhen Zhao and Amit Singer (2013). “Fourier–Bessel rotational invariant eigenimages”. *JOSA A* 30.5, pp. 871–877 (cit. on p. [4025](#)).
- (2014). “Rotationally invariant image representation for viewing direction classification in cryo-EM”. *Journal of structural biology* 186.1, pp. 153–166 (cit. on p. [4022](#)).

Received 2018-03-10.

AMIT SINGER  
DEPARTMENT OF MATHEMATICS  
PROGRAM IN APPLIED AND COMPUTATIONAL MATHEMATICS  
PRINCETON UNIVERSITY  
[amits@math.princeton.edu](mailto:amits@math.princeton.edu)



# STUDY AND RESEARCH PATHS: A MODEL FOR INQUIRY

MARIANNA BOSCH CASABÒ

## Abstract

This paper presents a line of research in didactics of mathematics developed during the past decade within the Anthropological Theory of the Didactic around what we call *study and research paths* (SRPs). SRPs are initially proposed as a study format based on the inquiry of open questions, which can be implemented at all educational levels, from pre-school to university, including teacher education and professional development. Additionally, they provide a general schema for analysing any kind of teaching and learning process, by especially pointing out the more or less explicit questions that lead the study process and the way new knowledge is built or introduced to elaborate answers to these questions. Current research on SRPs focuses on their *didactic ecology*, defined as the set of conditions required to generally implement SRPs at different educational levels, together with the constraints that hinder their development and dissemination.

## 1 Delimiting a unit of analysis

**1.1 The anthropological theory of the didactic.** Mathematics education—or *didactics of mathematics*, as we prefer to call it in many countries—is still a young field of research and comprises different approaches that do not always share their main assumptions or goals. The research here presented corresponds to the Anthropological Theory of the Didactic (ATD), a framework whose main creator, Yves Chevallard, received the ICMI Hans Freudenthal Medal in 2009 in recognition of the foundation and development of “[a very original, fruitful and influential research programme in mathematics education](#)”. To begin with, I will briefly explain how the ATD defines and delimits the object of study of didactics and the type of research questions that are primarily raised.

From the perspective of the ATD, the aim of didactics as a science is to elucidate the mechanisms by which, in a given society, knowledge is diffused within institutions and among persons. The conception of knowledge adopted is very broad. It embraces what

is usually considered as knowledge as such, in the sense of theoretical elaborations or constructions or, according to the dictionary, the “sum of what is known” in a given domain or discipline, close to what the French mathematician Georges Bouligand (1889-1979) defined as *syntheses*, which “keep track of new problems and assembles results known to coordinate an inventory of methods and operations” [Bouligand \[1957, p. 139\]](#). And it also includes the practical dimension of knowledge, the know-how that supports all kind of human activities.

Knowledge, in both the theoretical and practical sense, is modelled in the ATD through the notion of *praxeology*. The term is formed by a combination of *praxis*—the know-how or ways of doing—and *logos*—an organised discourse about the praxis. The praxis and the logos blocks of a praxeology are in turn made up of two distinct elements: *types of tasks* and *techniques* to carry them out, for the praxis; a *technology* or discourse about the technique, and a *theory* or justification of the technology, for the logos.

One of the main postulates of the ATD is that any kind of human activity, as well as the knowledge (in the broad sense) derived from it, can be described in terms of praxeologies. Therefore, group theory or complex analysis are praxeologies, made up of elaborated theoretical discourses that describe, justify and structure a wide array of problems and techniques. However, there also exist more humble praxeologies that are activated, for instance, when we wash dishes, ride a bike or give a lecture. Many of the praxeologies people enact are difficult to describe: they consist of informal techniques that do not always have a name and include poorly organised descriptions and justifications, based on implicit assumptions and concepts. The situation is a little different in the case of sciences or academic disciplines, since a great collective effort is regularly made to make them explicit, by describing the methods used, especially to test them and share them with the community; by specifying their main assumptions and organising them coherently; by defining and structuring the notions that constitute these assumptions, the results gathered and the methods used to produce them, that will soon become new assumptions to put to use.

The need to disseminate praxeologies clearly contributes to developing them, by enriching their description (logos) and by assembling amalgams of praxeologies to build new better organised bodies of knowledge. It also helps developing their praxis to adapt it to new situations and new users. Didactics as a research field is mainly concerned by the study of “the conditions and constraints under which praxeologies start to live, migrate, change, operate, perish, disappear, be reborn, etc. within human groups” [Chevallard \[2007\]](#).

The dissemination of praxeologies takes place through what we call *didactic systems*. A didactic system is a tern  $S(X, Y, \wp)$  which is formed any time a person or a group of persons  $Y$  (the teachers) does or do something to help a group of persons  $X$  (the students) to learn a given body of knowledge or praxeology  $\wp$ .  $X$  and  $Y$  can be reduced to single persons  $x$  and  $y$ , which can also coincide, thus forming an auto-didactic system  $S(x, x, \wp)$ .

The questioning about the delimitation, composition and origin of  $\wp$ —the knowledge or praxeology to be studied—is a core problem in didactics, and it leads to what we call the *epistemological or praxeological dimension* of the problem. The *dynamic* of didactic systems—what  $X$  and  $Y$  do to make it evolve—and the *conditions and constraints* that enable and hinder this dynamic are also important dimensions at the centre of the didactic questioning. They correspond to what we call the *economy* and the *ecology* of didactic systems.

**1.2 The scale of levels of didactic codeterminacy.** Didactic systems do not exist in a vacuum. In order to facilitate the analysis of their ecology, we consider a *scale of levels of didactic codeterminacy* Chevallard [2002]. The higher levels of the scale correspond to the conditions and constraints related to the general way of organising teaching and learning processes (Fig. 1). The level of *pedagogies* comprises everything  $X$  and  $Y$  do for the didactic system to run that does not depend on the particular praxeology  $\wp$  at play. For instance, many of the instructional formats that are usually proposed to improve university teaching practices (for instance, “interactive lectures”, “cooperative learning”, “discovery learning”, “participative tutorials”, etc.) are defined independently of the precise content that is to be taught and learn and can thus be located at the pedagogical level. Their specification to a given content is then left under the teachers’ own responsibility, even if it is not always a trivial affair...

The level of *schools* includes all the infrastructures provided by educational institutions to organise didactic systems and help them run: organisations of groups of teachers and students, structures in courses and modules, physical and virtual spaces, time schedules, final exam obligations, access to knowledge resources and experts, accreditations, etc. Depending on the school systems and traditions, some pedagogical resources will be easier to develop than others and, therefore, some types of praxeologies will be easier to disseminate than others.

The levels situated at the higher end of the scale include the way teaching and learning processes are conceived and managed in *societies* or, when shared by different societies,

in *civilisations*. The scale ends at the most general level, the level of *humanity* (Fig. 1).

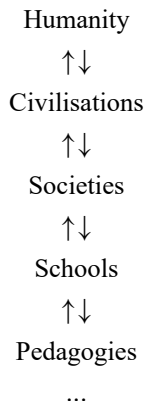


Figure 1. Scale of levels of didactic codeterminacy - Higher levels

In our societies, teaching and learning processes are mainly organised according to what has been called the *paradigm of visiting works* [Chevallard \[2015\]](#). In this paradigm, instructional processes are determined by the selection of a set of works or praxeological organisations—a curriculum—that students are asked to “visit” under the guidance of the teacher. The visit includes learning what those works are made of, which their main elements are and how they can be used, for instance to solve some given sets of problems—usually called “applications”. It not only comprises becoming aware of their existence, but also acknowledging their importance as historical productions.

In this paradigm, the selection of praxeologies that form a curriculum leads to specific knowledge organisations, which can vary depending on the society, school institution and historical period considered, but remain relatively stable over long periods of time. Think, for instance, in a first year university course of Calculus or Linear Algebra, or in the teaching of equations in secondary school. The lower levels of the scale of didactic codeterminacy take these structures into account by distinguishing different “sizes” of the praxeological organisations: *disciplines, sectors, domains, themes and questions* (Fig. 2). Thus, when a didactic system is formed in a regular school setting, the question about the delimitation and composition of  $\wp$  is answered internally:  $\wp$  corresponds to this or that type of tasks, or to this or that theme, domain, sector or discipline, which, in turn, belongs to (or is composed of) these themes, domains, sectors, etc.

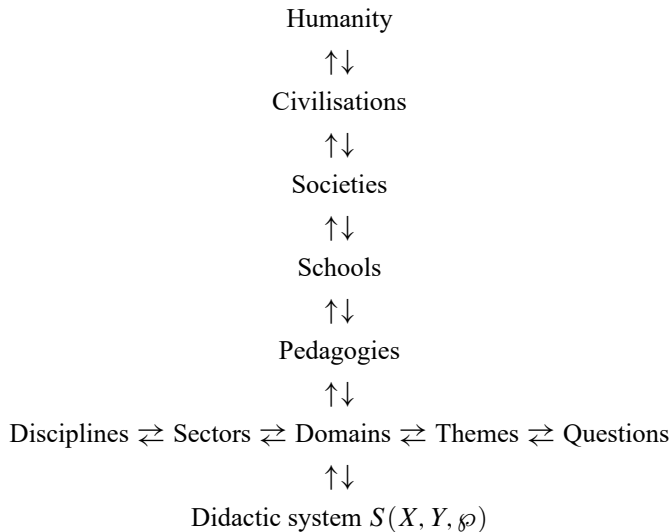


Figure 2. Scale of codeterminacy in the paradigm of visiting works

The scale of didactic codeterminacy is first of all a tool for researchers in didactics to question the reality they aim at studying. Its main utility is to enlarge our vision towards certain empirical fields that are traditionally kept outside the didacticians' perspective and are thus taken for granted. In effect, a great amount of research in didactics that focuses on the levels of the questions or the themes that are taught and learnt, rarely questions the specific structuring of disciplines, sectors or domains these questions or themes belong to. Therefore, many of the conditions and constraints that come from the lower levels remain hidden, as if they were not part of the problems addressed. Let us take a single example. Many studies about the teaching and learning of negative numbers assume that this work belongs to the school arithmetical domain, where numbers are introduced as measures of quantities—directional quantities, in this case. The possibility of introducing integers within the algebraic domain, for instance as necessary tools to give coherence to the work with equations and formulas [Cid \[2015\]](#), is rarely considered, mainly because this situation does not correspond to the current school structuring of the mathematical content. It is important for research in didactics to question the way mathematics is organised and considered at the different levels of the scale of didactic codeterminacy. At this respect, didacticians have to assume a different perspective from the one of the teachers who are asked to teach a body of knowledge that is already organised into sectors, domains and even themes. On the contrary, didactic analyses and interventions should dare approach



the higher levels of the scale, in spite of the important methodological difficulties they might encounter.

A first step in this direction is the study of what we call the *mathematical praxeological models prevailing in educational institutions* regarding the content to be taught: how is mathematics conceived, why is it structured in these sectors and domains, what is the role of such and such themes or domains in the whole organisation, which are and which could be their *raison d'être*. To avoid taking the vision proposed by the educational institution for granted, researchers in didactics had to build their own *praxeological reference models* of the mathematical works that are taught and learnt, if they want to be able to question them, as a first step towards proposing possible alternatives Bosch and Gascón [2006]. The consideration of integers as algebraic entities is a good example of these reference praxeological models: once it is specified as an alternative mathematical organisation (compared to the school ones), it can be used to point out different didactic phenomena that explain many difficulties encountered by secondary school teachers and students in relation to negative numbers—and to algebra Cid [2015].

Questioning the prevailing epistemological models in educational institutions is at the core of research in didactics of mathematics in the perspective proposed by the theory of didactic situations Brousseau [1997], which is at the origin of the ATD. The evolution of a didactic system  $S(X, Y, \wp)$  undoubtedly depends on what  $X$  and  $Y$  can do, but also on *how the praxeology  $\wp$  is delimited, conceived, considered, imported, used, legitimised* in school institutions—as well as at the other higher levels of codeterminacy. We talk about *didactic phenomena* to refer to all regular facts occurring in teaching and learning processes that are specific of the content. The frontier between *didactic* and *pedagogical* phenomena is not always clear and many teaching or learning difficulties—like the students' lack of motivation, for instance—are sometimes approached as *pedagogical* phenomena when they are of a clear *didactic* nature. Didactic research tries to overcome the frontiers between the pedagogical proposals that rarely take into account the specificity of the content to be taught and learnt, and the specific levels related to the structuring of this content. All in all, and in order to be operational, any general pedagogical change—such as the incorporation of competencies as a key tool to define educational objectives—have to be developed on close interaction with the specific praxeological organisations that are at the centre of didactic systems.

## 2 A change of paradigm: from “visiting works” to “questioning the world”

The paradigm of visiting works bases its legitimacy on the social importance given to the selected bodies of knowledge that constitute the curriculum. These are supposed to

have been chosen because of their utility in the future life of students, but this utility is more assumed than proved, since there is rarely the need in making it explicit. The main objective of the study process is defined by getting to know the selected praxeologies and being able to activate them—up to a given extent—, the *raison d'être* of these praxeologies, not only the reasons for learning them but also their reasons for existence, can remain in the shadow or simply be delayed, presented as something that will appear later on—if it does.

This does not mean that the paradigm of visiting works necessarily implies “transmissive” instructional formats, where the teacher presents or depicts a given body of knowledge for the students to acquire it and try to apply it through a given set of activities. It is also compatible with student-centred instructional formats or even with inquiry or problem-based learning. However, in these cases, the choice and role of the activities, problems or inquiries remains subordinate to the construction of the praxeologies.

This traditional way to disseminate mathematical knowledge is based on the transmission of *syntheses*—in the sense of Bouligand. The way mathematical knowledge has been selected and organised for schools, is structured as an already finished product, which includes precise terminology to describe its main notions, results and techniques. However, it leaves little room for the questions that motivated or could motivate their construction. This situation leads to what Chevallard [2015] has called the *monumentalisation of curriculum*, where each selected mathematical work appears as “a monument, a masterpiece even, that, however impudently, we are expected to revere and bow towards”. It also leads to a *sacralisation of syntheses*, that is, of the praxeological organisations elaborated to structure the bodies of knowledge in themes, sectors and domains. The monumentalisation of curriculum goes together with the unquestionability of the lower levels of the scale of didactic codeterminacy.

To avoid assuming this state of things, as researchers—but also as citizens—, the paradigm of visiting works is subsumed into a larger pedagogical paradigm, the *paradigm of questioning the world*, which can also appear as a counter-paradigm because of the important changes it requires in the scale of codeterminacy, from the lowest to the highest levels. The main element to define the paradigm of questioning the world is the notion of *study and research path* (SRP) based on the so-called *Herbartian schema*:

$$[S(X; Y; Q) \curvearrowright M] \hookrightarrow A^\heartsuit.$$

In this paradigm, the didactic system  $S(X; Y; Q)$  is not formed around a given praxeology  $\wp$  to be studied, but around a question  $Q$  to which  $X$ , with the help of  $Y$ , has to provide an answer  $A^\heartsuit$ . The study of  $Q$  generates an inquiry process involving a didactic milieu  $M$  made up of different types of objects or tools for the inquiry:

$$M = \{A_1^\diamond, A_2^\diamond, \dots, A_m^\diamond, W_{m+1}, W_{m+2}, \dots, W_n, Q_{n+1}, Q_{n+2}, \dots, Q_p, D_{p+1}, D_{p+2}, \dots, D_q\}.$$

The heart in superscript in  $A^\heartsuit$  means both that  $A^\heartsuit$  is dear to the didactic system's "heart" and will be "at the heart" of the didactic system's activity during the inquiry process: it will be the *official answer* to  $Q$  in the class  $[X, Y]$ .

The  $A_i^\diamond$  are "ready-made" answers that seem helpful to answer  $Q$  (or to answer some questions  $Q_k$  derived from  $Q$ ) that the investigators  $X$ , supervised by  $Y$ , have discovered in the institutions around them: they are institutional answers to  $Q$ , and the lozenge  $\diamond$  in superscript indicates that this answer  $A^\diamond$  is labelled or "hallmarked" by the institution that presents it as the "official" answer to  $Q$ . The  $W_j$  are works drawn upon to make sense of the  $A_i^\diamond$ , analyse and "deconstruct" them, and to build up  $A^\heartsuit$ . The  $Q_k$  are the questions induced by the study of  $Q$ , the  $A_i^\diamond$ , and the  $W_j$ , as well as the questions raised by the construction of  $A^\heartsuit$ . Finally, the  $D_l$  are sets of data of all natures gathered in the course of the inquiry.

In this schema, the "visit of works" does not disappear: in order to find the appropriate labelled answers  $A_i^\diamond$  that would turn out to be productive for the inquiry, it can be sometimes necessary to explore large domains of knowledge and requiring the help of experts guides. However, the visit in this case is motivated, not by the importance or prestige of  $A_i^\diamond$ , but only by its productivity in the construction of  $A^\heartsuit$ .

The Herbartian schema indicates the main elements of the inquiry process. Its dynamics is captured in terms of some *dialectics* that describe the production, validation and dissemination of  $A^\heartsuit$ . We will consider three of them here. The first one is the *question-answer dialectic*, which will provide a first description of the structure of the process as well as a number of milestones on the paths followed or foreseen during the inquiry. The dialectical character of the questions and answers is related to the notions of study and research: to approach a question  $Q$ , one usually searches for available answers  $A_i^\diamond$  and has to *study* them: that is, to deconstruct and reconstruct to adapt them to  $Q$ . This study generates new questions about the validity and limitations of  $A_i^\diamond$ , its adequacy to  $Q$ , the adaptations required, etc. The question-answer dialectic is the one that provides visible proof of the progress of the inquiry and contributes to what is called the *chronogenesis* of the process.

Another crucial element of the dynamics of inquiry processes is the *media-milieu dialectic*. *Media* refers to any system emitting messages. A *milieu* in didactics is a system that is supposed to be devoid of intention with respect to the question studied and to which elements of response can be "extorted" [Brousseau 1997]. To put the media-milieu dialectic into play, any message from the media has to be confronted with the milieu to test its validity and to collect critical elements providing new information. In a sense, the answers supplied by the media have to be integrated in the milieu—turning into "sure" knowledge—and the elements of the milieu have to be worked out in order to make it

send new messages—to become a media. The evolution of the milieu by the incorporation of new objects and partial answers constitutes the *mesogenesis* of the inquiry (the generation of the milieu).

The third dialectic is the one of *the individual and the collectivity*, which reminds us that the inquirers  $X$  act jointly and in cooperation with  $Y$ , while the production of the group will also depend on the capacity of each member  $x$  and  $y$  to contribute to the common project. The way responsibilities are shared in the process and how each member assumes different roles is called the *topogenesis* of the inquiry (the generation of different places or *topos* to teachers and students).

**2.1 The herbartian schema as an analytical tool.** If we take the chronogenesis, mesogenesis and topogenesis as the main dimensions of the inquiry, we can obtain an outline of the study and research process. Its first description can take the form of a tree or arborescence of derived questions raised and partial answers obtained till the elaboration of the final answer  $A^\heartsuit$  Bosch and Winsløw [2015], Hansen and Winsløw [2011], Jessen [2014], and Winsløw, Matheron, and Mercier [2013] (Fig. 3). This first model, which shows the progress of the inquiry, including its possible detours and dead-ends, can then be enriched by the description of the evolution of the milieu  $M$ . Finally, it is possible to incorporate the didactic sub-systems created and the position and responsibilities of their actors into the model obtained.

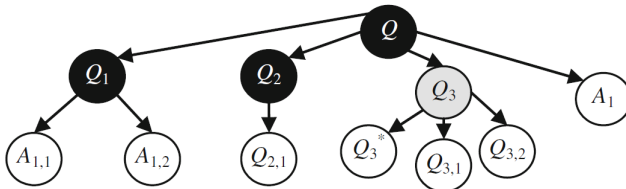


Figure 3. Example of questions-answers map Winsløw, Matheron, and Mercier [2013, p. 271]

With these elements, the Herbartian schema can be used as a tool to describe any kind of study and research process, from the most “transmissive” formats to the most “constructivist” ones; from a traditional course to a PhD research. For instance, the traditional case of a course based on lectures can be depicted with the following schema:

$$[S(X; y; \emptyset) \curvearrowright \mathcal{M}] \hookrightarrow A^\heartsuit = A_y^\diamond = \wp.$$

In this case, the question  $Q$  is usually kept in the shadow (to be discovered later on) and the answer provided by the class  $[X, y]$  corresponds to the teacher’s answer  $A_y^\diamond$  which is

supposed to reproduce a pre-established praxeology  $\wp$ . At the other extreme, there may be a purely problem-based activity in which the students are supposed to work out the answer from the sole exploitation of the available milieu, without the possibility to access any external media. In between, one can imagine several different inquiry formats, the most common one being:

$$[S(X; y; Q_{\wp}) \curvearrowright M] \hookrightarrow A^{\heartsuit} = \wp.$$

where the teacher poses a question  $Q_{\wp}$  which is supposed to lead to a previously established praxeology  $\wp$  and progressively incorporates some derived questions, data and pieces of answers into the milieu for the students to reproduce  $\wp$ .

The Herbartian schema can thus be exploited as an analytical tool, not only to describe any kind of study and research process, but also to question the choices made in its organisation with respect to many possible alternatives, by asking questions such as:

- What is  $Q_0$ ? Where does it come from? Who poses it? Why do we need to answer it? What kind of answer is required? To whom is it addressed?
- What derived questions  $Q_j$  appear during the process? Which ones fail to do so? Who poses them? How are they addressed? By whom? What kind of answers are searched for? In what types of media? How are they incorporated in the milieu  $M$ ?
- What are the main initial elements of the milieu  $M$ ? How are they used? How do they evolve?
- How is the sharing of responsibilities among teacher(s) and students established? Who does what?

Besides these basic methodological elements for the analysis of the pedagogy of the inquiry, the Herbartian schema also appears as a productive tool for the didactic design and explorations of the possibilities open by the new paradigm, as we will see in the next sections.

### 3 Implementing study and research paths

During the past decade, the Herbartian schema has been used in several research studies to design and implement various types of study and research paths (SRPs) in different educational institutions. These experimentations provide interesting empirical material to study the *economy* of SRPs—the possible ways to implement and manage them—and especially their *ecology*—the institutional conditions and constraints that facilitate or hinder their implementation. They also show possible ways of approaching certain previously

identified didactic phenomena originated in the paradigm of visiting works. A detailed account of these investigations cannot be done here, but partial reviews can be found in [Parra and Otero \[2017\]](#) and [Jessen \[2017\]](#).

**3.1 Commonalities in the design and implementation of SRPs.** The experimented SRPs were mainly implemented at secondary and tertiary levels—with an interesting exception at pre-school we are not considering here [Ruiz-Higueras and García García \[2011\]](#). They adopt different instructional formats depending on the school conditions assumed and those that were modified. There are, however, some commonalities that should be emphasised.

**3.1.1 The generating question.** The starting point of all SRPs was a generating question  $Q_0$  posed by the teacher who was in most cases also a researcher in didactics. When the teacher was not a researcher, there was always a team of researchers closely collaborating with her. In order to give predominance to  $Q_0$ —and not to the hypothetical knowledge that was supposed to be activated during the study and research process—a role-playing activity was carried out, in which the class  $(X, Y)$  acted as a consultancy team, the teachers  $Y$  assuming the role of the leaders and the students  $X$  of junior consultants.  $Q_0$  was then introduced as an assignment that came from an instance external to the class—a client—to whom an answer in the form of a report had to be handed in after a given period of time (some weeks or months). During the SRP, some interactions with the client were made possible (for instance requesting more information by e-mail), and some intermediate reports were occasionally required.

**3.1.2 Collective work.** It is important to point out that there was a unique question addressed to the whole class and, therefore, there was also only one joint report to elaborate at the end of the SRP. However, during the SRP, students could be organised in small teams  $X_i$  and different responsibilities could be assigned to each team, according to the derived questions  $Q_i$ ’ generated by the SRP. This organisation differs from many inquiry-based learning formats where students work individually on their own project under the supervision of a teacher. The way to manage the dialectic between the collective and the individual work thus appears as an important condition to take into account, and it soon revealed the lack of pedagogical resources available (at least in the mathematics class) to address it.

In most of the cases, student teams were asked to present weekly reports of the work done, including the questions addressed, the partial results obtained, the difficulties met and the new questions raised. These reports were shared with the large group in different ways, such as oral presentations, peer-reviews or simply saved in a shared web folder.

Sometimes the teacher would name a “secretary” of the class to elaborate a synthesis of the joint work carried out during the session. On some occasions, the students’ lack of experience and strategies to work in small teams, for instance to collectively perform a mathematical task, caused difficulties and some simple devices such as what [Liljedahl \[2016\]](#) calls “vertical learning” (letting student teams work in different corners of the room, using the blackboard or blank posters) turn out to be very useful.

**3.1.3 Chronogenesis: managing long study processes.** With regard to the length of the whole process, all SRPs took place during several sessions that, depending on the school constraints, varied from 50 minutes to 2 hours. This situation was completely new to the students, who are used to being asked to solve several problems within one session, especially in mathematical classrooms. They very rarely have to approach the same problematic question for weeks or months. At times, a feeling of tiredness was observed in some students—“Again the duck populations?”—that can be interpreted in two ways. On the one hand, it attests their poor engagement in the activity and their difficulty in getting rid of the traditional *didactic contract* [Brousseau \[1997\]](#) where students rarely assume the responsibility of “carrying with them” a given problem till they are finally able to solve it... On the other hand, the students’ tiredness can also be seen as a consequence of the teacher’s failure to keep the chronogenesis of the process alive. An important constraint here is the scarcity of mathematical and didactic resources available—for both the teacher and the students—to describe the inquiry process, identify the results obtained and establish milestones for the work to do. We will come back to this point later on.

**3.1.4 Topogenesis: sharing responsibilities between the teacher and the students.**

The most evident constraint that appeared in almost all SRPs was the difficulty for the teacher to share responsibilities with the students beyond those (few) assigned to them in the traditional didactic contract. At the beginning, students were easily involved in the process, but they were not used to lead the questioning, not even to raise the questions to follow during the inquiry, to select the ways that seem more promising and discard possible dead-ends. At the beginning, even us as researchers had difficulties in assuming the new contract. For instance, in the design of one of the first SRPs, we struggled with the elaboration of a realistic schedule (since we had no elements of contrast) before realising that planning the work was not necessarily the teacher’s responsibility, that students could also contribute to it. From then onwards, students were always asked to keep a logbook with a planning that was regularly updated. Again, the lack of words to describe the steps followed and those foreseen appeared as an important constraint for both the teacher and the students.

**3.1.5 Mesogenesis: the media-milieu dialectics.** The disappearing of the teacher as the main *media* and *milieu of validation* for the students (to tell them what is wrong or correct) appeared as another difficulty—especially for the teacher. Therefore, the inquiry work also required to enrich the traditional media-milieu dialectic to obtain new sources of proof and information. In what concerns the access to the media, in all SRPs, the choice of the initial question was based on a priori analyses that showed the inclusion of some empirical work available to the students and the need of new information to be searched for (in Internet, textbooks, by asking experts, etc.). Again a lack of pedagogical resources appeared: for the teacher to manage new situations requiring knowledge she may not master; for the students to be critical of the information obtained, independently of the source consulted; for both of them to create ad hoc ways to (in)validate answers  $A_i^\diamond$  when they came from “expert sources” but were not necessarily appropriate to the specific needs of the inquiry. Very elementary strategies of validation had to be established, besides the classic mathematical ones, like the comparison of different sources, the questions to experts, the rejection of useless answers, etc. On some occasions, however, it was the teacher who finally ended up introducing some key elements to let the inquiry progress.

**3.1.6 Openness and assessment.** The different levels of openness of the inquiry process are what make it at the same time exciting and disturbing for the study group. The design strategy followed by the research team consists in elaborating an *a priori questions-answers map* (Q-A map) of some expected questions derived from  $Q_0$  and possible available answers, while keeping the end of the story obviously open. This previous analysis of the generating power of  $Q_0$  ensures a minimal viability of the inquiry process and also gives the teacher a first insight of the students’ possible proposals. In many SRPs, and in spite of the normal resistance experienced by teachers to let the process advance towards unexpected—and sometimes dead-end—paths, the more freedom was given to the students, the richer the inquiry became. A possible reason is that strong guidance reveals that the initial question is not the real goal of the study, as if the means were more important than the end. This of course contributes to weaken the initial question and, therefore, frustrates the students’ efforts made to elaborate the final answer  $A^\heartsuit$ .

The importance given to the initial question should be made visible at the end of the process with the type of assessment method applied. In the experienced SRPs, the assessment strategy included part of the intermediate oral or written reports required of the students, as well as the final presentation of the answer given to  $Q_0$  or to the derived questions  $Q_i$  assigned to each team. A peer-review process among the teams of students was sometimes organised, but on most occasions the feedback was mainly given by the teacher acting as the leader of the consultant group or as the client’s representative. Panels with



external teachers and experts were organised at times, and they included oral or poster presentations depending on the number of students to be assessed.

**3.2 Questioning the sector and domain levels.** Some of the first experiences of the SRPs came as the response to previously identified didactic phenomena that were assumed to be intimately correlated to monumentalism. In the case at hand, this phenomena were also related to some key curricular contents—proportionality, algebra, functions and derivatives—and can thus be located at the level of the sector or domain in the scale of codeterminacy.

In the case of lower secondary school, a first proposal by [García \[2005\]](#) addressed the question of the isolation of proportionality from other functional relationships and the implicit preponderance given to linear growth in school mathematics. In this context, the proposed SRP started with a question that motivated the construction and comparison of different possible relationships:

How can we save money for the end-of-the-year trip, or any other trip we plan to go on in 6 to 9 months time?

Students were invited to propose different saving strategies, starting with a regular constant instalment (linear growth) and comparing it with other possible plans of increasing and decreasing instalments. It was also part of the SRP to determine the periodicity of the instalments and to make a final decision in accordance with the priorities assumed by the students and the consideration of possible unexpected events (withdrawals, newcomers, etc.). Students were led to carry out an algebraic modelling of the proposed saving plans using Excel simulations as a milieu and study the characteristics of each proposal through the adequate manipulation of equations with parameters (formulas), a work that is very unusual in secondary school mathematics. The design of the SRP was based on a previous reconstruction of the school mathematical domain of proportionality and functions to relate proportionality with other relationships between quantities and present it as a possible model among many others [García, Gascón, Ruiz Higuera, and Bosch \[2006\]](#).

A similar type of SRP was proposed by [Ruiz-Munzón \[2010\]](#) at upper secondary school to facilitate the passage from the algebraic modelling of relationships between quantities to the functional one. This SRP addressed the question of how much money a group of students can make by selling one-print T-shirts, taking into account their unit cost, the selling price and some fixed expenses due to the rent of a stand and/or a store. The data provided to the students were the sales of the previous years and they had to elaborate a plan to reach a given amount of money for an end-of-the-year trip. In this case, students had to search for information about the T-shirts production cost and possible prices, and use functional tools to solve algebraic inequalities with three or four parameters [Ruiz-Munzón, Matheron, Bosch, and Gascón \[2012\]](#). The important aspect of the work carried

out was to introduce and use functions and functional graphs to answer questions that were not initially formulated in the functional domain. Functions were not to be studied *per se* (as it is usually the case at secondary level) but because they were needed to solve inequalities that could not be solved algebraically.

In continuity with this work, [Oliveira Lucas \[2015\]](#) designed and implemented an SRP aimed at connecting functional modelling with elementary calculus. As in the previous cases, the research was based on an empirical study of the prevailing epistemological model at secondary level about elementary calculus and the elaboration of an alternative praxeological organisation to reconstruct the *raison d'être* of elementary calculus in a functional modelling context. The SRP started with some data about a Dengue epidemic and requested a forecast for the following weeks. An extensive *a priori* analysis of the generating question considering two main inquiry paths, depending on the type of data and quantities (discrete/continuous), and the variable used (the original one, its rate of change, its relative rate of change) was performed. Then, continuous models were introduced because of their technical facilities to calculate the variations of the variables considered, reversing in a way the traditional organisation of concepts, where the rate of change appears as a previous step to defining the derivative: here, it is the derivative that is at the service of the rate of change.

In summary, these types of SRPs started by the questioning of the epistemological models about certain themes, sectors or domains that prevail at secondary school and that can be interpreted as a consequence of monumentalism. In these models, mathematical praxeologies like proportionality, algebraic equations, elementary functions and derivatives are organised according to their theoretical components—their logos—, and the types of questions addressed—the praxis—are always presented as applications of already introduced notions or properties. Even in the case where real questions are proposed, their final aim is always to illustrate or construct the previously established praxeologies and, especially, their theoretical components: the concepts of proportionality, function, derivative, etc. In this school organisation of mathematics, modelling activities cannot find their real place, since the resolution of problems is always subordinated to the construction of the notions that structure the curriculum.

These examples illustrate how the problem of the ecology of some mathematical activities taking place at the level of the theme or the domain has to be addressed by questioning the higher levels of the scale of didactic codeterminacy because it is in these higher levels where some *raison d'être* of the aimed praxeologies can be found. This questioning leads to a redefinition, from the research perspective, of the praxeological organisations that conform the knowledge to be taught and learn. In this context, the role of SRPs is to serve as a study format that breaks some of the main assumptions of monumentalism and offers better conditions of existence for the alternative praxeological organisations.

**3.3 Questioning the discipline level at the university.** The previous SRPs take as a starting point a previous praxeological analysis of the content to be taught and learnt in order to overcome some identified didactic phenomena, especially related to the isolation and loss of the *raison d'être* of some curricular praxeologies. In a way, the generating question of the SRP is not the core objective of the inquiry process; the elements produced during the inquiry are. In this case, the SRP is not open, but *finalised*. The previous examples addressed a given theme or domain of school mathematics in order to reconstruct it in a more functional way. It was always foreseen that some or most of the praxeologies that constitute the curriculum would appear as answers to the questions raised during the inquiry. If a given syllabus is defined as a set of mathematical praxeologies  $\{A_1^\diamond, A_2^\diamond, \dots, A_r^\diamond\}$ , the design of a *finalised SRP* consists in finding a sequence of questions  $\{Q_1, Q_2, \dots, Q_k\}$  that could be derived from an initial question  $Q_0$ , the study of which is highly likely to activate a subset of the targeted praxeologies  $A_i^\diamond$ .

In the case of university education, the constraints imposed by the curricula are usually weaker than in secondary education, and lecturers have a greater degree of freedom to select and rearrange the subject matter content. The research carried out by Barquero [2009] proposes a finalised SRP that covers almost all the mathematical content of a first year course of Natural Sciences degrees. In this case, the level of the scale of didactic codeterminacy addressed is the discipline one. A unique question about the study of the dynamics of populations was proposed to a group of first year students of a degree in Chemical Engineering during four consecutive academic years. The SRP was proposed as a “mathematical modelling workshop” running parallel to the normal course during the whole year. Its main aim was to establish appropriate conditions for mathematics to be learnt as a modelling tool, starting from a generating question in the domain of natural science and using some of the main praxeologies included in the course syllabus. The initial question was formulated as follows:

Given the size of a population over previous periods of time, how can we predict the long-term behaviour of its size? What sort of assumptions about the population, its growth and its surroundings should be made? What kind of forecasts can be made and how to test them?

This question was specified with different populations: pheasants, fish and yeast. The first ones were modelled with discrete models and considered two cases: independent and mixed generations; with the third population a similar path was reproduced for the continuous case. The proposed SRP was divided into four branches (Fig. 4).

The design of the SRP also included the elaboration of a productive enough milieu to produce the emergence of the derived questions and the deconstruction and reconstruction of the new praxeological organisations that were required to help the inquiry progress. Some of these praxeological organisations were introduced by the lecturer in the “normal course”

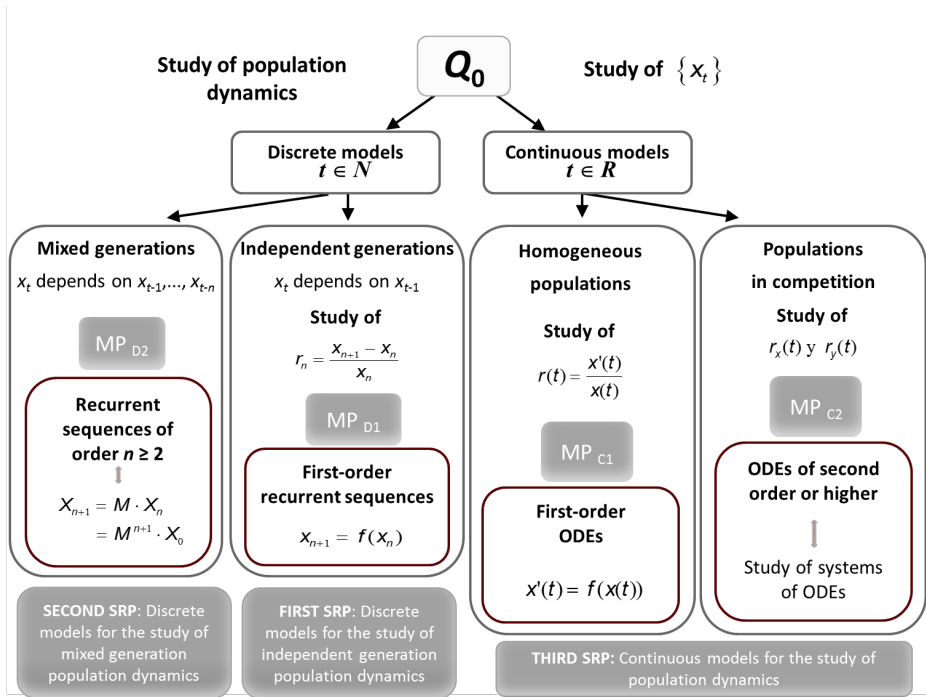


Figure 4. Structure of an SRP on population dynamics [Barquero and Bosch \[2015, p. 265\]](#)

(sequences, functions, derivatives, matrices, ordinary differential equations, etc.) and the students were asked to search for the more specific ones (Malthus and Verhulst models, transition and Leslie matrices, etc.) in the media available. During the four implementations of the SRP, the relationship between the course and the workshop evolved towards greater integration. It seemed as if the logic of the questions and answers derived from the SRP was gradually modifying the traditional organisation of the course contents since the lecturer agreed to introduce them when needed at the workshop, sometimes modifying the traditional organisation based on the theoretical coherence of the *syntheses*. Therefore, a workshop that was initially implemented as a complement to the course—to illustrate the main applications of the mathematical content introduced—started to acquire a prominent role, letting the course run as a nourisher of the inquiry process. This is in fact the ideal situation of the paradigm of questioning the world: a curriculum structured around a set of crucial questions to be studied and some flexible modules “on demand”, aimed at supplying the inquiry processes, when required, with some of the more basic needed praxeologies.

Unfortunately, this was not the destiny of the SRP on population dynamics, even if its evolution towards a more traditional teaching format can be easily explained by some constraints originated at the higher levels of didactic codeterminacy (see [Barquero, Bosch, and Gascón \[2013\]](#)).

**3.4 Weakening the school and pedagogical constraints.** The previous examples of SRPs were designed to ensure their viability by introducing new conditions and didactic resources given the common constraints encountered at the levels of schools, pedagogies and disciplines. What happens when we try to weaken these constraints and implement *more open* SRPs?

[Quintana \[2005\]](#) implemented an SRP at upper secondary level generated by the question of how to determine which of the three mobile companies operating at this moment in Spain was giving the best rate plans depending on the customer profile. This question was chosen by the research team to propose a situation where students had to build their own functional models and use the graph sketching as a tool to solve a problem (inequalities) that was too complex to be solved algebraically [Rodríguez, Bosch, and Gascón \[2008\]](#). The SRP lasted for 18 two-hour sessions and was implemented as an after-class activity for two consecutive academic years with groups of 10-12 volunteer students. The mathematics teacher presented it to the students as a mathematical workshop that would help them with the subject.

It is worth mentioning that this SRP, one of the first to be experienced, revealed important features about the traditional didactic contract that were implicitly assumed by the research team. Maybe the fact that the SRP was proposed as an optional after-class activity facilitated the participation of motivated students and weakened the school, the pedagogical and even the didactic constraints. Many unexpected outcomes appeared thanks to the students' initiative. For instance, during the very first session, once the class had raised some initial derived questions to approach the problem, and because there was no empirical information available (no Internet connection in the class), the students decided to invent some basic cases to start with. They were spontaneously creating their own exercises for a functional purpose! Another interesting anecdote to report is the teacher asking the students to stop comparing rates, once she saw that many of the most interesting functions had already been used and that the comparison was taking a lot of time. "There is no point in making comparisons if we do not compare everything with everything!", the students answered, thus reminding the teacher that the SRP was about providing an answer to the initial question, and not only about using functions to solve inequalities...

Another interesting and unexpected outcome was the fact that the students proposed to present the final report as an open interactive internet site where people could enter

their regular consumption and get advice in return. Three options were established depending on the kind of information required from the consumer: “normal people”, “lazy people” and “very lazy people” —the last one only asking for the minutes of conversation per day and the number of messages sent. Also, in one of the editions, the students decided to use their own invoices as a validation of their final answers, thus enriching the expected work with functions with some basic statistics description. Finally, surprised at the complexity of the work involved, they decided to write a letter to the Minister for Economy and Finance to complain about the consumers’ situation and the opacity of the information provided by companies. This illustrates how the inquiry into a question soon penetrates activity domains of different sorts, breaking the limits established at schools between disciplines and between domains within a discipline.

#### **4 The evolution of SRPs: integrating didactic tools into the inquiry process**

It seems clear that the paradigm of questioning the world represents important changes in the organisation of study processes at the different levels of the scale of didactic codeterminacy. Those at the level of civilisations are possibly the most hidden ones, since they correspond to beliefs or assumptions that are difficult to identify, unless we move to another civilisation, through the space or the time. The last case presented shows that the act of questioning, of posing queries about any aspect of our surrounding reality, has not always been assumed with normality by all civilisations and is still not clearly approved in all domains. It is not clear either that the access to any kind of labelled answer  $A^\diamond$  is seen as possible or appropriate for everybody at any time and in all the domains: each civilisation has its own forbidden spaces and implicit regulations. The hierarchy established among different types of knowledge—some being noble, others plebeian—is another variable to take into account. Societies also establish certain compartments in the organisation of knowledge and not everybody can easily move from one part to another. They also promote a given model of teacher that makes erudition prevail over inquiry, and do not succeed in making education evolve beyond the paradigm of visiting works: curricula formulated as lists of works, individual conception of learning, final school examinations based on tightly identified content, etc.

However, changes performed at the higher levels of the scale will remain limited if they do not come with the corresponding modifications at the lower levels. The inclusion of “competencies” as a key tool to impulse university teaching renovation shows the limitations of proposals that do not easily surpass the pedagogical level... But the lower levels of the scale introduced in figure 2 correspond to the paradigm of visiting works, where didactic systems  $S(X, Y, Q)$  are established around previously determined pieces of knowledge

located in relation to a given discipline. This cannot be assumed in the paradigm of questioning the world, where didactic systems are not formed around some selected works but a set of selected questions. In this case, the scale of levels of codeterminacy should end at the level of the didactic system, since what appears below will not only depend on the type of question addressed, but also on the decisions made by the inquirers about the possible works that are candidates to provide partial answers to the derived questions. Moreover, the choice of disciplines, domains or sectors where some existent labelled answers  $A_i^\diamond$  could be found is not a simple issue and, in any case, should be included as a question to deal with in the inquiry process. And it is not always a simple issue: let us just remember that, in World War 2, it took the British secret intelligence services a certain amount of time to associate the problem of deciphering the Nazi codes with the discipline of mathematics, a question that was traditionally associated with linguistics...

Therefore, in the paradigm of questioning the world, questions do not belong to any pre-established field of knowledge. Moreover, it is part of the inquiry process to investigate the possible sources of useful answers and, in particular, to mix praxeologies of a different nature, size and degree of “honourability”. Thus, the specific levels corresponding to the given disciplines have to be located below—or after—the didactic system (Fig. 5).

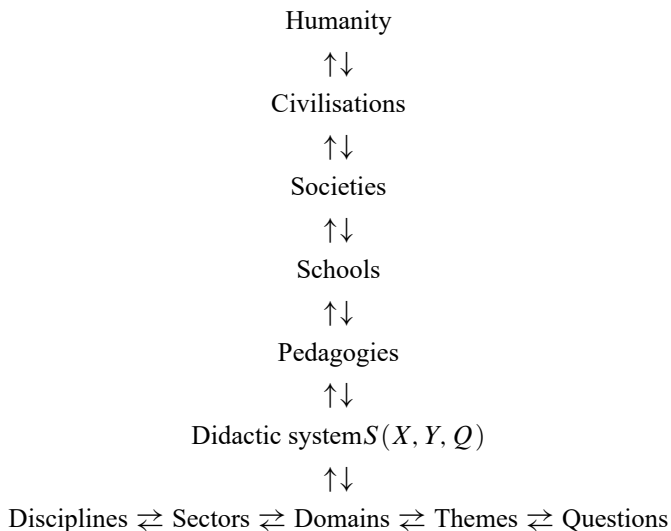


Figure 5. Scale of codeterminacy in the paradigm of questioning the world

An important constraint the experimented SRPs have revealed is precisely related to this intermediate level of the didactic system in the scale. It refers to a certain lack of knowledge resources to guide the inquiry that are not easy to locate in any of the official

descriptions of the disciplines. We have already mentioned this issue in the description of the SRPs' chronogenesis, concerning the difficulties of proposing explicit milestones to structure the inquiry process: teachers, together with students, need to discern where they are, what has been done so far and what seems to remain to be done at some crucial moments of the inquiry. When the inquiry is not a pre-established path—which it rarely is—the words and concepts to describe its main steps are not always available. In the paradigm of visiting works, the evolution of teaching processes can be formulated in terms of the praxeologies that have been visited, by naming their main components, the practical ones as well as the theoretical ones. The logos of the praxeologies provide descriptions of their main elements and present already-made assemblages of types of tasks and techniques, of techniques and properties or theorems, etc. As the study programme is predetermined, there are words, expressions and labels to designate the paths followed during the visits carried out in the study process: “We have covered limits of functions, we can now start with derivatives”.

The situation is very different when one is in the middle of a long inquiry process. There is no “official” discourse to *name* the elements of the inquiry process: the provisional results obtained, the questions derived, the paths selected and those that have been ruled out, etc. The work carried out in an SRP is always in need of new words, concepts and discourses. It is also in need of didactic—or epistemological—resources, for instance to manage the media-milieu dialectic, now that the school can no longer ignore that the access to external and unfiltered information is absolutely unavoidable.

The currently available school, pedagogical and didactic resources are unable to support the teachers' and the students' work in the same way textbooks, treatises, encyclopaedias, documentaries, etc. do in the paradigm of visiting works. In our first explorations of the ecology of the paradigm of questioning the world, it was up to teachers and students to elaborate their own narrative of the inquiry process, and to establish their own milestones to mark the path and set the pace. In the last experimented SRPs [Florensa, Bosch, and Gascón \[2016\]](#), some of the elements of the Herbartian schema, such as the questions-answers maps, are starting being used as explicit tools for teachers and students to manage the inquiry process. It is not impossible than others, like the media-milieu dialectics for instance, will also turn out to be productive at this respect. The lack of epistemological resources to manage inquiry processes seems to be one of the main open problems raised by the research on SRPs. Its solution will certainly require the contribution of scholars of different fields—and also mathematicians—to create new knowledge means to fill the gap. Our current efforts go in this direction.



## References

- B. Barquero (2009). “Ecología de la modelización matemática en la enseñanza universitaria de las matemáticas”. PhD thesis. Universitat Autònoma de Barcelona, Spain (cit. on p. 4048).
- B. Barquero and M. Bosch (2015). “[Didactic engineering as a research methodology: From fundamental situations to study and research paths](#)”. In: *Task design in mathematics education*. Switzerland: Springer, pp. 249–271 (cit. on p. 4049).
- B. Barquero, M. Bosch, and J. Gascón (2013). “The ecological dimension in the teaching of mathematical modelling at university”. *Recherches en Didactique des Mathématiques* 33 (3), pp. 307–338 (cit. on p. 4050).
- M. Bosch and J. Gascón (2006). “25 years of didactic transposition”. *ICMI Bulletin* 58, pp. 51–64 (cit. on p. 4038).
- M. Bosch and C. Winsløw (2015). “Linking problem solving and learning contents: the challenge of self-sustained study and research processes”. *Recherches en Didactique des Mathématiques* 35 (3), pp. 357–401 (cit. on p. 4041).
- G. Bouligand (1957). “L’activité mathématique et son dualisme”. *Dialectica* 11, pp. 121–139 (cit. on p. 4034).
- G. Brousseau (1997). *Theory of didactical situations in mathematics*. Dordrecht, NL: Kluwer (cit. on pp. 4038, 4040, 4044).
- Y. Chevallard (2002). “Organiser l’étude. 3. Écologie & régulation”. In: *Actes de la 11e École d’Été de didactique des mathématiques*. Grenoble: La Pensée Sauvage, pp. 41–56 (cit. on p. 4035).
- (2007). “Passé et présent de la Théorie Anthropologique du Didactique”. In: *Sociedad, escuela y matemáticas. Aportaciones de la Teoría Antropológica de lo Didáctico*. Jaén, Spain: Publicaciones de la Universidad de Jaén, pp. 705–746 (cit. on p. 4034).
- (2015). “[Teaching Mathematics in tomorrow’s society: a case for an oncoming counter paradigm](#)”. In: *Proceedings of the 12th International Congress on Mathematical Education*. Springer International Publishing, pp. 173–187 (cit. on pp. 4036, 4039).
- E. Cid (2015). “Obstáculos epistemológicos en la enseñanza de los números negativos”. PhD thesis. Universidad de Zaragoza, Spain (cit. on pp. 4037, 4038).
- I. Florensa, M. Bosch, and J. Gascón (2016). “SRP design in an Elasticity course: the role of mathematic modelling”. In: *Proceedings of INDRUM 2016*. Montpellier, France: Université de Montpellier & INDRUM, pp. 191–200 (cit. on p. 4053).
- F. J. García (2005). “La modelización como herramienta de articulación de la matemática escolar. De la proporcionalidad a las relaciones funcionales”. PhD thesis. Universidad de Jaén, Spain (cit. on p. 4046).

- F. J. García, J. Gascón, L. Ruiz Higuera, and M. Bosch (2006). “[Mathematical modelling as a tool for the connection of school mathematics](#)”. *ZDM – International Journal on Mathematics Education* 38 (3), pp. 226–246 (cit. on p. 4046).
- B. Hansen and C. Winsløw (2011). “Research and study diagrams as an analytic tool: The case of bi-disciplinary projects combining mathematics and history”. In: *Un panorama de la TAD*. Barcelona: CRM, pp. 685–694 (cit. on p. 4041).
- B. E. Jessen (2014). “How can research and study courses contribute to the teaching of mathematics in an interdisciplinary setting?” *Annales de Didactique et de Sciences Cognitives* 19, pp. 199–224 (cit. on p. 4041).
- (2017). “Study and Research Paths at Upper Secondary Mathematics Education: a Praxeological and Explorative study”. PhD thesis. University of Copenhagen (cit. on p. 4043).
- P. Liljedahl (2016). “[Building Thinking Classrooms: Conditions for Problem-Solving](#)”. In: *Posing and Solving Mathematical Problems*. Springer International Publishing, pp. 361–386 (cit. on p. 4044).
- C. Oliveira Lucas (2015). “Una posible razón de ser del cálculo diferencial elemental en el ámbito de la modelización funcional”. PhD thesis. Universidad de Vigo (cit. on p. 4047).
- V. Parra and R. Otero (2017). “Enseñanza de la matemática por recorridos de estudio e investigación: indicadores didáctico-matemáticos de las “dialécticas””. *Revista de Educación Matemática* 29 (3), pp. 9–49 (cit. on p. 4043).
- E. R. Quintana (2005). “Metacognición, resolución de problemas y enseñanza de matemáticas una propuesta integradora desde el enfoque antropológico”. PhD thesis. Universidad Complutense de Madrid (cit. on p. 4050).
- E. Rodríguez, M. Bosch, and J. Gascón (2008). “[A networking method to compare theories: metacognition in problem solving reformulated within the anthropological theory of the didactic](#)”. *ZDM Mathematics Education* 40, pp. 287–301 (cit. on p. 4050).
- L. Ruiz-Higuera and F.J. García García (2011). “Análisis de praxeologías didácticas en la gestión de procesos de modelización matemática en la escuela infantil”. *Relime* 14 (1), pp. 41–70 (cit. on p. 4043).
- N. Ruiz-Munzón (2010). “La introducción del álgebra elemental y su desarrollo hacia la modelización funcional”. PhD thesis. Universitat Autònoma de Barcelona, Spain (cit. on p. 4046).
- N. Ruiz-Munzón, Y. Matheron, M. Bosch, and J. Gascón (2012). “Autour de l’algèbre: les entiers relatifs et la modélisation algébrique-fonctionnelle”. *Recherches en Didactique des Mathématiques*. Numéro spécial hors-série : Enseignement de l’algèbre élémentaire. Bilan et perspectives, pp. 87–106 (cit. on p. 4046).

- C. Winsløw, Y. Matheron, and A. Mercier (2013). “[Study and research courses as an epistemological model for didactics](#)”. *Educational Studies in Mathematics* 83 (2), pp. 267–284 (cit. on p. 4041).

Received 2017-12-01.

MARIANNA BOSCH CASABÒ

[marianna.bosch@iqs.url.edu](mailto:marianna.bosch@iqs.url.edu)

## ON THEORIES IN MATHEMATICS EDUCATION AND THEIR CONCEPTUAL DIFFERENCES

LUIS RADFORD

### Abstract

In this article I discuss some theories in mathematics education research. My goal is to highlight some of their differences. How will I proceed? I could proceed by giving a definition,  $T$ , of the term *theory* and by choosing some differentiating criteria such as  $c_1$ ,  $c_2$ , etc. Theories, then, could be distinguished in terms of whether or not they include the criteria  $c_1$ ,  $c_2$ , etc. However, in this article I will take a different path. In the first part I will focus on a few well-known theories in Mathematics Education and discuss their differences in terms of their *theoretical stances*. In the last part of the article, I will comment on a sociocultural emergent trend.

### Introduction

In order to make sense of problems around the teaching and learning of mathematics, mathematics educators have come up with different theories. Currently, there is a large number of theories in use. My goal is to highlight some of their differences. How will I proceed? I could proceed by giving a definition,  $T$ , of the term *theory* and by choosing some differentiating criteria such as  $c_1$ ,  $c_2$ , etc. Theories, then, could be distinguished in terms of whether or not they include the criteria  $c_1$ ,  $c_2$ , etc. see Radford [2008a, 2017a]. In this article, however I will take a different path. In the first part of the article, I will focus on a few well-known theories in Mathematics Education and discuss their differences in terms of their *theoretical stances*. In the last part of the article, I will comment on a sociocultural emergent trend.

My choice of theories has been guided by what may be termed their historical impact in the constitution of mathematics education as a research field. By historical impact I do not mean the number of results that a certain theory produced in a certain span of time. Although important, what I have in mind here is something related to the foundational principles of a theory. The foundational principles of a theory determine the research

questions and how to tackle them within a certain research field, thereby helping to shape the form and determine the content of the research field itself.

To discuss the types of theories in our field is to discuss their differences and, more importantly, what accounts for these differences. My argument is that these differences are better understood in terms of *theoretical suppositions*. [Sriraman and English \[2005\]](#) argued that the variety of frameworks in mathematics education is directly related to differences in their epistemological perspectives. I suggest that, in addition to the underpinning corresponding epistemologies, differences can also be captured by taking into account the cognitive and ontological principles that theories in mathematics education adopt.

Obviously, I will neither be able to present a rich sample of theories in mathematics education nor will I be able to delve deeply into the intricacies of any of them. I hope, nonetheless, that by focusing on a few theories, and contrasting their theoretical suppositions, we may gain a sense of their distinctiveness and thereby better understand the notion and the types of theories in our field.

Because of space constraints, I will deal with three theories. Although other choices are certainly possible, I will deal with Constructivism, the Theory of Didactic Situations, and Socio-Cultural Theories.

## 1 Constructivism

**1.1 The Theoretical Principles.** During the 1980s and 1990s, Constructivists introduced their theory as based on two main principles:

*p1*: knowledge is not passively received but built up by the cognizing subject;  
and

*p2*: the function of cognition is adaptive and serves the organization of the experiential world, not the discovery of ontological reality. [von Glasersfeld](#) [see [1995](#), p. 18]

Principle *p1* stresses constructivism's opposition to teaching by transmission. Constructivism, indeed, emerged as an option against behaviourism and its pedagogy of direct teaching. It is in this context that Paul Cobb remarked some twenty years ago that

An abundance of research indicates that students routinely use prescribed methods to solve particular sets of tasks on which they have received instruction without having developed the desired conceptual knowledge. [Cobb](#) [1988, p. 90]

However, although historically important, the true novelty of the constructivist perspective does not rest on the first principle. It rests, rather, as von Glaserfeld claims, on the

epistemic and ontological attitudes conveyed by the second principle and its concomitant concept of *knowledge*. Without necessarily denying the existence of a pre-existent reality, and in a move consistent with Kant's theory of knowledge, constructivism does not claim that the knowledge constructed by the cognizing subject corresponds to such a reality; its epistemology rests precisely on the denial of the possibility of any certain knowledge of reality Ernest [1991].

In the beginning, constructivism envisioned the goals of mathematics instruction along the lines of Piaget's epistemology. At the end of the 1980s, Cobb argued that the goal of instruction is or should be to help students build [mental] structures that are more complex, powerful, and abstract than those that they possess when instruction commences Cobb [1988, p. 89]. The pedagogical problem was then to create the classroom conditions for the development of complex and powerful mental structures.

The constructivist research was oriented to a great extent to the study of the development of the students' mental arithmetic and other mathematical structures and to the investigation of the students' difficulties in developing them. Particular attention was paid to the students' counting types and construction of arithmetic units see e.g. Cobb [1985], Steffe and von Glasersfeld [1983] and Steffe, von Glasersfeld, Richards, and Cobb [1983].

The creation of the classroom conditions for the development of mental structures led unavoidably to the question of the role of the teacher. Cobb said:

The teacher's role is not merely to convey to students information about mathematics. One of the teacher's primary responsibilities is to facilitate profound cognitive restructuring and conceptual reorganizations. Cobb [1988, p. 89]

A close examination of the role of the constructivist teacher shows that the constructivist epistemic and ontological principles were underpinned by a general concept of the cognizing subject that framed the specific role of the student and the teacher. For constructivism, the epistemic and ontological principles  $p1$  and  $p2$  make sense only in the context of a self that is autonomously constructing her knowledge. If we remove the autonomy principle, constructivism becomes simply a variant of certain socio-cultural approaches. This third principle can be formulated as follows:

$p3$ : the cognizing subject not only constructs her own knowledge but she does so in an autonomous way.

Intellectual autonomy was in fact part of two of the general goals identified by constructivism from the outset:

teaching by imposition is incompatible with two general goals of mathematics instruction that follow from constructivism, the construction of increasingly powerful conceptual structures and the development of intellectual autonomy. Cobb [ibid., p.100]

As I argued elsewhere [Radford \[2008c\]](#), the idea of the autonomous cognizing subject conveyed by constructivism was not a novelty in education. In fact, just such an idea is at the heart of the concept of the self of Western modernity—an idea that goes back to the very roots of Kant’s theory of knowledge and its related epistemic subject. Kant’s epistemic subject is not one that receives knowledge but one that produces it. It is a constructor that epitomizes the idea of man as *homo faber*. However, as we shall see later, although interesting from a historic viewpoint, this epistemic concept of the cognizing subject as an autonomous constructor of its own knowledge is considered too restrictive to account for the concrete processes of learning in the classroom and constitutes a point of divergence of theories in mathematics education.

**1.2 The Ontology of Constructivism.** The constructivist denial of the possibility of knowledge of reality is not mere fancy nor extravagant ontological position. It is, rather, one of the consequences of the remarkable subjectivism in which it was rooted from the start. The cognizing subject of modernity found itself in a world whose understanding was no longer assured by tradition and the interpretations offered by religion. The understanding of the world could only come from what the cognizing subject could accomplish through its sensing body and its intellect. Starting from the senses as the basic structure of knowledge, David Hume argued in the 18th century that the establishment of logical necessity was impossible to ascertain, for all that we can witness are particular associations occurring among events. Hume was perhaps the first thinker to express in the clearest way the finitude of the human condition that results from a subjectivism that started to arise from the Renaissance and that was clearly articulated by the philosophers of the Enlightenment. The long period that followed Kant’s *Inaugural Dissertation*, published in 1770 (for a modern translation see [Kant \[1894\]](#)) and the first critique, that is the *Critique of Pure Reason*, published in 1781 (for a modern translation see [Kant \[2003\]](#)), the so-called silent decade, is explained by the intense cogitations in the course of what Kant sought for a solution to Hume’s problem. This decade of intense cogitations led Kant to the development of his ontology [Goldmann \[1971\]](#), a neutral ontology, the main feature of which is, as von Glasersfeld noted, the abandonment of claims about the knowability of reality – i.e., an ontology that neither asserts that knowledge is about reality nor that it is not.

However, Kant’s neutral ontology has an exception: the neutral ontology of Kant does not apply to mathematical knowledge. For Kant, mathematics was the paradigmatic example of certain knowledge. This is what Kant meant by the a priori status of mathematics, a status that put mathematical objects (in opposition to phenomenological objects such as chairs and dogs) within the realm of the truly knowable.

Kant’s ontology rests on a form of *a priorism* that Piaget did not endorse. For Piaget, and for the ensuing constructivism in education, knowledge (mathematical or not) has to

be constructed. Since there was no way to check the correspondence between subjective constructs produced by the cognizing subject and reality, von Glasersfeld suggested that knowledge is not about *certainty* but about *viability*. A piece of knowledge is kept by the cognizing subject as long as it seems to work. All knowledge is hypothetical.

This concept of knowledge has some interesting corollaries. One of them is that since everyone constructs his or her own knowledge, we can never be sure that we are talking about the same things. We can just assume or pretend that we are perhaps sharing something. For constructivists, we take knowledge and meanings as *taken-as-shared*. Naturally, one question that has been raised in this regard is whether or not the subjectivist idea of knowledge and meaning conveyed by constructivism is a form of solipsism. Constructivists answer negatively, stressing the role of social interaction in the cognizing subject's construction of viable knowledge.

**1.3 Social Knowledge in Constructivism.** Although some mathematics educators were intrigued by the extreme relativism of the Kantian constructivist neutral epistemology see e.g. [Goldin \[1990\]](#), ontological questions seemed to recede into the background as constructivist teachers and researchers were preoccupied with the understanding of good practices to ensure the students' development of mental structures. Naturally, the search for solutions was framed by constructivism's principles. In particular, the question was to devise pedagogical actions coherent with the idea of avoiding teaching the answers and influencing the student's reasoning. In short, the question was how to teach without trespassing into the domains of the student's self-determination. The solution was sought in the idea of the classroom as a space of *negotiation* of meanings.

Later on, this idea was developed further, perhaps as a result of the dialogue between constructivists and the German interactionists [Bauersfeld \[1980\]](#), [Voigt \[1985\]](#), etc. Thus, in the early 1990s, constructivism was formulating the learning-teaching process as a process that is interactive in nature and involves the implicit and explicit negotiation of mathematical meanings. In the course of these negotiations, the teacher and students elaborate the *taken-as-shared* mathematical reality that constitutes the basis for their ongoing communication [Cobb, Yackel, and Wood \[1992, p. 10\]](#).

Through the insertion of the idea of mathematics as a social practice and the classroom as a space of negotiation of meanings, constructivism moved into a new direction. In an article published in 1994, Cobb described two different constructivist research lines. The first remained centred around the investigation of the students' development of mental structures. The second focused rather on the evolution of meanings in the course of the students' interaction in the classroom [Cobb \[1994\]](#).

One of the challenges for this second line of research was to make the idea of interaction operational within the constraints imposed by their three basic principles. The



operationalization was made through a clear distinction between: (1) the students' psychological processes, on the one hand, and (2) the social processes of the classroom, on the other. While the investigation of students' psychological processes went along the lines of Piaget's concept of reflective abstraction, the social processes were related to the idea of *collective classroom reflection* Cobb, Boifi, McClain, and Whitenack [1997].

Certainly, developing the new research line was not an easy move. It had to take into account social interaction in a context where, as a result of the theoretical principles, constructivism found itself with not too much room left. Indeed, interaction had to be devised in such a way that the inclusion of the Other in the cognizing subject's act of knowing left no room for interference with the autonomous constructivist cognizing subject. From the outset, there was a vivid tension between the students' mathematical meanings and those of the teacher: "The teachers' role in initiating and guiding mathematical negotiations is a highly complex activity that includes ... implicitly legitimizing selected aspects of contributions" Cobb, Wood, Yackel, Nicholls, Wheatley, Trigatti, and Perlwitz [1991, p. 7]. To explicitly legitimize selected students' contributions would jeopardize, indeed, the constructivist project and its principle that knowledge construction is a personal and self-determining matter.

The dichotomy that constructivism erects between its culturally detached autonomous cognizing subject and the socio-cultural historical traditions in which this cognizing subject thinks and acts, turns out to be, as many find, an unsatisfactory solution. Thus, given the theoretical principles adopted by constructivism, Waschescio [1998] argues that a link between the individual and the cultural realm is certainly missing. Actually, as Lerman claims, such a link is simply impossible to find Lerman [1996].

To sum up, constructivism is a student-centred theory. Its influence in education has been very impressive, not only in North America but all over the world. The detailed analyses of classroom interaction and the sophisticated methodologies designed to scrutinize the negotiation of meanings underpinning the students' conceptual growth have helped the community of mathematics educators become aware of the variety of meanings that the students mobilize in tackling mathematical problems. Constructivism has certainly helped us to better understand the complexities surrounding the students' processes of learning and provides us with an alternative to direct teaching.

## 2 The Theory of Didactic Situations

The Theory of Didactical Situations (TDS) seeks to offer a model, inspired by the mathematical theory of games, to investigate, in a scientific way, the problems related to the teaching of mathematics and the means to enhance it.

In the beginning, the term *situation* referred to the student's environment as handled by the teacher for whom it appears as a tool in the process of teaching. Later, the situation was enlarged in order to include the teacher herself and even the educational system as a whole Brousseau [1997a].

As any theory, the TDS works on a set of principles, among them the following epistemic ones:

*p1*: knowledge results as the —optimal solution to a certain situation or problem.

*p2*: learning is —in accordance to Piaget's genetic epistemology— a form of cognitive adaptation.

As in the case of constructivism, these principles are supplemented by a conception of the roles that teacher and students have to play in the classroom:

**2.1 The Role of the Teacher.** An essential part of the teacher's role is not to show the students how to solve the problems, but rather to let the students deal with them, for doing mathematics does not consist only of receiving, learning, and sending correct, relevant (appropriate) mathematical messages Brousseau [1997b, p. 15]. Like Constructivism, the TDS is opposed to direct teaching. The teacher's role is rather to identify the problems or situations that will be given to the students and that will provoke the expected learning.

**2.2 The Role of the Student.** The student which the TDS talks about is an epistemic subject, a sort of ideal model of the individual, conceived of as behaving (or having to behave) in a rational manner, in a way close to the behaviour of the mathematician. Her role is to engage in mathematical problems in a way that is coherent with the professional scientific practice. In the course of a faithful reproduction of scientific activities, the student is required to produce, formulate, prove, and construct models, languages, concepts and theories. Brousseau [*ibid.*, p.22].

The roles of the teacher and the student are explained in the following passage:

The modern conception of teaching ... requires the teacher to provoke the expected adaptation in her students by a judicious choice of problems that she puts before them. These problems, chosen in such a way that the students can accept them, must make the students act, speak, think, and evolve by their own motivation. Brousseau [*ibid.*, p. 30]

The judicious choice of problems is, of course, a delicate part of the teaching process. Its concrete possibility rests on the following epistemological assumption:

*p3*: for every piece of mathematical knowledge there is a family of situations to give it an appropriate meaning.

This family is called a *fundamental situation*. For Brousseau [1997b, p. 24], the search for fundamental situations and their insertion into the more general classroom project of teaching and learning requires at least two elements: a good epistemological theory, which would reveal the deepness of mathematical knowledge and positively inform the teaching process, and a good didactic engineering, which would be oriented to the design of situations and problems to be solved by the students.

A fourth principle specifies further the concept of learning in the TDS. The general epistemic principle *p2* tells us that learning is of an adaptive nature; it consists of the students' adaptations to a milieu, but it does not say anything about the socio-interactional conditions to be fulfilled for it to occur. Principle four fills the gap and gives an impeccable theoretical consistency to the TDS –although, as we will see, some paradoxes will appear later on:

*p4*: the student's autonomy is a necessary condition for the genuine learning of mathematics.

Thus, if the process of learning was not accomplished autonomously vis-à-vis the teacher, learning could not have happened. For “if the student produces her answer without having had herself to make the choices which characterize suitable knowledge and which differentiate this knowledge from insufficient knowledge, the evidence [of learning] becomes misleading” Brousseau [*ibid.*, p. 41]. In other words, “if the teacher teaches her [the student] the result, she does not establish it herself and therefore does not learn mathematics” Brousseau [*ibid.*, pp. 41-42].

The student is hence expected to engage with a fundamental situation in a particular type of game that gives rise to another situation, called *adidactic* Brousseau [*ibid.*, p. 30], characterized by the student's autonomy vis-à-vis the teacher. What makes the *adidactic* situation different is the fact that it is partially freed from the teacher's direct interventions Brousseau [2003, p. 2]. This is why, referring to the *adidactic* situations –the only one through which true knowledge acquisition can be said to happen (knowledge by adaptation)– Brousseau asserts that “Between the moment the student accepts the problem as if it were her own and the moment when she produces her answer, the teacher refrains from interfering and suggesting the knowledge that she wants to see appear” Brousseau [1997b, p. 30].

Within this context, the teacher's mission is not only to ensure the successful devolution of the fundamental situation to the student in the *adidactic* situation, but also to maintain a fruitful interaction with the milieu (i.e., the antagonist system of the actors) in an encompassing context called the *didactic* situation. As Brousseau puts it,

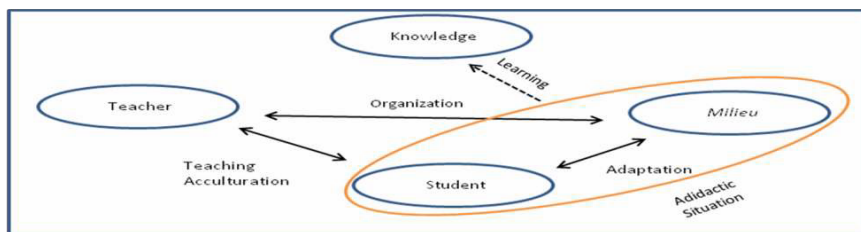


Figure 1: The four-pole (simplified) diagram shows the basic components of a Didactic Situation.

This situation or problem chosen by the teacher is an essential part of the broader situation in which the teacher seeks to devolve to the student an adidactical situation which provides her with the most independent and most fruitful interaction possible. For this purpose, according to the case, the teacher either communicates or refrains from communicating information, questions, teaching methods, heuristics, etc. She is thus involved in a game with the system of interaction of the student with the problem she gives her. This game, or broader situation, is the *didactical situation*. Brousseau [ibid., pp. 30–31]

Figure 1 (which is a simplified and modified version of Perrin-Glorian and Hersant [2003] diagram) conveys the complexity of a didactic situation.

The didactic situation is in the end a model that can be better conceptualized as a game see Brousseau [1988]. The situation models the interaction of a subject with a milieu by a game (e.g. a problem to solve) where players have to take decisions: some states of the game are more favourable than others to win; thus the situation defines a piece of knowledge as a means for the subject to reach or maintain a favourable state (for the game) in this milieu Perrin-Glorian [1994]

In practice, however, the game does not necessarily proceed smoothly. The student may fail to solve the problem or simply may avoid it. A negotiation takes place:

Then a relationship is formed which determines – explicitly to some extent, but mainly implicitly – what each partner, the teacher and the student, will have the responsibility for managing and, in some way or other, be responsible to the other person for. This system of reciprocal obligation resembles a contract. What interests us here is the *didactical contract*, that is to say, the part of this contract which is specific to the “content”, the target mathematical knowledge. Brousseau [1997b, pp. 31–32]

Brousseau acknowledges that this system of reciprocal obligations is not exactly a contract in so far as it is not fully explicit. It is rather something like a flexible, ongoing negotiation. However, this is not a negotiation in the sense of constructivism, for what is being negotiated in the TDS is neither the mathematical meanings constructed in the classroom by the students and the teachers nor the mathematical forms of proving, arguing, etc. For the TDS, in opposition to constructivism, mathematical meanings and the mathematical forms of proving are not negotiable: they are part of the target knowledge, the cultural knowledge of reference. Negotiation is about the fluctuating borders of a teacher-student division of labour that seeks to ensure that the teacher's devolution of the fundamental situation is accepted by the student; that is to say, that the student takes responsibility for the solution of the problem and enters into an adidactic situation.

Because of its own nature, the unavoidable fuzzy didactic contract is haunted by some paradoxes. Let me dwell briefly on this point.

**2.3 The Paradoxes of Learning.** Teachers have the social obligation to make sure that learning is happening in the classroom.

What to do, then, if the student fails to learn? The student will ask the teacher to be taught. But

the more the teacher gives in to her demands and reveals whatever the student wants, and the more she tells her precisely what she must do, the more she risks losing her chance of obtaining the learning which she is in fact aiming for. Brousseau [1997b, p. 41]

Brousseau does not consider this paradox as a contradiction. The paradox reveals the tricky situation that the teacher will be often called upon to live in the classroom. If the teacher gives up, knowledge attainment will be compromised:

everything that she [the teacher] undertakes in order to make the student produce the behaviours that she expects tends to deprive this student of the necessary conditions for the understanding and the learning of the target notion; if the teacher says what it is that she wants, she can no longer obtain it. Brousseau [ibid., p. 41]

Another paradox may arise when it is not possible to find a fundamental situation that would fit the students' intellectual possibilities at a certain point of their development. In this case, the teacher

gives up teaching by adaptation; she teaches knowledge directly in accordance with scientific requirements. But this hypothesis implies that she must give up providing a meaning to this knowledge and obtaining it as an answer

to situations of adaptation because then the students will colour it with false meanings. Brousseau [ibid., p. 42]

According to Brousseau, the student is also put in a paradoxical situation: “she must understand AND learn; but in order to learn she must to some extent give up understanding and, in order to understand, she must take the risk of not learning” Brousseau [ibid., p. 43].

For the TDS, these (and other paradoxes) are an intrinsic part of didactic situations. They are part of the teaching of mathematics and knowledge acquisition. However, these paradoxes can also be seen as the result of a tension in the TDS’ account of teaching and learning—a tension that results from a particular conception of learning, an epistemological and ontological rationalist view of mathematics and its adherence to a classical concept of the cognizing subject.

**2.4 The Idea of Learning.** As seen previously, for the TDS, genuine learning can only arise from the individual’s own deeds and reflections. It is this theoretical stance on learning that gives sense to the ideas of didactic situation and devolution. Although the TDS involves social interaction at different levels Kidron, Lenfant, Bikner-Ahsbabs, Artigue, and Dreyfus [2008], knowledge acquisition is, in the end, the result of the student’s personal relationship with the object of knowledge. There is no difference in this respect between constructivism and the TDS. Differences arise in terms of, for example, the epistemic role of the problem: while in the former, a problem may lead to diverse, equally genuine *viable* pieces of knowledge, in the latter, in contrast, the design of the didactic situation should lead to the target cultural knowledge.

As we will see in the next section, the road taken by Vygotskian Sociocultural contemporary approaches to the problem of teaching and learning is different in important ways.

The TDS has had a significant influence in France and French-speaking communities around the world. It has also had an important influence in Spain and Central and South America. The detailed epistemic analyses of fundamental situations, their engineering and control in the classroom by the teacher, have helped mathematics educators understand the key role of suitable problems in the development of students’ mathematical thinking.

### 3 Sociocultural Approaches

We have seen that for both constructivism and the TDS, the autonomy of the cognizing subject vis-à-vis the teacher, is a prerequisite for knowledge acquisition. For sociocultural approaches, autonomy is not the prerequisite of knowledge acquisition. Autonomy is, in fact, its result. This is one of the central ideas of Vygotsky’s concept of *zone of proximal*

*development*. Therefore, it is easy to imagine that, within sociocultural approaches of Vygotskian ascent, the roles of the teacher and the students are theorized along very different lines from what is found in other theories. This important difference will become clearer as I present a summary of the main principles of sociocultural approaches.

**3.1 The Ontological and Epistemological Principles.** The *ontological* position of a theory consists of specifying the sense in which the theory approaches the nature of conceptual objects (in our case, the nature of mathematical objects, their forms of existence, etc.). The *epistemological* position consists of specifying the way in which, according to the theory, these objects can (or cannot) end up being known.

One of the most popular ontologies is Realism. Realists consider that the existence of mathematical objects precedes and is independent from the activity of individuals and that they exist independently of time and culture. Contemporary sociocultural approaches take a different route:

*pl*: knowledge is historically generated during the course of the mathematical activity of individuals.

The principles of the TDS and constructivism seem to be in agreement with this ontological stance. If there is not a discrepancy in the “mode of being” of mathematical knowledge, there might be nonetheless some discrepancies in terms of its “modes of production.” As seen earlier, the TDS and constructivism consider knowledge as the result of the adaptive actions of the cognizing subject. For socioculturalists, however, adaptation is insufficient to account for the production of knowledge. One of the reasons is that socioculturalists consider cognition as a cultural and historically constituted form of reflection and action embedded in social praxes and mediated by language, interaction, signs, and artifacts. As a result, knowledge is produced by cognizing subjects who are, in their productive endeavours, subsumed in historically constituted traditions of thinking. The cognizing subject of sociocultural theories is a subject that thinks within a cultural background and that, in so doing, goes beyond the necessities of mere ahistorical adaptive urges. In other terms, the “will to knowledge” (to borrow Foucault’s term) and the way knowledge comes into being are neither driven nor shaped by adaptive needs or impulses to produce “viable” hypotheses or “optimal” results. The “will to knowledge” and knowledge itself are rather mediated by cultural forms of thinking and values (scientific, aesthetic, ethic, etc.) that orient (without imposing) the growth of knowledge into certain new directions. Within sociocultural contexts, viability cannot be understood as a mere subjective game of hypothesis generation by a cognizing subject in its attempt at getting around its environment. Much in the same way, *optimality* cannot be understood in terms of some universal, intrinsic mechanisms of mathematical knowledge. Mathematical thinking and mathematical responses are always framed by the particular rationality

of the culture where they take place; within these cultures optimality can have different meanings and may not be the main drive to move mathematical thinking to new levels of development [Radford \[1997b\]](#), [Radford \[2008a\]](#).

For instance, the ways of dealing with the prediction of future events or the understanding of past events in early 20th century Azande culture was not at all moved by questions of optimality. The Azande reasoning was inscribed in a different worldview from the Ho versus Ha view of hypotheses testing of Western mathematics. And yet, like the latter, the Azande's ceremonial procedures were clear processes of understanding and making sense of their reality [Evans-Pritchard \[1937\]](#), [Feyerabend \[1987\]](#), and [Radford \[2017b\]](#).

We can summarize this discussion in the following principle:

*p2*: the production of knowledge does not respond to an adaptive drive but is embedded in cultural forms of thinking entangled with a symbolic and material reality that provides the basis for interpreting, understanding, and transforming the world of the individuals and the concepts and ideas they form about it.

**3.2 Learning.** In the previous section it was argued that socioculturalists claim that from a phylogenetic point of view, conceptual objects are generated in the course of human activity. From an ontogenetic point of view, the central problem is to explain how acquisition of the knowledge deposited in a culture can be achieved: this is a fundamental problem of mathematics education in particular and of learning in general.

The metaphor of knowledge construction seems to convey very well the idea that knowledge is not something transcendental to the human sphere and that knowledge is rather something made by human beings. Constructivism, the TDS, and sociocultural perspectives agree on this point.

However, from a sociocultural perspective, the extrapolation of this metaphor to the ontogenetic dimension leads to a series of important irresolvable problems. Instead of talking about students constructing knowledge, some socioculturalists prefer to talk about students making sense of, and becoming fluent with, historically constituted modes of thinking. One of the advantages in putting the problem of learning in this way is that the student's knowledge is not seen as something coming from *within* (a kind of private or subjective construction endlessly seeking to reach a culturally-objective piece of knowledge) but from *without*. Principle 3 summarizes this idea:

*p3*: learning is the reaching of a culturally-objective piece of knowledge that the students attain through a social process of *objectification* mediated by signs, language, artifacts, and social interaction as the students engage in cultural forms of reflecting and acting.



The idea of learning as the reaching of cultural knowledge should not be interpreted as if the students reach knowledge in a passive way. Unfortunately, we have become used to making a dichotomy and to thinking that either students construct their own knowledge *or* knowledge is imposed upon them. This is a too easy and misleading oversimplification –what Lerman has termed the absolutist view about learning [Lerman \[1996\]](#). Learning, from a sociocultural perspective, is the result of an active engagement and self-critical, reflexive, attitude towards what is being learned. Learning is also a process of transformation of existing knowledge. And perhaps more importantly, learning is a process of the formation of subjectivities, a process of agency and the constitution of the self.

Sociocultural approaches resist indeed the idea that learning is just the uncritical appropriation of existing knowledge absorbed by a passive student-spectator. Knowledge has a transformative power: it transforms the object of knowledge and, in the course of knowing and learning, the subject is itself transformed. There is a dialectical relationship between subject and object that can be better understood by saying that learning is a process of objectification (*knowing*) and subjectification (or *agency*), that is a process of *being* [Radford \[2008c\]](#).

**3.3 The Role of the Teacher and the Students.** The role of the teacher is not, as it can be imagined from what we just said, to dispense knowledge. Since sociocultural approaches argue that knowledge cannot be injected into the students' mind<sup>1</sup>, in order to get the students to know (in the sociocultural transformative sense) objects and products of cultural development, one of the roles of the teacher is to offer students rich classroom activities featuring, in a suitable manner, the encounter with the various layers of generality of historical cultural objects and the encounter with other voices and forms of understanding.

The configuration of these activities (both in terms of the mathematical content and its social- interactive dimension) is framed by the ultimate socioculturalists' idea of how learning occurs. As already mentioned, for socioculturalists, learning will not necessarily or uniquely occur as the result of the student's autonomous cogitations in her attempt to create viable hypotheses or to give optimal solutions to a problem. Learning, in fact, very often starts when the student is no longer able to continue by herself and requires the active participation of the teacher (this is one of the ideas of Vygotsky's *zone of proximal development*). This participation may become apparent in terms of questions and clues to redirect the student's attention to certain unattended features of the problem under consideration and that are vital to the attainment of a certain form of mathematical thinking. But it also can result from actively and critically interacting with the teacher while both

---

<sup>1</sup>Knowledge does not spring up in the individual as a result of a direct projection on his brain of the ideas and concepts worked out by preceding generations [Leont'ev \[1978, p. 19\]](#).

teacher and students solve the problem *together*. Of course, such a way of doing cannot be accounted for as an instance of learning in other theories, where the intellectual autonomy of the student plays the role of a prerequisite for learning. For sociocultural theories, however, autonomy is not a prerequisite, but, as already mentioned, its result.

The nuance is in fact subtler, for the idea of autonomy is not taken by sociocultural perspectives as something that develops from within the individual, or as something latent that the subject manages to expand: autonomy is not seen as my capability to do things without the help of others: autonomy is a social relation that I acquire as I engage in social praxes, and as such, is always a commitment to others [Radford \[2008c, 2012\]](#).

Sociocultural approaches to teaching and learning are younger than the other two approaches discussed in this paper. They were introduced in the early 1990s into mathematics education by mathematics educators, such as Ubiratan D'Ambrosio, Alan Bishop, Steve Lerman, and Mariolina Bartolini Bussi. The sociocultural approaches have gained some impetus in the past few years and shed some light on the problem of the cultural nature of mathematics [D'Ambrosio \[2006\]](#) and [Bishop \[1991\]](#), classroom interaction and discourse [M. G. Bartolini Bussi \[1998\]](#) and [Lerman \[1996, 2001\]](#), classroom conceptualization [Radford \[2000, 2008d\]](#) and [Radford, Bardini, and Sabena \[2007\]](#), semiotic mediation [Arzarello and Robutti \[2004\]](#), [M. G. Bartolini Bussi and Mariotti \[1999\]](#), [M. Bartolini Bussi and Mariotti \[2008\]](#), and [Radford \[2005\]](#), and the question of culture and cognition [Radford \[1997a, 2008b,e\]](#).

## 4 A New Trend

In this last section, I want to briefly mention a new trend as observed in the Fifth Congress of the European Society for Research in Mathematics Education (CERME-5, February 22-26, 2007). The European Society for Research in Mathematics Education organizes bianual conferences that are designed to encourage an exchange of ideas through thematic working groups. A few plenary activities take place, yielding most of the space to group work. One of the recurring CERME working groups is the one devoted to theories in mathematics education. For instance, in the CERME-5 conference held in the city of Larnaca, Cyprus, the working group 11 *Different Theoretical Perspectives / Approaches in Research in Mathematics Education* was one of the most popular, which attests to the interest in understanding that which makes theories different. However, the goal of this working group was not just to understand differences, but to seek new forms of linking and connecting current theories. More specifically, the idea was to discuss and investigate theoretical and practical forms of *networking* theories. Most of the papers presented at the meetings of working group 11 will appeared in an issue of the journal *ZDM - The International Journal on Mathematics Education*. As I mention in the commentary paper written

for this ZDM issue [Radford \[2008a\]](#), this new trend consisting of investigating ways of connecting theories is explained to a large extent by the rapid contemporary growth of forms of communication, increasing international scientific cooperation, and the attenuation of political and economic barriers in some parts of the world, a clear example being, of course, the European Community.

This new trend is leading to an enquiry about the possibilities and limits of using several theories and approaches in mathematics education in a meaningful way. The papers presented at the conference provided an interesting array of possibilities.

Depending on the goal, connections may take several forms. [Prediger, Bikner-Ahsbabs, and Arzarello \[2008\]](#) identify some of them, like comparing and contrasting, and define them as follows. In comparing, the goal is to find similarities and differences between theories, while in contrasting theories, the goal is to stress big differences. [Cerulli, Georget, Maracci, Psycharis, and Trgalova \[2008\]](#) is an example of comparing theories, while [Rodríguez, Bosch, and Gascón \[2008\]](#) is an example of contrasting theories. These forms of connectivity are distinguished from others like coordinating and combining. In coordinating theories, elements from different theories are chosen and put together in a more or less harmonious way to investigate a certain research problem. Halverscheid's article (2008) is a clear example of an attempt at coordinating theories, in that, the goal is to study a particular educational problem (the problem of modelling a physical situation) through the use of elements from two different theories (a modelling theory and a cognitive one). In combining theories, the chosen elements do not necessarily show the coherence that can be observed in coordinating connections. It is rather a juxtaposition of theories (see Prediger et al.'s paper, (2008)). [Maracci \[2008\]](#) and [Bergsten \[2008\]](#) furnish examples of combining theories.

At least in principle, comparing and contrasting theories are always possible: given two mathematics education theories, it is possible to seek out their similarities and/or differences. In contrast, to coordinate or to integrate theories, which is another possible form of connection [Prediger, Bikner-Ahsbabs, and Arzarello \[2008\]](#) paper, seems to be a more delicate task.

Connecting theories can, in sum, be accomplished at different levels (principles, methodology, research questions), with different levels of intensity. Sometimes the connection can be strong, sometimes weak. It is still too early to predict how this new trend will evolve. What is clear, in contrast, is that the investigation of integration of theories and their differentiation is likely to lead to a better understanding of theories and richer solutions to practical and theoretical problems surrounding the teaching and learning of mathematics.

**Acknowledgments.** This article is a result of a research program funded by The Social Sciences and Humanities Research Council of Canada / Le Conseil de recherches en sciences humaines du Canada (SSHRC/CRSH).

## References

- F. Arzarello and O. Robutti (2004). "Approaching functions through motion experiments". *Educational Studies in Mathematics (PME Special Issue of Approaching functions through motion experiments) with R. Nemirovsky, M. Borba and C. DiMattia (Eds.)* 57.3. CD-Rom, chapter 1 (cit. on p. 4069).
- M. G. Bartolini Bussi (1998). "Verbal interaction in the mathematics classroom: A Vygotskian Analysis". In: *Language and Communication in the Mathematics Classroom*. Reston, Virginia: National Council of Teachers of Mathematics, pp. 65–84 (cit. on p. 4069).
- M. G. Bartolini Bussi and M. A. Mariotti (1999). "Semiotic Mediation: from History to the Mathematics Classroom". *For the Learning of Mathematics* 19.2, pp. 27–35 (cit. on p. 4069).
- M. Bartolini Bussi and M. A. Mariotti (2008). "Semiotic mediation in the mathematics classroom: Artefacts and signs after a Vygotskian perspective". In: *Handbook of international research in mathematics education (2nd edition)*. New York: Routledge, Taylor and Francis, pp. 746–783 (cit. on p. 4069).
- H. Bauersfeld (1980). "Hidden dimensions in the so-called reality of a mathematics classroom". *Educational Studies in Mathematics* 11, pp. 23–41 (cit. on p. 4059).
- C. Bergsten (2008). "On the influence of theory on research in mathematics education: The case of teaching and learning limits of functions". *ZDM - the International Journal on Mathematics Education* 40.2, pp. 189–199 (cit. on p. 4070).
- A. J. Bishop (1991). *Mathematical enculturation: A cultural perspective on mathematics education*. Dordrecht: Kluwer (cit. on p. 4069).
- G. Brousseau (1988). "Le contrat didactique: Le milieu". *Recherches en Didactique des Mathématiques* 9.3, pp. 309–336 (cit. on p. 4063).
- (1997a). "La théorie des situations didactiques". Cours donné lors de l'attribution du titre de Docteur Honoris Causa de l'Université de Montréal, Montréal (cit. on p. 4061).
  - (1997b). *Theory of Didactical Situations in Mathematics*. Dordrecht: Kluwer (cit. on pp. 4061–4065).
  - (2003). "Glossaire de quelques concepts de la théorie des situations didactiques en mathématiques". Retrieved on January 20, 2007 (cit. on p. 4062).

- M. Cerulli, J. P. Georget, M. Maracci, G. Psycharis, and J. Trgalova (2008). “Comparing theoretical frameworks enacted in experimental research: Telma experience”. *ZDM - the International Journal on Mathematics Education* 39.2, pp. 201–213 (cit. on p. 4070).
- P. Cobb (1985). “An investigation of young children’s academic arithmetic contexts”. *Educational Studies in Mathematics* 18, pp. 109–124 (cit. on p. 4057).
- (1988). “The tension between theories of learning and instruction in mathematics education”. *Educational Psychologist* 23.2, pp. 87–103 (cit. on pp. 4056, 4057).
- (1994). “Where Is the Mind? Constructivist and Sociocultural Perspectives on Mathematical Development”. *Educational Researcher* 23.7, pp. 13–23 (cit. on p. 4059).
- P. Cobb, A. Boifi, K. McClain, and J. Whitenack (1997). “Reflective Discourse and Collective Reflection”. *Journal for Research in Mathematics Education* 28.3, pp. 258–277 (cit. on p. 4060).
- P. Cobb, T. Wood, E. Yackel, J. Nicholls, G. Wheatley, B. Trigatti, and M. Perlwitz (1991). “Assessment of a Problem-Centered Second-Grade Mathematics Project”. *Journal for Research in Mathematics Education* 22.1, pp. 3–29 (cit. on p. 4060).
- P. Cobb and E. Yackel (1996). “Constructivist, Emergent, and Sociocultural Perspectives in the Context of Developmental Research”. *Educational Psychologist* 31.3/4, pp. 175–190.
- P. Cobb, E. Yackel, and T. Wood (1992). “A Constructivist Alternative to the Representational View in Mathematics Education”. *Journal for Research in Mathematics Education* 23.1, pp. 2–33 (cit. on p. 4059).
- U. D’Ambrosio (2006). *Ethnomathematics*. Rotterdam: Sense Publishers (cit. on p. 4069).
- P. Ernest (1991). “Constructivism, the Psychology of Learning, and the Nature of Mathematics: Some Critical Issues”. In: *Proceedings of 15th International Conference on the Psychology of Mathematics Education*. Vol. 2. Assisi, Italy, pp. 25–32 (cit. on p. 4057).
- E. E. Evans-Pritchard (1937). *Witchcraft, Oracles and Magic among the Azande*. Reprinted in 1963. Oxford: Clarendon Press (cit. on p. 4067).
- P. Feyerabend (1987). *Farewell to Reason*. Reprinted in 1994. London: Verso (cit. on p. 4067).
- E. von Glasersfeld (1995). *Radical Constructivism: A Way of Knowing and Learning*. London: The Falmer Press (cit. on p. 4056).
- G. A. Goldin (1990). “Epistemology, Constructivism, and Discovery Learning in Mathematics”. *Journal for Research in Mathematics Education* 4, pp. 31–47 (cit. on p. 4059).
- L. Goldmann (1971). *Immanuel Kant*. London: NLB (cit. on p. 4058).
- S. Halverscheid (2008). “Building a local conceptual framework for epistemic actions in a modelling environment with experiments”. *ZDM - the International Journal on Mathematics Education* 40.2, pp. 225–234 (cit. on p. 4070).
- I. Kant (1894). *Inaugural dissertation*. Original work published in 1770. New York: Columbia College (cit. on p. 4058).

- (2003). *Critique of pure reason*. Original work published in 1781. New York: St. Martin's Press (cit. on p. 4058).
- I. Kidron, A. Lenfant, A. Bikner-Ahsbahs, M. Artigue, and T. Dreyfus (2008). “[Toward networking three theoretical approaches: The case of social interactions](#)”. *ZDM - the International Journal on Mathematics Education* 40, pp. 247–264 (cit. on p. 4065).
- A. N. Leont'ev (1978). *Activity, Consciousness, and Personality*. New Jersey: Prentice-Hall (cit. on p. 4068).
- S. Lerman (1996). “[Intersubjectivity in Mathematics Learning: A Challenge to the Radical Constructivist Paradigm?](#)” *Journal for Research in Mathematics Education* 27.2, pp. 133–150 (cit. on pp. 4060, 4068, 4069).
- (2001). “[The Function of Discourse in Teaching and Learning Mathematics: A Research Perspective](#)”. *Educational Studies in Mathematics* 46, pp. 87–113 (cit. on p. 4069).
- M. Maracci (2008). “[Combining different theoretical perspectives for analyzing students' difficulties in vector spaces theory](#)”. *ZDM - the International Journal on Mathematics Education* 40.2, pp. 265–276 (cit. on p. 4070).
- M.-J. Perrin-Glorian (1994). “Théorie des situations didactiques: naissance, développement et perspectives”. In: *Vingt ans de didactique des mathématiques en France*. Grenoble: La pensée sauvage, pp. 97–147 (cit. on p. 4063).
- M.-J. Perrin-Glorian and M. Hersant (2003). “Milieu et contrat didactique, outils pour l'analyse de séquences ordinaires”. *Recherches en Didactique des Mathématiques* 23 (2), pp. 217–276 (cit. on p. 4063).
- S. Prediger, A. Bikner-Ahsbahs, and F. Arzarello (2008). “[Networking strategies and methods for connecting theoretical approaches: First steps towards a conceptual framework](#)”. *ZDM - the International Journal on Mathematics Education* 40.2, pp. 165–178 (cit. on p. 4070).
- L. Radford (1997a). “L'invention d'une idée mathématique: la deuxième inconnue en algèbre (The invention of a mathematical idea: the second unknown in algebra)”. *Repères – Revue des instituts de Recherche sur l'enseignement des Mathématiques* 28, pp. 81–96 (cit. on p. 4069).
- (1997b). “On Psychology, Historical Epistemology and the Teaching of Mathematics: Towards a Socio-Cultural History of Mathematics”. *For the Learning of Mathematics* 17.1, pp. 26–33 (cit. on p. 4067).
- (2000). “[Signs and meanings in students' emergent algebraic thinking: A semiotic analysis](#)”. *Educational Studies in Mathematics* 42.3, pp. 237–268 (cit. on p. 4069).
- (2005). “The semiotics of the schema. Kant, Piaget, and the Calculator”. In: *Activity and Sign. Grounding Mathematics Education*. New York: Springer, pp. 137–152 (cit. on p. 4069).

- L. Radford (2008a). “[Connecting theories in mathematics education: Challenges and possibilities](#)”. *ZDM - The International Journal on Mathematics Education* 40 (2), pp. 317–327 (cit. on pp. [4055](#), [4067](#), [4070](#)).
- (2008b). “Culture and cognition: Towards an anthropology of mathematical thinking”. In: *Handbook of international research in mathematics education*. 2nd ed. New York: Routledge, Taylor and Francis, pp. 439–464 (cit. on p. [4069](#)).
- (2008c). “Di Sé e degli Altri: Riflessioni su un problema fondamentale dell’educazione (The self and the other: Reflections on a fundamental problem in education)”. *La Matematica e la sua didattica* 22.2, pp. 185–205 (cit. on pp. [4058](#), [4068](#), [4069](#)).
- (2008d). “[Iconicity and Contraction: A Semiotic Investigation of Forms of Algebraic Generalizations of Patterns in Different Contexts](#)”. *ZDM - The International Journal on Mathematics Education* 40.1, pp. 83–96 (cit. on p. [4069](#)).
- (2008e). “The Ethics of Being and Knowing: Towards a Cultural Theory of Learning”. In: *Semiotics in mathematics education: epistemology, history, classroom, and culture*. Rotterdam: Sense Publishers, pp. 215–234 (cit. on p. [4069](#)).
- (2012). “[Education and the illusions of emancipation](#)”. *Educational Studies in Mathematics* 80.1, pp. 101–118 (cit. on p. [4069](#)).
- (2017a). “Mathematics education theories: The question of their growth, connectivity, and affinity”. *La Matematica e la sua Didattica* 25.2, pp. 217–228 (cit. on p. [4055](#)).
- (2017b). “Réflexions sur l’ethnomathématique”. In: *Actes du colloque du groupe de didactique des mathématiques du Québec 2016*. Ottawa: GDM, pp. 168–177 (cit. on p. [4067](#)).
- L. Radford, C. Bardini, and C. Sabena (2007). “Perceiving the General: The Multisemiotic Dimension of Students’ Algebraic Activity”. *Journal for Research in Mathematics Education* 38, pp. 507–530 (cit. on p. [4069](#)).
- E. Rodríguez, M. Bosch, and J. Gascón (2008). “[A networking method to compare theories: Metacognition in problem solving reformulated within the anthropological theory of the didactic](#)”. *ZDM - the International Journal on Mathematics Education* 39.2, pp. 287–301 (cit. on p. [4070](#)).
- B. Sriraman and L. English (2005). “[Theories of Mathematics Education: A global survey of theoretical frameworks/trends in mathematics education research](#)”. *Zentralblatt für Didaktik der Mathematik* 37.6, pp. 450–456 (cit. on p. [4056](#)).
- L. P. Steffe and E. von Glasersfeld (1983). “The construction of arithmetical units”. In: *Proceedings of the 5th annual meeting of the North American Chapter of the International Group of the Psychology of Mathematics Education*. Montreal: Université de Montréal: Faculté de Science de L’Éducation, pp. 292–304 (cit. on p. [4057](#)).
- L. P. Steffe, E. von Glasersfeld, E. Richards, and P. Cobb (1983). *Children’s counting types: Philosophy, theory, and applications*. New York: Praeger Scientific (cit. on p. [4057](#)).

- J. Voigt (1985). “Patterns and routines in classroom interaction”. *Recherches en Didactique des Mathématiques* 6.1, pp. 69–118 (cit. on p. [4059](#)).
- U. Waschescio (1998). “The missing link: Social and cultural aspects in social constructivist theories”. In: *The Culture of the Mathematics Classroom*. Cambridge: Cambridge University Press, pp. 221–241 (cit. on p. [4060](#)).

Received 2018-01-15.

LUIS RADFORD  
ÉCOLE DES SCIENCES DE L'ÉDUCATION  
UNIVERSITÉ LAURENTIENNE ONTARIO  
CANADA  
[lradford@laurentian.ca](mailto:lradford@laurentian.ca)





# IN SEARCH OF THE SOURCES OF INCOMPLETENESS

JAN VON PLATO

## Abstract

Kurt Gödel said of the discovery of his famous incompleteness theorem that he substituted “unprovable” for “false” in the paradoxical statement *This sentence is false*. Thereby he obtained something that states its own unprovability, so that if the statement is true, it should indeed be unprovable. The big methodical obstacle that Gödel solved so brilliantly was to code such a self-referential statement in terms of arithmetic. The shorthand notes on incompleteness that Gödel had meticulously kept are examined for the first time, with a picture of the emergence of incompleteness different from the one the received story of its discovery suggests.

1. PRELUDE TO INCOMPLETENESS. Kurt Gödel’s paper of 1931 about the incompleteness of mathematics, *Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I*, belongs to the most iconic works of the first half of 20th century science, comparable to the ones of Einstein on relativity (1905), Heisenberg on quantum mechanics (1925), Kolmogorov on continuous-time random processes (1931), and Turing on computability (1936). Gödel showed that the quest of representing the whole of mathematics as a closed, complete, and formal system is unachievable.

It has been said repeatedly that Gödel’s discovery has had close to no practical effect on mathematics: the impossibility to decide some questions inside a formal system has not surfaced except in a few cases, such as the question of the convergence of what are known as Goodstein sequences. On the other hand, the effect of methods Gödel developed to prove his theorems has been immeasurable: He invented the idea of a *formal syntax* coded through primitive recursive functions, from which Turing’s idea of machine-executable code arose.

---

I thank Thierry Coquand, Martin Davis, John Dawson, Warren Goldfarb, and Bill Howard for their suggestions and generous encouragement, Maria Hämeen-Anttila for moral support in moments of doubt during the study of Gödel’s manuscripts, and the splendid class of participants in my “Gödel detective” lecture course at the University of Helsinki in 2017, during which the results reported here were achieved. Unpublished works of Kurt Gödel (1934–1978) are Copyright Institute for Advanced Study and are used with permission. All rights reserved by Institute for Advanced Study.

MSC2010: primary 03F40; secondary 01A60, 03-03, 03A05.

Gödel was enormously lucky, or sagacious, to find in 1928 a very precise problem in logic to work with. It occurs in the textbook *Grundzüge der theoretischen Logik* (Basic traits of theoretical logic) by David Hilbert and Wilhelm Ackermann, actually written by Hilbert's assistant Paul Bernays (for which see [Hilbert \[2013, p. 49\]](#)). It gave, for the first time, a complete system of axioms and rules for the logic of the connectives and quantifiers and was the first important step in Hilbert's program that had set as its aim the formalization of mathematical reasoning within a logical language, with proofs of the consistency and completeness of the formalization. Predicate logic is such a language in which proofs in elementary arithmetic can be expressed as *derivations* in a formal system. Derivations, alongside *expressions* of a formal language, form an inductively defined class of objects. Hilbert's inspiration for the formalization of mathematical proofs within a logical language came from the three-volume *Principia Mathematica* 1927 that became known and read in Hilbert's Göttingen from 1917 on.

In [Hilbert and Ackermann 1928], the completeness of predicate logic is given as an important open problem (p. 68):

Whether the axiom system is complete in the sense that all logical formulas correct in each domain of individuals really are derivable in it, is a question still unresolved. One can say only purely empirically that the axiom system has been always sufficient in all applications. The independence of the single axioms has not been studied yet.

Gödel set out to solve the completeness problem, a work that led to his doctoral thesis of 1929, with even proofs of the independence of the axioms and rules included.

I recently found in Gödel's Nachlass an 84 page notebook entitled *Übungsheft Logik* (Logic exercise notebook) based on his reading of Hilbert-Ackermann. Besides predicate logic in which one quantifies only over individuals, the book also presents higher-order logic, with short explanations of how the formalism could be applied in arithmetic and set theory. By the comprehension principle, higher-order predicates correspond to sets, or functions as well over which one can apply the quantifiers, with a very powerful formalism in which to express the principles and postulates used in mathematical theories as a result. One thing that transpires from the *Übungsheft* is that Gödel's main initial objective was to use higher-order logic for a formalization of proofs in arithmetic and set theory. The formalizations are in the "logician" tradition of Gottlob Frege and Russell ( see [von Plato \[2018\]](#) for a detailed account).

Another, astonishing feature of the *Übungsheft* is the way Gödel overcomes the main difficulty of the axiomatic logic of Frege and Russell which is that formal derivations are, as a practical matter, impossible to construct in axiomatic logic. With no explanation or hesitation whatsoever, he starts to use a *system of natural deduction* in place of the hopelessly clumsy axiomatic calculus for his formal proofs, in a format in which the formulas

follow in a linear vertically arranged succession. Gerhard Gentzen is generally considered to be the inventor of natural deduction, a logical calculus in which derivations are presented in a tree-form that allows for a deep analysis of their structure. His way to natural deduction is detailed out on the basis of his manuscripts in the book *Saved from the Cellar* [von Plato 2017a].

Gödel changed his objective from the actual formalization of mathematics as in the *Übungsheft* into a study of the properties of any such formalization: On 6 July 1929, Gödel handed in a brilliant doctoral thesis with the title *Über die Vollständigkeit des Logikkalküls* (On the completeness of the calculus of logic). A shorter version got published in 1930.

Gödel used the term *completeness* (Vollständigkeit) for the case of predicate logic. The contrary was not Unvollständigkeit, but *undecidability* (Unentscheidbarkeit). In the 1931 article, he uses “formal undecidability,” even in the title and writes then specifically of a sentence  $A$  “axiomatically undecidable,” i.e., undecidable or unsolvable in an axiomatic system in the sense of unprovability of both  $A$  and its negation  $\neg A$  within the system. There is some danger of confusing this term with another notion of undecidability, the one of Hilbert’s *Entscheidungsproblem*, as found in Turing’s 1936 title. The two senses of undecidability have a somewhat tricky relation with all four combinations possible: Classical propositional logic is decidable and complete, but if you leave out one axiom you get an incomplete though still decidable axiom system. Similarly, classical predicate logic is complete but undecidable, and Peano arithmetic incomplete and undecidable.

2. THE SOURCES OF INCOMPLETENESS. Gödel was a maniacal keeper of notebooks in which he recorded his thoughts, from his earliest school years on, and there is an enormous amount of material left behind and kept in Princeton. These notebooks are written in the Gabelsberger shorthand that was regularly taught at schools at the time, but rendered obsolete since its substitution by the “unified shorthand” in 1925. Work on these archival sources was done in connection with the publication of the third of Gödel’s five-volume *Collected Works*, which led to the incorporation of two very important manuscripts in the volume: a lecture on the consistency of the continuum hypothesis of December 1939, and another on Gödel’s well-known *functional interpretation* of arithmetic in April 1941. Since those days, with publication in 1995, Gödel’s notebooks have lain dormant except for his collection of fifteen philosophical notebooks, the *MaxPhil* series (Philosophical Maxims).

I started to work with the Gödel notes on 20 March, 2017, after a volume of shorthand notes by Gödel’s younger contemporary Gentzen, the mentioned *Saved from the Cellar*. Here I report on my findings that concern the sources of Gödel’s work on incompleteness, with a typewritten manuscript of *Über formal unentscheidbare Sätze der Principia*

*Mathematica und verwandter Systeme* submitted on 17 November and published in March 1931.

The Gödel archives contain three suites of notes in preparation of the incompleteness paper. The third version comes close to the published paper and bears the cover title *Unentsch. unrein*. This text comes first in the microfilm and is in 39 pages. There follow about the same number of additional pages, some of which with remarkable connections to the main text. A second notebook follows that has the same title, this time written inside, and it is even a bit longer, with a break and what seems a new start in the middle. Finally, there is a third notebook with no title and with the first preserved notes on incompleteness, some 45 pages followed by a dozen pages that are very similar to the introduction of the printed paper.

3. THE CRUCIAL POINT OF THE COMPLETENESS PROOF. Gödel's proof of completeness for the "narrower functional calculus," i.e., first-order classical predicate logic, has disjunction, negation, and universal quantification as the basic notions. The simplest case of quantification is the formula  $\forall x F(x)$  with  $F$  propositional. Gödel states in a shorthand passage that if such a formula is "correct," i.e., becomes true under any choice of domain of individuals and relations for the relation symbols of the formula, then the instance *with a free variable  $x$*  must be a "tautology" of propositional logic.<sup>1</sup> In the usual "Tarski semantics" that is—unfortunately—included in almost every first course in logic, the truth of universals is explained by the condition that *every instance* be true, an explanation that with an infinite domain of objects leads to circles.

In Gödel, in contrast, with the free-variable formula  $F(x)$  a tautology, it must be provable in propositional logic by the completeness of the latter, a result from Bernays' *Habilitationschrift* of 1918 and known to Gödel from Hilbert-Ackermann. That book is also the place in which the rules of inference for the quantifiers appear for the first time in an impeccable form (p. 54, with the acknowledgment that the axiom system for the quantifiers "was given by P. Bernays"). With the free-variable formula  $F(x)$  provable in propositional logic, the rule of universal generalisation gives at once that even  $\forall x F(x)$  is derivable. The step is rather well hidden in Gödel's proof in the thesis that proceeds in terms of satisfiability. At one point, he moves to provability of a free-variable formula, then universally quantified "by 3," the number given for the rule of generalisation.

Gödel's profound understanding of predicate logic, especially the need for rules of inference for the quantifiers without which no proof of completeness is possible, is evident through comparison: Rudolf Carnap, whose course he followed in Vienna in 1928, published in 1929 a short presentation of Russell's *Principia*, the *Abriss der Logistik*, but one searches in vain for the quantifier rules in this booklet. Other contemporaries who failed

<sup>1</sup> My notes are incomplete at this point and I have so far not found again this passage in Gödel's manuscripts.

in this respect include Ludwig Wittgenstein and Alfred Tarski. The former was a dilettante in logic who thought that truth-tables would do even for predicate logic. With the latter, no trace of the idea of the provability of universals through an arbitrary instance is found in his famous tract on the concept of truth of 1934. We shall see that Gödel was way ahead of him in understanding these matters by the summer of 1930.

Gödel's thesis, but not its short published version of 1930, contains a deep remark by which the proof of completeness cannot be finitary because such a proof would give a decision method for predicate logic.

4. ENCOUNTER IN KÖNIGSBERG. There exists a very short and readable lecture about completeness in Gödel's hand, namely the one he gave in a conference in Königsberg in early September 1930. Close to the end of that lecture, we find the following passage [Gödel 1930c, p. 28]:

If one could prove the completeness theorem even for the higher parts of logic (the extended functional calculus), it could be shown quite generally that from categoricity, definiteness with respect to decision [Entscheidungsdefintheit] follows. One knows for example that Peano's axiom system is categorical, so that the solvability of each problem in arithmetic and analysis expressible in the *Principia Mathematica* would follow. Such an extension of the completeness theorem as I have recently proved is, instead, impossible, i.e., there are mathematical problems that can be expressed in the *Principia Mathematica* but which cannot be solved by the logical means of the *Principia Mathematica*.

It is clear from these remarks that Gödel's first thought was to extend the completeness result to higher-order logic, a point emphasised in [Goldfarb 2005]. The above is an indication of his way to the first incompleteness theorem from the time when the actual work was done, not later reconstruction.

The second version of the incompleteness paper has, after some fifteen pages, the title "Meine Damen und Herren!" Then comes the text of the Königsberg lecture on completeness in shorthand. The ending is:

I have succeeded, instead [of extending the completeness theorem to higher-order logic], in showing that such a proof of completeness for the extended functional calculus is impossible or in other words, that there are arithmetic problems that cannot be solved by the logical means of the PM even if they can be expressed in this system. These things are, though, still too little worked through to go into more closely here.

The last sentence reads in German: “Doch sind diese Dinge noch zu wenig durchgearbeitet um hier näher auf einzugehen.” In the typewritten version, we read somewhat differently about his proof of the failure of completeness:

In this [proof], the reducibility axiom, infinity axiom (in the formulation: there are exactly denumerable individuals), and even the axiom of choice are allowed as axioms. One can express the matter also as: The axiom system of Peano with the logic of the PM as a superstructure is not definite with respect to decision. I cannot, though, go into these things here more closely.

The German is: “Auf diese Dinge kann ich aber hier nicht näher eingehen.” Then this last sentence is cancelled and the following written: “Doch würde es zu weit führen, auf diese Dinge näher einzugehen” (It would, though, take us too far to go more closely into these things). It would seem that matters concerning the incompleteness proof had cleared in Gödel’s mind between the writing of the shorthand text for the lecture and the typewritten version. This must have been in the summer of 1930.

Just a few pages before the Königsberg outbreak, Gödel writes that the formally undecidable sentences have “the character of Goldbach or Fermat,” i.e., of universal propositions that can be refuted by a numerical counterexample. A formally undecidable proposition  $\forall x F(x)$  can have each of its numerical instances  $F(n)$  provable, but still, addition of the negation  $\neg \forall x F(x)$  does not lead to an inconsistency. Were the free-variable instance  $F(x)$  provable, universal generalisation would at once give a contradiction.

Among Gödel’s audience in Königsberg sat Johann von Neumann, who reacted at once and wanted more explanations. Gödel gave such in a discussion among the two and most likely during his stay in Berlin immediately after. The most detailed account of these events is [Wang 1996], section “Some facts about Gödel in his own words” [*ibid.*, p. 82–84]:

I represented real numbers by predicates in number theory and found that I had to use the concept of truth to verify the axioms of analysis. By an enumeration of symbols, sentences, and proofs of the given system, I quickly discovered that the concept of arithmetic truth cannot be defined in arithmetic.

...

Note that this argument can be formalised to show the existence of undecidable propositions without giving any individual instances.

Von Neumann suggested in the discussion to transform undecidability “into a proposition about integers.” Gödel then found “the surprising result giving undecidable propositions about polynomials.”

Von Neumann lectured from late October 1930 on in Berlin on “Hilbert’s proof theory” of which Carl Hempel, later a very famous philosopher, has recollected the excitement

created, even evidenced by contemporary letters for which see [Mancosu 1999]. The account is [Hempel 2000, pp. 13–14]:

I took a course there with von Neumann which dealt with Hilbert's attempt to prove the consistency of classical mathematics by finitary means. I recall that in the middle of the course von Neumann came in one day and announced that he had just received a paper from... Kurt Gödel who showed that the objectives which Hilbert had in mind and on which I had heard Hilbert's course in Göttingen could not be achieved at all. Von Neumann, therefore, dropped the pursuit of this subject and devoted the rest of the course to the presentation of Gödel's results. The finding evoked an enormous excitement.

These are later recollections; for example, it is known that von Neumann got the proofs of Gödel's paper around the tenth of January 1931. The lectures of late 1930 were based on other sources to be presented below.

Jacques Herbrand was born in 1908 and received his education at the prestigious *Ecole normale supérieure* of Paris. He finished his thesis *Recherches sur la théorie de la démonstration* at the precocious age of 21 in the spring of 1929. He went to stay for the academic year 1930–31 in Germany, first Berlin from October 1930 on, then from late spring 1931 to July in Hamburg and Göttingen. These stays were in part prompted by his work on algebra, where Emil Artin in Hamburg and Emmy Noether in Göttingen were the leading figures.<sup>2</sup>

There is a letter of Herbrand's of 28 November 1930 to the director of the *Ecole normale* Ernest Vessiot in which he mentions von Neumann's "absolutely unexpected results," then writes that for now he will tell about the

extremely curious results of a young Austrian mathematician who succeeded in constructing arithmetic functions  $Pn$  with the following properties: one calculates  $Pa$  for each number  $a$  and finds  $Pa = 0$ , but it is impossible to prove that  $Pn$  is always zero.

Gödel's account, as reported by Wang, suggests that he had found this result right after the Königsberg meeting; it is further clear that he must have explained it to von Neumann during his visit to Berlin right after.

Eight days before Herbrand's letter, von Neumann had written to Gödel about his proof:

It can be expressed in a formal system that contains arithmetic, on the basis of your considerations, that the formula  $1 = 2$  cannot be the endformula in a proof that starts from the axioms of this system—and in this formulation in fact a formula of the formal system mentioned. Let it be called  $\mathfrak{A}$ .

<sup>2</sup>[von Plato 2017b], section 8.3 on two "Berliners" contains a detailed account of Herbrand's stay in Germany and his relation to von Neumann.



...

I show now:  $\mathfrak{W}$  is always unprovable in systems free of contradiction, i.e., a possible effective proof of  $\mathfrak{W}$  could certainly be transformed into a contradiction.

Gödel must have explained to von Neumann the essential point, not just a blunt statement of incompleteness, namely that provability of a formula in a system can be expressed as a formula of that system, here provability of  $1 = 2$ .

Von Neumann writes next that if Gödel is interested, he would send the details once they are ready for print. He asks further when Gödel's treatise will appear and when he can have proofs, with the wish to relate his work "in content and notation to yours, and even the wish for my part to publish sooner rather than later."

5. GÖDEL'S LOST REPLY TO VON NEUMANN'S LETTER. Gödel's final shorthand version for his incompleteness paper occupies the first 39 pages of a notebook. It begins very closely the way the typewritten version does. Even footnotes are numbered consecutively until number 29 on page 24 of the manuscript. The impressive list of 45 recursive relations in the published paper matches a similar list of 43 items, some ten pages, followed by the upshot of the laborious work in the form of a theorem:

*VI. Every recursive relation is arithmetic.*

After the text proper of the manuscript for the article ends, there are two attempts at a formulation of a title, like this:

On the existence of undecidable mathematical propositions in the system of  
*Principia Mathematica*

On unsolvable mathematical problems in the system of *Principia Mathematica*

There follow five pages with formulas, recursive definitions of functions, elementary computations, and a stylish layout for a lecture on the completeness of predicate logic given in Vienna on 28 November. Next the title "Lieber Herr von Neumann" hits the eye, with the following letter-sketch:<sup>3</sup>

---

<sup>3</sup> A word about the nature of shorthand sources is in place here: The transcription of shorthand is by the very nature of the script, with missing endings of words and abrupt shortenings—a single letter can stand for different words that have to be figured out from the context—also error-bound interpretation and guesswork. There are in addition uncertainties for reasons such as faded sources, badly written or heavily cancelled passages, etc. I have no pretense to a grammarian's exact reading, word for word, but offer my English translations as accounts of what Gödel wrote down about 87 years ago, in the hope that they appear consonant with Gödel's thought, with the suggestion to anyone who should like to quote them to give their own interpretation of the text. At places, I

Dear Mr von Neumann

Many thanks for your letter of [20 November]. Unfortunately I have to inform you that I am already since about three months in possession of the result you communicated. It is also found in the attached offprint of a communication to the Academy of Sciences. I had finished the manuscript for this communication already before my departure for Königsberg and had presented it to Carnap. I gave it over for publication in the *Anzeiger* of the Academy on 17 September. [Cancelled: The reason why I didn't make any presentation [written heavily over: didn't tell anything] of the above result is that the precise proof is not suited to oral communications and an approximate indication could easily arouse doubts about the correctness...that would not convince]. As concerns the publication of this matter, there will be given only a shorter sketch of the proof of impossibility of freedom from contradiction in the *Monatshefte* that will appear in the beginning of 1931 (the main part of this treatise will be filled with the proof of existence of undecidable sentences). The detailed carrying through of the proof appears in a *Monatsheft* only in July or August. I can send you proofs in a few weeks.

I shall include a part of my work that concerns the proof of freedom from contradiction so that you can state to what extent your proof matches mine.

The carrying out of the proof appears together with my proof of undecidability in the next volume of the *Monatshefte*. I didn't want to talk about it further provisionally because this thing (even more than the proof of undecidability) must arouse doubt about its executability before it is laid out in a concrete way.

There are eight pages between the first and second versions of the letter, filled with Gödel's attempts at formulating the second incompleteness theorem in various ways and how it should be proved, until a second letter sketch:

Dear Mr von Neumann!

Hearty thanks for your letter of 20-/XI. The result of which you write to me is known to me since already about three months, but I didn't want to talk anything about it before I had done it in a print-ready form. I send you enclosed an offprint in which the proved theorem gets expressed. The manuscript of

---

have left a question mark in place of a word or two that I failed to read. A sentence can be informative even if one doesn't see what precise verb is used in it, say. There are very many cancelled passages in the letter sketches; I have included just some of these details, with the aim of a readable result in mind, as close to something Gödel may have intended to send to von Neumann.

this communication to the Academy was finished already before my departure to Königsberg and presented it to Carnap. I gave it over for publication in September. The carrying through of the proof will appear together with the proof of undecidability in a near *Monatsheft* (beginning of 1931). I shall have proofs of this work in a few weeks and will then send them to you immediately.

Now to the matter itself. A basic idea of my proof can be described (quite roughly) like this. The sentence  $A$  that I have put up and that is undecidable in the formal system  $S$  asserts its own unprovability and is therefore correct. If one analyses precisely how this undecidable sentence  $A$  could still be metamathematically decided, it appears that this is possible only under the condition of the freedom of contradiction of  $S$ . That is, it is strictly taken not  $A$  but  $W \rightarrow A$  that is proved ( $W$  means the proposition:  $S$  is free from contradiction). The proof of  $W \rightarrow A$  lets itself be carried through, though, within the system  $S$ , so that if even  $W$  is provable in  $S$ , then also  $A$  which contradicts the undecidability of  $A$ .

As concerns the meaning of this result, my opinion is that *only* the impossibility of a proof of freedom from contradiction for a system *within this system* is thereby proved. For the rest, I am fully convinced that there is [cancelled: a finite] an intuitionistically unobjectionable proof of freedom of contradiction for classical mathematics [added above: and set theory], and that therefore the Hilbertian point of view has in no way been refuted. Only one thing is clear, namely that this proof of freedom from contradiction is in any case far more (?) complicated than had been assumed so far.

As concerns the question that remains, my opinion is that exists no formal system in which all [cancelled: intuitionistically unobjectionable constructive] finite proofs would be expressible.\*<sup>4</sup> Still, I would like very much to hear about your contrary argument concerning the matter. I would be further interested whether your proof is built on the same thought as mine, something I hope all the same from what you intend in relation to publication, namely that you relate your work to mine.

Unfortunately, nothing seems to come of my travel to Berlin this year.

In the hope of a swift reply, I remain with

---

<sup>4</sup> [Ed: The asterisk directs to an addition at the end of the letter sketch:]

\* From the treatise of P. Bernays on "Philosophie der Mathematik und die hilbertsche Beweistheorie" in the *Blätter für Deutsche Philosophie*, volume 4, issue 3/4, 1930, I gather that this is also the view of Hilbert and Bernays (cf. what is said on page 366).

best wishes, yours sincerely

6. HERBRAND'S TESTIMONY. Herbrand had explained the post-Königsberg statement of incompleteness in terms of polynomials to Vessiot, and five days later he writes another letter, to his friend Claude Chevalley, in the worst handwriting imaginable, but full of sparkling ideas that seem to spring from nothing. In the letter, Herbrand explains von Neumann's argument for the second incompleteness theorem as follows:

Let  $T$  be a theory that contains arithmetic. Let us enumerate all the demonstrations in  $T$ ; let us enumerate all the propositions  $Q$   $x$ ; and let us construct a function  $P$   $x$   $y$   $z$  that is zero if and only if demonstration number  $x$  demonstrates  $Q$   $y$ ,  $Q$  being proposition number  $z$ .

We find that  $P$   $x$   $y$   $z$  is an effective function that one can construct with arithmetic functions that are easily definable.

Let  $\beta$  be the number of the proposition  $(x) \sim P$   $x$   $y$   $y$  ( $\sim$  means: not); let  $A$   $x$  be the proposition  $\sim P$   $x$   $\beta$   $\beta$

$A$  the proposition  $(x).A$   $x$  ( $A$   $x$  is always true)

$A$   $x$ , equivalent to: demonstration  $x$  does not demonstrate the proposition  $\beta$ ;  
so

$A$   $x$ .  $\equiv$  . demonstration  $x$  does not demonstrate  $A$

Let us enunciate:

$A$   $x$ .  $\equiv$  .  $\sim D(x, A)$

1)  $A$   $x$  is true (for each cipher  $x$ ); without it  $D(x, A)$  would be true; therefore  $A$ ; therefore  $A$   $x$ ; therefore  $\sim D(x, A)$ .

2)  $A$  cannot be demonstrated

for if one demonstrates  $A$ ,  $A$   $x$  would be false; contradiction.

Therefore:  $A$  0,  $A$  1,  $A$  2 ... are true

$(x)A$   $x$  cannot be demonstrated in T

Next in Herbrand's letter comes von Neumann's striking addition to Gödel's first theorem: with  $D(x, A)$  standing as above for: proof number  $x$  demonstrates proposition  $A$ , Herbrand writes in the letter the magic formulas:

3)  $\sim A \rightarrow D(x, A)$  et  $D(z, \sim A)$

therefore:  $\sim(D(x, A)$  et  $D(z, \sim A)) \rightarrow A$

The conclusion, for the unprovable proposition  $A$ , is that “if one proves consistency, one proves  $A$ ”: Consistency requires that for any proposition  $A$ , there do not exist proofs of  $A$  and  $\sim A$ , i.e.,  $\sim \exists x \exists z (D(x, A) \text{ et } D(z, \sim A))$ , or in a free-variable formulation, for each  $x$  and  $z$ ,  $\sim (D(x, A) \text{ et } D(z, \sim A))$ .

At the time Herbrand wrote to Vessiot, 28 November, the “absolutely unexpected results” he alludes to are perhaps an indication of von Neumann’s version of the second theorem. By 29 November, von Neumann has read Gödel’s letter of reply and that shows in Herbrand’s letter to Chevalley of 3 December. Gödel had explained to von Neumann that the second theorem is proved by first showing an implication *within* the formal system. The details are found in the interim pages between the two letter sketches—with even references to the incompleteness paper. Here  $\mathcal{K}$  is any “recursive consistent class” of formulas:

Let us now turn back to the undecidable proposition *17Gen r*. The proposition that  $\mathcal{K}$  is free from contradiction will be denoted by  $Wid(\mathcal{K})$  for the proof that *17Gen r* is unprovable, and only the freedom of contradiction of  $\mathcal{K}$  is used (cf. 1.) on page 30) so we have

$$Wid(\mathcal{K}) \rightarrow \overline{Bew_{\mathcal{K}}}(17Gen r)$$

If now  $Wid(\mathcal{K})$  were provable within the system, also the unprovable sentence  $\overline{Bew_{\mathcal{K}}}(17Gen r)$  would, which is impossible.

In von Neumann’s second letter to Gödel, of 29 November, he writes:

I believe I can reproduce your sequence of thoughts on the basis of our communication and can therefore tell you that I used a somewhat different method. You prove  $W \rightarrow A$ , I show independently the unprovability of  $W$ , though with a different kind of inference that likewise copies the antinomies.

Von Neumann’s proof idea brings to mind Gödel’s early formulations of the unprovability of consistency. More cannot be said unless notes for the course are found somewhere. The lectures must have been widely attended, but I have been able to secure only Hempel, Herbrand, and B. H. Neumann, and very likely Gerhard Gentzen as participants.

There is a third letter of von Neumann’s of 12 January 1931, after he had received the page proofs of Gödel’s article, in which he sketches what he describes as a “somewhat shorter carrying out of the unprovability of freedom from contradiction.”

7. GÖDEL IN PANIC. The sequence of events in and around Gödel’s two sketches of letters is psychologically interesting. He was of course worried about von Neumann’s plans: First he wants to assure von Neumann that he had both results, even mentioning Carnap

as witness and quoting 17 September as the date he sent in the short note to press. He writes that he will copy a part of his manuscript for the incompleteness paper, about the second theorem, etc. Then come eight pages of attempts at a satisfactory formulation, and the second letter sketch in which just the proofs of the incompleteness article are promised once they arrive.

The pages between Gödel's two letter sketches to von Neumann are his notes for section 4 of his incompleteness paper. An inspection of his typewritten manuscript shows that the last three lines of page 41 have been cancelled. They contain the beginning of his closing paragraph as in the shorthand manuscript. Pages 42–44 contain the added section 4. The first proofs have a "I" added in the end of the title, a paragraph that explains the second theorem added at the end of the introduction, and a long footnote on the second theorem added in another place. The original proofs have no mention at all of the second theorem before section 4 that Gödel wrote, and must have taken directly to the printer's, some time after he had received von Neumann's letter.

A shadow is cast on Gödel's great achievement; there is no way of undoing the fact that Gödel played a well-planned trick to persuade von Neumann not to publish. In his letter of reply, he reproduced details from section 4, freshly written after von Neumann's letter, but he also included his short note of October 1930 that contains a statement of the second theorem. The latter would have been enough, but Gödel panicked at the prospect of von Neumann publishing his second theorem. The writing is quite nervous, with cancellations and additions all over. Moreover, the first proofs that reveal his trick must have caused him quite a stress; nothing he could send to von Neumann who would have wondered why the magnificent second incompleteness theorem is not even mentioned in the lengthy introduction. He got page proofs for the article only around the tenth of January.

Concerning the October 1930 one-page notice to the Vienna academy, the last page of the shorthand manuscript instructs to add to page 1 a reference to this note. There is in the title (!) of Gödel's article a footnote that points to it, without further explanations. The microfilms contain a typewritten copy with a stamp "Akademie der Wissenschaften in Wien, Zahl 721/1930 eingefangt: 21.X.1930." The wording of "Satz II" is well known:

Even when one allows in metamathematics all the logical means of the *Principia Mathematica* (especially therefore the extended functional calculus with the axiom of reducibility or without ramified type theory and the axiom of choice), there is *no proof of freedom from contradiction* for the system *S* (and even less if one restricts the means of proof in some way). Therefore, a proof of freedom from contradiction of the system *S* can be carried through only by methods that lie *outside* the system *S*, and the case is analogous for other formal systems, say the Zermelo-Fraenkel axiom system for set theory.

Having sent the note to von Neumann, it is clear that the latter had no new result to publish, and there would have been no need for Gödel to change anything, at most mention the results in the short notice. The formulation also confirms what I said above, namely that Gödel's early metamathematics used strong methods. Moreover, the printed text mentions  $\omega$ -consistency, but in the manuscript and in the notes before Königsberg, Gödel always wrote  $\aleph_0$ -consistency, the latter a distinctly set-theoretic notation.

The typewritten manuscript with the typesetters' leaden fingerprints on it contains three lines at the end of page 41, and the rest exists only in his shorthand:

To finish, let us point at the following interesting circumstance that concerns the undecidable sentence  $A$  put up in the above. By a remark made right in the beginning [page 41 ends here, in the shorthand the letter  $S$  is used instead of  $A$ ],  $S$  claims its own unprovability. Because  $S$  is undecidable, it is naturally also unprovable. Then, what  $S$  claims is correct. Therefore the sentence  $S$  that is undecidable in the system has been decided with the help of metamathematical considerations. An exact analysis of this state of affairs leads to interesting results that concern a proof of freedom from contradiction of the system  $P$  (and related systems) that will be treated in a forthcoming continuation of this work.

Gödel shows here a cautiousness the editor of his *Collected Works* Sol Feferman liked to emphasise about him, just "interesting results" about consistency. The thought of von Neumann publishing the second theorem must have haunted him and led to the hasty addition of a section on results so far "zu wenig durchgearbeitet" as he put his closing words in the Königsberg lecture. In fact, Gödel was unable to prove the second theorem to his satisfaction and no "Part II" of the incompleteness paper ever appeared, neither do the shorthand notes suggest any such work even in manuscript form.

8. INCOMPLETENESS BEFORE THE SECOND THEOREM? We have now looked ahead from the Königsberg meeting; let's look back also. Among Gödel's preserved notes for the incompleteness article, the last one is, as noted, very close to the printed paper. The first of these notebooks is a rather carefully composed set that seems to have been written for an article before Gödel had found the second theorem, so before August 1930 and before the part of the second notebook that was written down before the notes for the Königsberg lecture. This timing is in accordance with what Gödel wrote to von Neumann toward the end of November, namely that he had been in possession of the second theorem for some three months, and with what he told Hao Wang in 1976–77. Therein we find Gödel state that he discovered his second theorem "shortly after the Königsberg meeting." Therefore,

anything that precedes the Königsberg lecture notes in the second suite of notes for incompleteness, must be before the second theorem about the unprovability of consistency had surfaced.

The first version of the incompleteness article opens with the words:

The question whether every mathematical problem is solvable, i.e. whether for every mathematical proposition  $A$  either  $A$  or non  $A$  is provable, was so far devoid of a concrete sense, because the words “mathematical proposition” and “mathematically provable” had not been made precise. The opinions of various mathematicians diverge strongly on this point, as is shown sufficiently by the discussions over the axiom of choice and the law of excluded middle. The way to make for precision that is at the basis of the investigation at hand is essentially the one given in the *Principia Mathematica*.

A detailed examination of this early incompleteness work has to await another occasion. Let us note two interesting remarks therein. Page 32 has:

It is easy to convince oneself by complete induction about the correctness of the following theorem:

*Every provable formula is true* because the axioms are obviously true and this property is not destroyed by the rules of inference. This result can be proved, though, only with the help of the axiom of choice.

The passage refers to a formal system that contains higher-order logic. On page 18, the nature of metamathematics is described:

No limitations in the means of proof are required. One can use all the theorems and methods of analysis, set theory, etc in metamathematical proofs. A proof of a metamathematical theorem conducted in such a way is comparable to a proof in analytical number theory.

At the end of the more than forty pages of notes, there is a text for an introduction that begins with:

In what follows, a proof is sketched in coarse outline by which Peano’s axioms with the logic of the *Principia Mathematica* (natural numbers as individuals) don’t form any system definite with respect to decidability, even allowing the axiom of choice. In other words, there are in the system unsolvable problems, even of a relatively simple structure.

The second set of Gödel’s notes for incompleteness is, as its first page tells, “a provisional version.” It gets a fairly good dating by the presence of the Königsberg lecture in it:



anything before the lecture text is before the Königsberg meeting. An “exact definition” of the notion of truth is given in the earlier part and the theorem stated that all provable formulas are true. A proposition is then constructed by arithmetic coding that states its own unprovability. If it is provable, it is true, so must in fact be unprovable.

The critical point in the truth definition is with the universal quantifier:

$(x)F(x)$  shall be called true when and only when for every number  $n$  [of right type],  $F(n)$  is true. This definition fails in that it presupposes that there are names for all classes and relations which certainly is not the case (there are especially only denumerably many names).

...

When one asks by what means not contained in the system  $S$  undecidability was concluded, the answer can be only: through the definition of truth that extends type theory into the transfinite.

A footnote tells: “The idea of such a definition has been expressed [cancelled: simultaneously] independently by Mr A. Tarski of Warsaw.” Tarski had lectured in Vienna in February 1930 and a letter of Gödel’s of 2 April 1931 to Bernays even recollects a discussion on the topic with Tarski. It seems clear that at that time, Gödel was trying to prove the completeness of higher-order logic and needed a truth definition for the soundness part. The other direction failed, though.

Gödel has seen clearly the critical point, namely that the syntactic condition of provability of  $F(x)$  with a free variable suffices for  $(x)F(x)$  in predicate logic, whereas universal quantification in higher-order logic becomes a transfinite notion.

Soon after the Königsberg lecture notes break in the shorthand, Gödel saw that one can restrict the methods used in metamathematics. This change was prompted by von Neumann’s suggestion in Königsberg. Close to fifty years later, Gödel regretted not having mentioned the suggestion [Wang 1996, p. 84]. The proof of the first theorem in the final version is, as Gödel emphasised, carried out “constructively,” and he planned undoubtedly to do the same with the second theorem. There would then be two versions each of the two incompleteness theorems.

9. GÖDEL’S SOURCES, CITED AND UNCITED. Later in his life, Gödel gave various explanations of how he found the incompleteness results. He often repeated that he was thinking of self-referential statements, as in the liar paradox: *This sentence is false*. Substituting unprovable for false, one gets a statement that expresses its own unprovability. The explanation is good, and indeed given as a heuristic argument in Gödel’s 1931 paper, but it gives little clue as to how one would start thinking along such lines in the first place. Another explanation was that he tried to prove the consistency of analysis relative to first-order

arithmetic. This explanation has an affinity with the early formulations of incompleteness.

The middle version, where the definition of truth is given, makes the following comment after the proposition that states its own unprovability is shown to be true and therefore unprovable:

One recognises a close connection of this proof to the Richard antinomy and it can be expected that even other epistemic antinomies can be reorganized into analogous proofs, something that actually is the case.

Hilbert-Ackermann contains a lengthy discussion of such paradoxes and the analogy was therefore fresh in Gödel's mind. The effect of starting in the 1931 paper with the heuristic analogy gave the impression—whether planned or accidental—that the paradoxes were his way to the result, an impression that created an unprecedented aura of genius around the discovery and around him, shared by von Neumann and everyone else who read his finished paper.

Gödel's meticulously kept notes and other material point at interesting circumstances that concern his discovery of the undecidable sentences, to be treated in a forthcoming continuation of this work. Let me just refer to a couple of unmentioned sources: Gödel had begun work on incompleteness in the summer of 1930 by [*ibid.*, p. 82]; I would say perhaps May). Gödel's library request cards show that he had taken out in April Fraenkel's *Einleitung in die Mengenlehre* in which the question of completeness of mathematical theories is discussed. The most poignant remark is that “there should be nothing absurd in imagining that the unsolvability of a problem could even be *proved*” (p. 235).

On 13 May and again right after Königsberg on 12 September while in Berlin, Gödel borrowed an obscure Norwegian journal issue, Skolem's separately published 49 page “Über einige Grundlagenfragen der Mathematik,” of the previous year. There Skolem states a version of the “Skolem paradox,” namely that the theorems of a truly formal system are denumerable, indeed they can be ordered lexicographically, but that the properties of natural numbers cannot be in that way ordered, by which (p. 269):

It would be an interesting task to show that every collection of propositions about the natural numbers, formulated in predicate logic, continues to hold when one makes certain changes in the meaning of “numbers.”

Gödel wrote down detailed summaries of the works he read. In his three page summary of Skolem's paper, we read for Skolem's §7, with the condition  $ah - bk = 1$  pointing at the unique decomposition into prime elements in principal ideal domains:

§7 Example of a domain that is not isomorphic with the number sequence even if it is an integral domain and even if for every two relatively prime  $h, k$ ,  $ah -$

$bk = 1$ . Conjecture that the number sequence is not at all characterisable by propositions of first-order logic.

## References

- P. Bernays (1918). “Beiträge zur axiomatischen Behandlung des Logik-Kalküls”. Manuscript Hs. 973:193, Bernays collection, ETH-Zürich, printed in [Hilbert 2013].
- R. Carnap (1929). *Abriss der Logistik*. Springer.
- J. Dawson Jr. (1997). *Logical Dilemmas. The Life and Work of Kurt Gödel*. A K Peters, Ltd., Wellesley, MA, pp. xiv+361. MR: [1429389](#).
- A. Fraenkel (1928). *Einleitung in die Mengenlehre*. 2nd ed. Springer.
- K. Gödel (1930a). “Die Vollständigkeit der Axiome des logischen Funktionenkalküls”. *Monatsh. Math. Phys.* 37.1, pp. 349–360. MR: [1549799](#).
- (1930b). “Einige metamathematische Resultate über Entscheidungsdefinitheit und Widerspruchsfreiheit”. *Anzeiger Akad. Wiss. Wien* 67, pp. 214–215.
- (1930c). “Vorlesung über Vollständigkeit des Funktionenkalküls”. In: vol. III. First printed in [Gödel 1986], pp. 16–29 (cit. on p. 4079).
- (1931). “Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I”. *Monatsh. Math. Phys.* 38.1, pp. 173–198. MR: [1549910](#).
- (1986). *Collected works. Vol. I*. Publications 1929–1936, Edited and with a preface by S. Feferman. Oxford University Press, New York, pp. xvi+474. MR: [831941](#) (cit. on p. 4092).
- W. Goldfarb (2005). “On Gödel’s way in: the influence of Rudolf Carnap”. *Bull. Symbolic Logic* 11.2, pp. 185–193. MR: [2132114](#) (cit. on p. 4079).
- C. Hempel (2000). “An intellectual autobiography”. In: *Science, Explanation, and Rationality*. Springer, pp. 3–35 (cit. on p. 4081).
- D. Hilbert (2013). *David Hilbert’s Lectures on the Foundations of Arithmetic and Logic, 1917–1933*. Vol. 3. Edited by W. Ewald, W. Sieg and M. Hallett. Springer, pp. xxv+1062. MR: [3551948](#) (cit. on pp. 4076, 4092).
- D. Hilbert and W. Ackermann (1928). *Grundzüge der theoretischen Logik*. Die Grundlehren der mathematischen Wissenschaften, Band 27. Springer, pp. viii+120. MR: [0351742](#) (cit. on p. 4076).
- P. Mancosu (1999). “Between Vienna and Berlin: the immediate reception of Gödel’s incompleteness theorems”. *Hist. Philos. Logic* 20.1, pp. 33–45. MR: [1700712](#) (cit. on p. 4081).
- J. von Plato (2017a). *Saved from the Cellar; Gerhard Gentzen’s Shorthand Notes on Logic and Foundations of Mathematics*. Sources and Studies in the History of Mathematics and Physical Sciences. Springer, Cham, pp. x+315. MR: [3642843](#) (cit. on p. 4077).

- (2017b). *The Great Formal Machinery Works, Theories of Deduction and Computation at the Origins of the Digital Age*. Princeton University Press, Princeton, NJ, pp. viii+377. MR: [3676146](#) (cit. on p. [4081](#)).
- (2018). “Kurt Gödel’s first steps in logic: formal proofs in arithmetic and set theory through a system of natural deduction”. *Bull. Symb. Logic* 24 (cit. on p. [4076](#)).
- T. Skolem (1929). “Über einige Grundlagenfragen der Mathematik”. *Norsk videnskaps-akademi i Oslo. Skrifter I. Mat.-naturv. klasse* 4.
- A. M. Turing (1936). “On computable numbers, with an application to the Entscheidungsproblem”. *Proc. London Math. Soc. (2)* 42.3, pp. 230–265. MR: [1577030](#) (cit. on p. [4077](#)).
- H. Wang (1996). *A Logical Journey, From Gödel to Philosophy*. MIT Press, Cambridge, MA, pp. xiv+391. MR: [1433803](#) (cit. on pp. [4080](#), [4090](#), [4091](#)).
- A. Whitehead and B. Russell (1927). *Principia Mathematica*. 2nd ed. Vol. I–III. Cambridge (cit. on p. [4076](#)).

Received 2017-12-07.

JAN VON PLATO  
UNIVERSITY OF HELSINKI  
[jan.vonplato@helsinki.fi](mailto:jan.vonplato@helsinki.fi)



## IMPA'S COMING OF AGE IN A CONTEXT OF INTERNATIONAL RECONFIGURATION OF MATHEMATICS

TATIANA ROQUE

### Abstract

In the middle of the 20<sup>th</sup> century, the intimate link between science, industry and the state was stimulated, in its technical-scientific dimension, by the Cold War. Questions of a similar strategic nature were involved in the Brazilian political scene, when the CNPq was created. This presentation investigates the nature of the connection between this scientific policy and the presumed need for an advanced research institute in mathematics, that gave birth to IMPA. By retracing the scientific choices of the few mathematicians working at the institute in its first twenty years, we demonstrate how they paralleled the ongoing reconfiguration of scientific research. The development of dynamical systems theory provides a telling example of internationalization strategies which situated IMPA within a research network full of resources, that furnished, moreover, a modernizing drive adapted to the air of that time.

On October 19, 1952, *O Jornal do Comércio*, a newspaper published in Brazil's then capital city of Rio de Janeiro, reported that the National Research Council (CNP, Conselho Nacional de Pesquisas) had created an associated research arm called the Institute for Pure and Applied Mathematics (IMPA, Instituto de Matemática Pura e Aplicada). The National Research Council, later known as CNPq, had been created just a year beforehand and was directly connected to the country's government. On the same day, the first pages of the newspaper reported:

- The Soviet delegate presented a proposal for peace at the General Assembly of the United Nations. Stalin's Foreign Minister, Andrei Vyshinsky called for a reduction of one-third of the armaments of the great powers and an unconditional ban on atomic weapons. The "peace pact" would be a condition to stop the then ongoing Korean War;

- As the *de facto* leader of the Republican Party, Eisenhower's campaign to become US president had become fierce. His disagreements with Truman included an alleged covert plan by the General to end the Korean War. Foreign policy was presented as a weakness of the Democrats. Eisenhower's focus was to defeat the Communists, maintain pressure on the USSR and expand the American atomic arsenal.

The reaction of the Soviet envoy had been motivated by the signing of the North Atlantic Treaty in Washington in April 1949, designed to contain an armed attack by the Soviet Union against Western Europe. The key section of the treaty was Article V, which commits each member state to consider an attack against another member state as an attack against all members. In 1951, the treaty gave rise to NATO, with General Eisenhower at the head.

In the same year of 1951, the coordination to establish the CNPq was finalized, in large part due to the stubborn efforts of Admiral Álvaro Alberto da Motta e Silva, who acted as spokesman for the interests of a small but significant group of scientists. They wanted to overcome the country's chronic backwardness and boost economic development, and attributed a strategic role to nuclear power for both industry and national security. In the early years, the National Research Council's investments concentrated on infrastructure for the nuclear sector, which was closely connected with the field of atomic physics. The Brazilian Center for Research in Physics (CBPF, Centro Brasileiro de Pesquisas Físicas) had been created in 1949, contributing significantly to the foundation of CNPq and IMPA. Due to their international recognition, particle physicists such as César Lattes and José Leite Lopes played a central role in the public discussion on science policies [Vieira and Videira \[2014\]](#).

This was the time when Big Science began reshaping the very meaning of science and came to symbolize modernity, occupying the center of a new social contract between scientists and the state. As Pestre and Krige propose, after the Second World War, the intimate link between science, industry and the state was stimulated, in its technical-scientific dimension, by the Cold War [Krige and Pestre \(eds.\) \[1997\]](#). Even if "the desire to produce knowledge, to know more about 'nature' still remained the main motive of the practitioners", scientists pragmatically exploited the possibilities that only the state had to provide material resources – producing a "new identification between science and technology and state power and prestige" [Krige and Pestre \(eds.\) \[ibid., p.xxxiii\]](#).

Negotiations of a similar strategic nature were involved in the Brazilian political scene. The CNPq was meant to lay the groundwork for the purchase of reactors, assessing international scientific cooperation agreements, as well as helping to fight against monazite and thorium oxide exports, which could be useful to Brazil in the development of its own nuclear program. The goals of Álvaro Alberto were clearly to construct a sovereign position to the country in a Cold War background [de Andrade \[1999\]](#).

The question arises, however, as to the nature of the connection between this scientific policy and the presumed need for an advanced research institute in mathematics. What was the relationship between the mathematics stimulated by IMPA and nuclear physics that was at the center of the political power of CNPq? The institute could have been expected to focus on the mathematical foundations of atomic physics, yet this was not the case. The relationship between the political atmosphere and the core subject matter of mathematical research in the first years of IMPA is a complex question that so far has not been explored.

In fact, this weakness fits into a broader historiographical problem. Valuing the social and political context is now common practice in the history of science. What is less common, however, is to convincingly show how this context of knowledge production really matters to the knowledge being produced. This is one of the questions raised in the book *Science and Technology in the Global Cold War* Oreskes and Krige (eds.) [2014]. The authors ask how Cold War patronage specifically affected the patterns and priorities of scientific research, and seek to determine what role national ambitions played in fostering, enabling, or disabling certain lines of investigation. These questions are even more difficult to answer in the case of mathematics (not covered in the mentioned book).

There was a change in the direction of research at the end of the 1960s at IMPA, with greater focus being put on dynamical systems theory. We will show, in the final sections, how this reorientation occurred and explore the possible relationship with the development of applied mathematics in the US, which was itself related to priorities adopted due to the Cold War. This study can thus be considered as a step towards understanding the roles that patronage and Cold War geopolitics played in shaping mathematicians choices, defining spheres of possibility for concrete research (related to a question raised in Oreskes and Krige (eds.) [ibid., p.7]). This means understanding why some lines of research in mathematics were pursued while others were left out, and to what extent these choices were driven by the possibilities of patronage and international connections. In order to emphasize the changing priorities that would take place in the end of the 1960s, the first two sections describe the beginnings of IMPA, the political forces leading the project and its main directions of research before this turning point. By retracing the scientific choices of the few mathematicians working at the institute in its first twenty years, we will demonstrate how they paralleled the ongoing reconfiguration of scientific research. The development of dynamical systems theory provides a telling example of internationalization strategies which situated IMPA within a research network full of resources, that furnished, moreover, a modernizing drive adapted to the air of that time.



## 1 Periodization and institutional dimensions

The period of interest for this study begins in 1949, when the CBPF was founded, and ends in 1971, when IMPA began having a formally established post-graduate program and a stable research team<sup>1</sup>. In this same year, one of its founders, Leopoldo Nachbin, left IMPA. Nachbin's departure has not been sufficiently explored from a historiographical point of view. Testimonies evoke personal reasons and disagreements with colleagues<sup>2</sup>. These explanations are not good enough for a historian. The controversy surrounding Nachbin's departure is of particular interest because it opens the possibility of mapping the distinct mathematical influences that contributed to the consolidation of specific fields of research at IMPA in the 1960s.

Around 1970, the mathematical community became more numerous and institutionalized in IMPA. From then on, the organization of research changed. IMPA started its activities in a room of the CBPF. There was a director, Lélío Gama, and researchers who gave courses more or less regularly: Mauricio Peixoto, Leopoldo Nachbin and Paulo Ribenboim. Even after moving to a new building in 1957, the institute consisted of a small number of professors and students. Elon Lages Lima joined as a researcher in 1956 and helps to give an idea of the dimensions IMPA had at this time:

At the end of every month, Mr. Antonio came, he was the one who looked after the building at the corner of Sorocaba with São Clemente. Mr. Antonio was the guardian of the building, he lived there with his wife, Dona Maria. At the end of the month, Mr. Antonio came with a paper bag containing several parcels of money, which were our salaries, and he said: sign here, professor! This money came from the CNPq. He received the money and gave it to us, that's all [Lima \[11 May 2016\]](#).

Since 1939, there was a department of mathematics in the Philosophy Faculty of the University of Brazil (now, Universidade Federal do Rio de Janeiro)<sup>3</sup>. But, in the minds of IMPA's early supporters, advanced research was associated with the possibility of creating spaces outside the university. When the CBPF was founded, Leite Lopes was convinced that Brazil had to become something other than a "science-starving country", by strongly associating this possibility with the creation of research centers outside the university: "Our hopes turned to the university where unfortunately, by virtue of the lack of

<sup>1</sup>Before this, the degrees depended on an agreement with the University of Brazil.

<sup>2</sup>Nachbin himself says: "I was one of the founders of IMPA, and I left IMPA for personal reasons, these fights can occur everywhere". In: Interview with Leopoldo Nachbin by Roberto Martins and Hiro Kumasaka. CLE/Unicamp Historical Archives.

<sup>3</sup>In 1964, this department has been merged with the mathematics department of the Faculty of Engineering and others in the same university, giving birth to the Institute of Mathematics of the UFRJ.

understanding and intolerance of our statesmen, science advances slowly and painfully” [Videira \[2004\]](#). IMPA was the result of the efforts of a small group of very well organized and politically articulated scientists who convinced the government that a key factor in economic development is the encouragement of autonomous research institutions.

While debating the creation of IMPA as an entity directly linked to the National Research Council, Baptista Pereira, a member of the Council, asked whether it was better to create, instead of a new institute, a new course at the university. But another member, Cândido Dias, responded by stressing that independence was justified by the fact that universities could only have a very small number of professors, which made it difficult to hire all the mathematicians then dedicated to research. Dias specifically invoked the situation in Rio de Janeiro, where some prominent researchers (as Leopoldo Nachbin and Mauricio Peixoto) were not full professors at the university. The creation of an independent institute, he added, would give them a stable form of support. This also followed the understanding that high-level research needed “protected spaces”, detached from the constraints imposed by universities:

When there was a competition for a full position in mathematical analysis at the National Faculty of Philosophy, in 1950, José Abdelhay and Leopoldo Nachbin were candidates. The difference in titles between Abdelhay (baccalaureate in mathematics) and Nachbin (engineer) served as the basis for challenging Nachbin’s registration, who filed an appeal and thus the competition was suspended pending the court’s decision. This has become one of the longest known academic disputes at any Brazilian university. On the initiative of physicist José Leite Lopes, who would become one of the most distinguished Brazilian scientists, Monteiro<sup>4</sup> had been hired at the Centro Brasileiro de Pesquisas Físicas (CBPF), which was founded in Rio de Janeiro. Leopoldo Nachbin was also hired at this center. Thus, the CBPF set up the first “protected space” for mathematical research supported by the federal government<sup>5</sup>.

The same argument applied to the need to create IMPA as an independent institute. As Nachbin says, the CBPF was established because at the University of Brazil “there were no conditions to create a post-graduate program in physics”. He and Cândido Dias talked often about the necessity to have also an independent institute of mathematics. In fact, Joaquim Costa Ribeiro, the scientific director at the time, preferred to create a program

<sup>4</sup> Antonio Aniceto Monteiro was a Portuguese mathematician that stayed in Brazil from 1945 to 1949.

<sup>5</sup>Proceeding 112: meeting held at 17/10/1952; Proceeding 117: meeting held at 15/10/1952. Rio de Janeiro. MAST. Archives CNPq.

inside the university, but was convinced by Dias (with the support of the president of the Academy of Science, Artur Moses)<sup>6</sup>.

Nachbin and Peixoto were very influential in the decision processes of the National Research Council. During the 1950s, the CNPq complemented the salaries of researchers from different universities (where there were very few stable positions), gave scholarships for students and promoted travels of Brazilian researchers inside the country or abroad, besides inviting foreign mathematicians. The decisions were centralized in the hands of the Orientation Committee (that linked IMPA to the CNPq, being above the Scientific Local Committee). A rejected proposal from the University of Brazil shows that it was probably not accommodating the role community of mathematicians.

In 1959, Carlos Alberto Aragão de Carvalho, a professor from the University of Brazil, presented to the CNPq a project to create a National Commission of Mathematics. It was conceived as a means to unify different programs of research in mathematics and to have an impact on the training of future engineers. The commission would establish an Inter-American Instituto de Matemática Pura e Aplicada, integrating other Latin American countries, giving grants and stimulating visits from foreign researchers<sup>7</sup>. Charged with making a report on the project for the president of CNPq, Lélío Gama saw no interest in the proposal. It would be too complex and its goals “constitute regional problems, with singular characteristics in each country, and so demanding national solutions”<sup>8</sup>. One significant point of Aragão’s project was the assimilation of all mathematical institutes into the university, a problem that, Gama said, “in our view must be examined in light of particular circumstances in each case presented, in the sense to verify if this assimilation would imply, really, greater facilities for mathematical research in the region considered”. In Brazil, the only mathematical institute was IMPA. And it is clear that the project aimed to incorporate it into the university, an idea promptly rejected by the CNPq.

The development of advanced institutions outside the university was not a Brazilian exception. The evolution of the research systems in India has been described in terms of a *dualism*, as suggested by Raina and Jain [1997]. This notion characterizes the institutions of science and technology as structured by the requirements of a rapidly evolving knowledge standards as much as by the imperatives of modernization. The role of science in constructing a sovereign and modern state had similar flavors in Brazil and India during the 1950s<sup>9</sup>, and scientific institutions played a central role in reconfiguring the nation-state. In both countries, forming a scientific elite was seen as a key strategy to

---

<sup>6</sup>As Nachbin tells himself in an interview given to Elisabete Burigo in June 1988, available at <http://www2.unifesp.br/centros/ghemat/paginas/teses.htm>

<sup>7</sup>In 1958, a mathematical center had been created in Buenos Aires, with the support of UNESCO, that Nachbin visited in 1959.

<sup>8</sup>Processo 3595/59 CNPq. Arquivos Lélío Gama, MAST (LG-T-05-065).

<sup>9</sup>Even if the two countries have different histories with regard to colonization.

the mission of building a new nation. As Raina and Jain affirm, “the emergence of Big Science required the emergence of new institutions and the concomitant supersession of the university considered as ‘the age-old site for the production of knowledge’.” [Raina and Jain \[ibid., p.859\]](#). The tasks of national development, linked to research in atomic energy, implied the creation of a solid infrastructure of research, conceived as elite institutions where: “young men of the highest intellectual calibre in a society” could be trained [Raina and Jain \[ibid., p.866\]](#).

Travels were a decisive mean to train researchers. Foundations like Guggenheim and Rockefeller, as well as the Department of State, played an important role in promoting scientific exchanges in the period. Between the two wars, a Europe weakened by reconstruction efforts and the rise of Nazism helped to explain the importance gained by American philanthropic foundations, as Reinhard Siegmund-Schultze shows, with special attention to the case of mathematics [Siegmund-Schultze \[2001\]](#), and John Krige develops for the period after the Second War [Krige \[2006\]](#). A considerable number of Brazilian mathematics-related researchers traveled to US with a Rockefeller fellowship grant – their travels have already been elucidated in [Trivizoli \[2011\]](#) and [Barany \[2016\]](#). Leopoldo Nachbin and Mauricio Peixoto both received grants from US foundations.

In the beginning, the University of Chicago was a preferred destination of Brazilian mathematicians. This is also due to the presence of André Weil and his close relationship with Nachbin, since the years Weil stayed at São Paulo. Marshall Stone had become chair of the Department of Mathematics at Chicago in 1946, spearheading its renewal. Policies during Stone’s tenure were aligned with post-war American politics, which included training a cadre of high-level students from different countries [Parshall \[2009\]](#). Nachbin has stayed in Chicago for a first time from 1948 to 1950 and Peixoto also was there from 1949 to 1951. After going to Chicago for a second time in 1957, Nachbin went to Princeton, as well as Peixoto.

During the second half of the 1960s, other Brazilian mathematicians went on to study in the United States. From 1968 onward, some of them returned and settled at IMPA, like Jacob Palis, in 1968<sup>10</sup>, and Manfredo Perdigão do Carmo, in 1969<sup>11</sup>. Jacob Palis underlines the changes going on, in the late 1960s, in the conduction of mathematical research: “In fact, in 1969, a group of researchers arrived from abroad with the intention of shaking up IMPA” [Palis, Camacho, and Lima \(eds.\) \[2003, p.125\]](#).

In a move potentially related to the transformation of mathematics research at IMPA, Nachbin left the institute soon afterwards. This question requires further historical analysis.

---

<sup>10</sup>Palis was hired as assistant researcher and promoted immediately to associate researcher and then tenured professor in 1970.

<sup>11</sup>Carmo was officially a researcher at IMPA since 1966, but he started effectively in 1969, after a stay in the US and at the University of Brasília.

## 2 Leopoldo Nachbin and the Bourbakist approach

Leopoldo Nachbin received an engineering degree from the University of Brazil in 1943, the same year as Mauricio Peixoto. He started working with Antonio Monteiro. Although more intensely devoted to logic, Monteiro also had a strong working knowledge of order structures. Connecting it to topology would be fundamental to Nachbin, who defended his *livre-docência* thesis on metrizable and pseudo-metrizable topologies in 1948. Just before, a version of this work had been sent by Dieudonné to the *Comptes Rendus de l'Académie des Sciences de Paris* [L. Nachbin \[1946\]](#).

In 1945, André Weil came to the University of São Paulo (where he stayed until 1947) and, after him, several mathematicians from the Bourbaki group had residences at Brazilian universities: Jean Dieudonné, Charles Ehresman, Alexander Grothendieck and Lawrence Schwartz [Pires \[2006\]](#). Nachbin was 26 years old in 1948, when he published a monograph on topological vector spaces, which became very useful for training researchers [L. Nachbin \[1948\]](#). Mário Carvalho de Matos credits Nachbin with having promoted the “Bourbaki spirit” in Brazil, specifically mentioning the theory of topological vector spaces, a characteristic of Nachbin’s “mathematical style” [Barroso and A. Nachbin \[1997\]](#).

Cândido Dias’ 1951 professorial thesis provides an example of how topological vector spaces were having influence in the practice of mathematics [Dias \[1951\]](#). Before 1945, the mathematics department at the University of São Paulo had been frequented by Italian mathematicians, the most famous being Luigi Fantappiè. However, according to Cândido Dias, the topological base of Fantappiè’s theory of linear functionals was “precarious”. It lacked an element that later proved to be indispensable: the generalization of normed spaces. Topological vector spaces were thus of special interest in the study of analytic functionals. The vector space that would serve as a basis and the class of functionals were perfectly clear elements, but it was still necessary to combine the two. That is to say, it was necessary to “put a topology in the vector space whose continuous functionals were the class of Fantappiè’s analytics”. This demonstrates how the theory of analytic functionals *gets along* with modern functional analysis.

“Writing in Bourbaki language”, as Cândido Dias puts it, was a trend within the small mathematical research community of the early 1950s. The question of adapting notations, definitions and demonstrations to such a language was a key one. Nachbin went frequently to São Paulo and was the main researcher at Rio de Janeiro working on related topics. In the early 1950s, a problem posed by Dieudonné (and signed by Bourbaki) drew Nachbin’s attention, it was the question of knowing if any bornological space is barreled [Bourbaki \[1950\]](#). A negative answer was published by the Brazilian in 1954 [L. Nachbin \[1954\]](#)<sup>12</sup>. During the years 1953 and 1954, Grothendieck gave a course on topological vector spaces

<sup>12</sup>Cited in Bourbaki’s book on topological vector spaces, edition of 1955, v.2, p.13.

at IMPA and published a book on the subject [Grothendieck \[1954\]](#). In the introduction, he announced a forthcoming work that would bring together his reflections and those of Nachbin – this never materialized, perhaps because the research interests of Grothendieck changed.

In parallel, Nachbin was doing research on a generalization of Hahn-Banach's theorem, about which he had published one of his most important articles in 1950 [L. Nachbin \[1950\]](#). Nachbin was also working on the theory of approximation and extended the Stone-Weierstrass theorem to differentiable functions, a result published in the *Annals of Mathematics* [L. Nachbin \[1959\]](#). This research, which developed from interactions with Marshal Stone, continued throughout the 1960s and involved some doctoral students<sup>13</sup>. However, most of Nachbin's PhD students at IMPA worked on a subject he inaugurated around 1963 and would engage him until the end of this life: the topology of spaces of holomorphic mappings.

After his second stay in Chicago, in 1957, Nachbin wanted to go to Paris to follow Schwartz's seminar on partial differential equations. Schwartz supported the idea<sup>14</sup>. Nachbin did not end up going to Paris, however, perhaps because he discovered, some months afterwards, that the plans for the seminar had changed, since Schwartz wanted to study the applications of his theory of distributions to theoretical physics.

After Schwartz visited Brazil, in 1961, Nachbin finally went to Paris. In the same letter in which he had confirmed the invitation and stipulated Nachbin's salary as associate professor, Schwartz presented the subject of the conferences he was planning to make in Brazil. Schwartz seems more interested in talking about the irreducible representation of Lorenz groups in spaces of distributions with vector values than about topics related to topological vector spaces or partial differential equations (as we see in [Figure 1](#)).

In the 1960s, the point of view of topological vector spaces was not unanimously recognized as being so interesting as it was before. In 1957, Schwartz proposed extending to distributions with vector values the main properties of ordinary distributions (scalar distributions) [Schwartz \[1957\]](#). In particular, he showed that the topological properties of spaces of distributions could be studied using similar tools to those already employed in the ordinary case. Other properties, however, were more difficult to extend. Nonetheless, it was important to study distributions with vector values, as Schwartz supports, because theoretical physics uses distributions with values in operator spaces. Schwartz's articles were deeply influenced by Grothendieck's works on kernel theorems and topological tensor products. However, as Anne Sandrine Paumier shows, the reception of this approach was controversial, since some mathematicians saw the introduction of topological vector

<sup>13</sup> As Silvio Machado, João Bosco Prolla and Guido Zapata.

<sup>14</sup> He says in a letter to Harry M. Miller of the 5th May 1956. I thank Lucieli Trivizoli for showing me some letters, found in the Rockefeller Foundation, suggesting the following version.

Les conférences de Berkeley contiennent notablement plus de matériel que celles de Buénos-Aires, et j'ai encore quelque chose à y ajouter. Je crois que la partie purement mathématique du problème, c'est-à-dire la recherche des représentations irréductibles du groupe de Lorentz dans des espaces de distribution à valeurs vectorielles, serait très adaptée pour le Brésil puisqu'il s'agit d'espaces vectoriels topologiques, des produits tensoriels, des distributions, convolutions, intégrales de Fourier et équations aux dérivées partielles, tous sujets qui ont été bien travaillés par les mathématiciens de Rio. Naturellement cela aura l'inconvénient de répéter un peu des choses déjà faites ailleurs, mais je crois que cela n'est pas grave.

Je crois que pour le moment, je n'ai rien à dire de nouveau sur les espaces vectoriels topologiques et équations aux dérivées partielles, qui d'ailleurs ont fait l'objet d'assez nombreuses publications pour être à la disposition de tout le monde.

Figure 1: Extracts from a letter Schwartz sent to Nachbin in November 23 1960

spaces in the theory of distributions as unnecessarily complicated. Paumier observes that: “the transformation of the kernel theorem into a nuclear property of certain topological vector spaces leads to a much more important imbrication of distributions with topological vector spaces; and this even ensues a transformation of the theory of distributions itself, as well as the creation of the theory of distributions with vector values (...) The objects considered are no more distributions that are represented by kernels but spaces of distributions to which we give a structural property of nuclearity. Developing the theory of distributions with vector values, Schwartz incorporates, in some way, the writing practices around topological vector spaces” [Paumier \[2014, p.170\]](#).

This point of view, namely the writing practices around topological vector spaces, may have influenced Nachbin. Beyond his usual domains of research, around 1963, he was investing on the study of topologies of spaces of holomorphic mappings. In this year, Nachbin gave a course on the theory of distributions at the University of Rochester (published in 1964 by the University of Recife as [L. Nachbin \[1964\]](#)). There, he treated distributions in a new way:

“In planning my course, I had to face the following dilemma. Should I teach distributions on  $R^n$  (by using the coordinatewise approach), or should I do it on a finite dimensional real vector space (by preferring the intrinsic viewpoint)? Many, many years ago,



algebraists used to find it more pedagogical to talk first of permutation groups, and next of the then called ‘abstract’ groups (...). This is no longer the usual attitude in Algebra courses; groups are introduced from the very start, and permutation groups are mentioned as a fundamental example (...). Surprisingly enough, analysts still find it more pedagogical to present firstly analysis in  $R^n$ , and next maybe talk about analysis on a finite dimensional real vector space (...). By following a recent trend, we believe that analysis on a finite dimensional vector space should get an increasing emphasis from the very beginning of graduate courses, and should prevail over analysis on  $R^n$ ” [p.3-4].

Besides this “pedagogical” reason, the adoption of an intrinsic approach in the case of finite dimensional vector spaces stems from the need for such an approach in dealing with infinite dimensional vector spaces. Nachbin wanted in fact to extend Schwartz’s results to infinite-dimensional spaces. He started to work hard and to direct theses mainly on this subject<sup>15</sup>. John Horváth says that, in 1965, Nachbin started to build a “very important theory, totally original, the Theory of Holomorphic Functions in Banach Spaces, with applications to convolution equations and to partial differential equations in these spaces”. He adds this new theory attracted some of his own students<sup>16</sup>. As we learn from Nachbin’s writings and from interviews with some of his colleagues and former students, Nachbin believed he was founding a new field of research. Foreign students, some of whom met Nachbin in Rochester, came to IMPA attracted by this domain. This was the case of Sean Dineen, Richard Aron, and Philip Bolan. The working conditions for the subject at IMPA were favorable. As Dineen attests: “they were interested in making IMPA a world class institute, so they paid quite good salaries to all staff, including the PhD students” [McGuire \[2009\]](#). In Brazil, they met other Nachbin students, such as Jorge Alberto Barroso, Mário Carvalho Matos, Soo Bong Chae and Jorge Mujica. Besides the three theses directed by Peixoto, there were six theses at IMPA in the period, that were all directed by Nachbin<sup>17</sup>. These theses were either about the theory of approximation or about the topology on spaces of holomorphic mappings. And there were also the thesis on this last subject that Nachbin directed in Rochester. These researches gave rise to a book published by Springer [L. Nachbin \[1969\]](#)<sup>18</sup>.

The number of people who continued working in this field after the end of the 1970s diminished markedly. A plausible hypothesis is that the study of topologies of spaces of holomorphic mappings, with a possible extension to infinite dimension, did not have the

<sup>15</sup>During his years in Paris, besides Schwartz’s seminar, Nachbin frequented Pierre Lelong’s seminar. He knew the works of Andre Martineau and attracted students to this field, as Philippe Noverraz, Gérard Coeuré, André Hirschowitz and Jean-Pierre Ramis. Nachbin’s contacts in Paris must have motivated him to move to this new area

<sup>16</sup>Testimony for the title of *honoris causa* to Nachbin given by UFPE.

<sup>17</sup>Just the one of Luiz Adauto Medeiros was signed by Nachbin but effectively directed by Felix Browder.

<sup>18</sup>Already presented in the *Sexto Colóquio Brasileiro de Matemática*, in 1967.



success and posterity expected by Nachbin. In 1972, questions were more numerous than answers and some actors admitted “the poverty of the theory in regard of the hopes we can put in it” [Hirschowitz \[1972, p.256\]](#). Dinamérico Pombo, one of Nachbin’s last students, chose another domain when he started, motivated by the recognition that holomorphy theory was not a very stimulating subject (a perception shared by Luiz Adauto Medeiros)<sup>19</sup>. When asked if he continued working on infinite-dimensional complex analysis, Sean Dineen answered:

In the 1970s it was very topological, locally convex spaces, pseudo-convexity, holomorphic convexity, analytic continuation and things like that. At the end of the 1970s Phil Boland moved into statistics, Richard Aron went permanently to Kent State, so that stream had sort of finished. But if you want to stay active as a research mathematician, you have to reinvent yourself regularly.

The progressive decline of what Paumier calls “writing practices of topological vector spaces” during the 1960s may have been one of the reasons for the departure of Nachbin and his group from IMPA in 1971<sup>20</sup>. In fact, the point of view proposed by Schwartz in 1957 had already met some resistances during the 1960s. This explains the success of the works of Lars Hörmander, since this author preferred considering the theory of distributions without any practices linked to topological vector spaces. Paumier observes that “the objects ‘distributions’ are very important, but the structure, mainly topological, of spaces under consideration are not essential in his work” [Paumier \[2014, p.170\]](#).

It is often said, mainly rooted in declarations of Elon Lima, that one reason for Nachbin’s depart was his attempt to hire one of his students at IMPA [Palis, Camacho, and Lima \(eds.\) \[2003\]](#). This explanation is not convincing *per se*. Nachbin sent a letter to Lima, in 18 September 1969, supporting the proposal to hire Jorge Alberto Barroso after the end of his thesis<sup>21</sup>. He strongly emphasized the key role Barroso would play in elucidating questions related to Nachbin’s own interests in that time, namely the extension to locally convex complex spaces of the theory for Banach spaces developed in [L. Nachbin \[1969\]](#). Beyond personal motivations, which certainly existed and had probably influenced the outcome, there were important changes going on in the global scene of mathematical research. In particular, the US was acquiring an increasingly prominent role in mathematics and, importantly, the preferred domains began shifting away from topics related to the

<sup>19</sup>Interviews done at Rio de Janeiro in 2017 in the writing of this article.

<sup>20</sup>We asked Mário Carvalho if he agrees that the field of holomorphy had a decline but he contests, naming researchers that continued working on related themes. [Toledo \[2012\]](#) can be consulted for a list of researches in the field afterwards.

<sup>21</sup>Letter found in the archives of Lélío Gama at MAST (LG-D10-138-0011).

Bourbaki lineage, which had been so influential until this time. In order to fully understand the context, it is necessary to look at the wider picture, as we do in the following sections.

### 3 The rise of dynamical systems theory in the United States and the role of Mauricio Peixoto

In the aftermath of the Second World War, a redistribution of scientific forces took place on a global scale with major repercussions in mathematical research. Amy Dahan-Dalmedico shows that applied mathematics gained much more importance [Dahan-Dalmedico \[1996\]](#) and the United States became the leading mathematical power by the sheer breadth of its scientific community, the variety of fields covered, and by the dynamism of its research systems. Solomon Lefschetz was an exemplary figure in this scenario [Dahan-Dalmedico \[1994\]](#), since he moved from topology to differential equations during the war. Mauricio Peixoto worked with him and played a key role in shaping dynamical systems theory to be adapted to a transition towards US dominance in mathematics.

Just after the war, Lefschetz started leading a research program on Nonlinear Differential Equations and Nonlinear Oscillations funded by the Office of Naval Research. That allowed him to translate important works of the Soviet school of research in the theory of oscillations. After his stay in Chicago, where he went to work in analysis, Peixoto went back to the US in 1957, to work with Lefschetz. Just after the launch of the Sputnik, it became clear that it was necessary to fill the “mathematical gap” between Russia and the West, as Lefschetz said, so he created a mathematical center in the Research Institute for Advanced Studies (RIAS), in Baltimore, which has gained worldwide recognition. The focus was on the theory of nonlinear oscillations.

Lefschetz’s laboratory is known for having introduced in the US concepts formulated in the Soviet Union by Andronov and his group. Most notably among them was the concept of structural stability, which later became central to Peixoto and Brazilian researchers. The notion of a “systèmes grossiers” was proposed by Andronov and Pontryagin in an article published in French in 1937 [Andronov and Pontryagin \[1937\]](#), and developed in a book that Andronov wrote (in Russian), with other researchers, about the requirements mathematical models should fulfill to be useful to physics (translated as [Andronov, Vitt, and Khaikin \[1949\]](#)). The mathematical definition of this idea became what is now known as *structural stability*. Mauricio Peixoto, as well as other researchers who worked with Lefschetz in the beginning, specially De Baggis and Marilia Peixoto, played a key role in these developments, giving mathematical consistence to the initial concepts [Roque \[2007\]](#).

In 1959, Mauricio Peixoto led a round table on structural stability in a symposium on differential equations in Mexico. Lefschetz collaborated with a PhD Program at UNAM (Universidad Nacional Autónoma de México) and organized this International Symposium on Ordinary Differential Equations. Thanks to his meeting with Peixoto, Stephen Smale came to Brazil in 1960 and began conducting a research at Berkeley. This research was an extension of the work already done by Peixoto in two dimensions to higher dimensions, and included other mathematicians at Berkeley such as Morris Hirsch and Abraham. Abraham claimed that “the new subject was well under way in the fall of 1960, when I arrived in Berkeley, and the golden age of global analysis began” [Abraham \[2009\]](#). It is interesting to note that “dynamical systems” was not the official name of the field in these times. In the proceedings of the 1962 Brazilian Colloquium of Mathematics, Mauricio Peixoto presented the question of structural stability, which he maintained was a fundamental problem in “the theory of differential equations” and he felt it necessary to add that such a theory “is also called theory of vector fields of dynamical systems” [Peixoto \[1961\]](#).

The first three theses done at IMPA were supervised by Mauricio Peixoto, all finished in 1964 on subjects related to structural stability or generic vector fields (Ivan Kuptka, Jorge Sotomayor and Aristides C. Barreto). In 1964, the center Lefschetz directed at the RIAS moved to Brown University, renamed as Center for Dynamical Systems. Peixoto went to work there and the research at IMPA was most conducted by Nachbin during the 1960s. It was only around 1970 after other researchers, who had done their PhDs in the United States, returned to Brazil, as Jacob Palis, that the field of dynamical systems began to establish itself. Since then, research in dynamical systems became increasingly valued and played a key role in the institutionalization of research at IMPA. Some intrinsic characteristics of dynamical systems theory help to explain why it was a better fit for the direction that mathematics was taking at the time across the world.

## 4 The Americanization of mathematics

Americanization is a controversial term. It has ambiguous meanings, representing either a deliberate action of conquest or the disinterested attitude of a country which sees itself as a benefactor vis-à-vis the rest of the world. These meanings were obviously constructed over time and have a long history. Here, Americanization refers to the meaning suggested by the historian Ludovic Tournès, who has studied the actions of philanthropic foundations in French science during the interwar period [Tournès \[2010\]](#).

Michael Barany speaks of “mathematical colonialism” to describe the scientific actions of the United States at the time [Barany \[2016\]](#):

On a geopolitical scale, postwar mathematical colonialism was an elite driven, internationally oriented endeavor that blended the lofty discourse of technical and moral development with the tangled bureaucratic negotiations that enabled substantive coordination among a diverse mix of governments, foundations, and other organizations.

Barany observes that special attention must be paid to the personal scale. In the above quote, the qualification of “elite driven” accurately highlights the key role of certain individuals during the period treated here. But the coordination between foundations, local governments and institutions that were being created in South America also played a key role in the reorientation of mathematical research. Upon closer investigation of the mathematics that was being done – that is, moving from the scale of major science policies to that of the mathematicians at work – it seems that it would be a mistake to characterize the action of US policies on Brazilian mathematics as colonialism. The notion of Americanization as a transnational action, proposed by Tournès, would be more appropriate. Analyzing the literature about the history of science in the period, we found this notion resonates better with what was going on in Brazil.

Tournès’ work discusses the biological sciences and the introduction of experimental methods in the social sciences. He shows that the actions of philanthropic foundations are not limited to a disinterested financing of research, but they do not follow acculturation strategies either. The consequences of financial priorities are to be searched on the subjects chosen and on the methodologies borrowed from interactions with foreign researchers. These foundations do not necessarily have any a priori goal aimed at acculturating other countries to an American ideology. Rather, financing constitutes a transnational action aimed at intervening in existing research environments that, based on common interests, may converge with the foundation’s investment priorities. It can be described as a kind of *seduction*, that is the term Antonio Pedro Tota uses to describe Americanization strategies in Brazil Tota [2009].

Within the history of mathematics, internationalization during the period under investigation is strongly linked to the action of US institutions that turned attention to Latin America after the Second World War. One of their main programs was to offer grants for young Latin American intellectuals to do internships at universities in the United States. The goal was to constitute an “invisible college”, with George D. Birkhoff as a prominent proponent for the development of mathematics in Latin America. In addition to increasing interactions among mathematicians, the program also invested in libraries and opened American newspapers to mathematicians from South American countries Ortiz [2003].

Other cultural arenas in Brazil experienced a similar process. Controversy over American influence on Bossa Nova was intense in the same period. The musical style that became a symbol of Brazil abroad emerged in 1958, after a period of complaints regarding

the golden age of the good neighbor policy implemented by the United States to influence Brazilian culture. Brazilian films and songs, symbolized by Carmen Miranda, had great impact in the US. But, since nationalism was a strong current in Brazilian thinking, Miranda was criticized for being “Americanized”, leading her even to write a song to contest such criticism. Unlike movies and songs from this first period, marked by a more offensive action, Bossa Nova could not be said to be a product of American acculturation. It was a new kind of synthesis, that intrinsically mixed elements of Brazilian music with characteristics of American jazz, associated with a modernization trend. This merging of different cultural influences, with jazz as one important component, could have contributed to the broad and increasing international recognition of Bossa Nova during the 1960s [Medaglia \[2013\]](#).

The relationship of Brazilian scientific research to economic and Cold War histories has been studied by few historians of science, and mainly in health sciences [Cueto \(ed.\) \[1994\]](#). Freire and Silva seek to remedy this situation by focusing on the role played by physicists in connecting their disciplinary communities and international scientific networks [Freire and Silva \[forthcoming\]](#).

In mathematics we can observe, at the same time, a decline of the highly abstract approach associated with Bourbakist (so, French) mathematics. This came along with an increasing vaporization of a more geometric point of view, symbolized by the works of René Thom and others <sup>22</sup>. There was a subtle malaise with the excess of formalism that the Bourbakist’s orientation reinforced. Diminishing the importance of Nachbin’s works as compared to Peixoto’s, Elon Lima associates Nachbin to an excessively formal “French style”:

At IMPA, researchers have always had a vision that it was not necessary to learn a ton of mathematics to do high-level research, meaningful research. Many of the formalisms, general, abstract and complex theories can be ignored, and one can focus on important, basic problems, and be successful in the same way – the greatest example of this is Professor Mauricio Peixoto. Professor Nachbin had a slightly different vision, because he had a more French-style training, that is to say, he had to learn a ton of things, but he still managed to do some good quality research. He had a vision of mathematics as a formal system, while Mauricio had a vision closer to that of an engineer [Palis, Camacho, and Lima \(eds.\) \[2003, p.119\]](#).

Taking into account the place Lima occupied at IMPA in subsequent years, we can infer that this opinion influenced the choices that the Institute would later take. Lima had gone

---

<sup>22</sup>See for instance the discourse of Hassler Whitney about one of the Fields Medalist, John Milnor, in the ICM of 1962 (the other Fields Medalist was Hörmander).

to the University of Chicago in 1954, under Nachbin's advice<sup>23</sup>. While talking to us about the years spent there, he remembered a little song: Analysts, topologists, geometers agree / if you go for generality / there's no one but Bourbaki / one theorem by them / is almost ten by you and me / Bourbaki goes marching on [Lima \[11 May 2016\]](#).

Lima quoted Morris Hirsch as one author of the song, which describes a somewhat ironic atmosphere involving what was perceived as an excess of generality of Bourbakist's concerns. Hirsch himself, to whom I wrote to ask about the song (that he remembers but says not to be of his own), describes the feelings related to Bourbaki in Chicago:

I didn't feel any unease about the Bourbaki approach, but as the song suggests, we thought it extremely, and perhaps unnecessarily, abstract. But we appreciated its logical and systematic treatment. I remember having difficulty finding standard results in Bourbaki – maybe in real analysis, or group theory – because their expositions started with the most general case, in which I had no interest or understanding, and only after many pages getting down to what I considered to be the real subject.

Bourbaki's mathematics was associated with being “too general”, and therefore too restrictive for “finding standard results”. At the same time, applied mathematics was acquiring greater importance. Peixoto, with his approach to dynamical systems, synthesized two strands, the declining and the ascending one, as will be explained below.

When discussing his motivation for proposing structural stability and genericity as key notions for the development of a theory of dynamical systems, Peixoto says that he was convinced that the main goal of the mathematics of his time was to classify mathematical objects, by means of equivalence relations between them, putting emphasis in their structures [Peixoto \[2000\]](#). He thought that it would be fruitful to express the theory of differential equations in a set-theoretic language. The suggestion already given by Poincaré (to classify the functions defined by differential equations) had to be fulfilled with notions from set theory. In order to do that, Peixoto sought to introduce two new elements [Peixoto \[1987\]](#):

1. A space of differential equations, or dynamical systems, possessing a topological structure;
2. A notion of qualitative equivalence between two differential equations.

Both requirements were fulfilled in [Peixoto \[1959\]](#) and [Peixoto \[1962\]](#). Peixoto defined the space of dynamical systems by considering a dynamical system as a point of a Banach space, and proposed that an equivalence relation between two systems in this space should

---

<sup>23</sup>Lima was supervised by Nachbin at the beginning of his research in analysis, but ended up getting his doctorate in topology, in 1958, with a thesis directed by Edwin Spanier in Chicago.

be a homeomorphism transforming trajectories of one system into trajectories of the other. This last definition was inspired by the work of Andronov and Pontryagin. This confirms that Americanization cannot be defined as the action of a nation towards another. Indeed, one of Peixoto's major innovations was the adaptation of a proposal first introduced by – great irony – soviet mathematicians.

After Peixoto's results, an analogous program seeking to generally describe dynamical systems, in higher dimensions, was proposed, with a special role played by Smale in the beginning of the 1960s. During this decade, some counterexamples, came directly from the modeling of physical phenomena (like the works of Lorenz on meteorology), challenged the theory and made it advance in forging new definitions [Aubin and Dahan-Dalmedico \[2002\]](#). The tension between physical examples and mathematical categories that could or could not express some kind of generality has been a driving force in the development of dynamical systems theory (as I show in [Roque \[2016\]](#)). A research program was then conceived that synthesized, on the one hand, a less formal, more applied, and more “oriented to specific problems” mathematics and, on the other, abstract and general concerns.

This program facilitated the relations with institutions in the United States and also served the project of Brazilian mathematicians to develop an autonomous and internationalized research center. The geopolitical situation restructured the relationship between governments, patronage agencies and scientists, the whole process being governed by informal evaluations and political compromises with the aim of building a new world-class institution. The focus on dynamical systems took on a strategic role, as it enabled a combination of various elements:

- Connection with research centers in the United States, guaranteeing the means for Brazilian mathematicians to access and take part in ongoing changes at the core of mathematics;
- Flexibility of an autonomous institution to build a modern image, particularly associated, at this time, with the research done in the US;
- Construction of a new field of research that did not require prior knowledge of a great number of mathematical results and that, being less formal, was adapted to the profiles of researchers;
- Association with the applied trend that was becoming dominant in the US. The field was seen as being useful to applied domains, even if it did not always focus on effective applications.

Americanization was a strategy of transnational appropriation of both the research questions and the means of making their development possible. It was a trend observed in the

core of mathematics, as well as in more subtle extra-mathematical motivations (associated with scientific policies and patronage), stimulating mathematicians to follow certain directions of research instead of others.

## References

- Ralph Abraham (Jan. 2009). “Recent Progress in Dynamical Systems Theory”. *Journal of the Calcutta Mathematical Society* 5 (cit. on p. 4106).
- Ana Maria Ribeiro de Andrade (1999). *Físicos, Mésons e Política: A Dinâmica da Ciência na Sociedade*. São Paulo: Hucitec-MAST (cit. on p. 4094).
- Aleksandr Andronov and Lev Pontryagin (1937). “Systèmes Grossiers”. *Doklady Akademii Nauk SSSR* 14.5, pp. 247–250 (cit. on p. 4105).
- Aleksandr Andronov, A. Vitt, and S. Khaikin (1949). *Theory of Oscillations*. Princeton: Princeton University Press (cit. on p. 4105).
- David Aubin and Amy Dahan-Dalmedico (2002). “Writing the History of Dynamical Systems and Chaos: *Longue Durée* and Revolution, Disciplines and Cultures”. *Historia Mathematica* 29, pp. 273–339 (cit. on p. 4110).
- Michael Barany (2016). “Fellow Travelers and Traveling Fellows: The intercontinental shaping of modern mathematics in mid-twentieth century Latin America”. *Historical Studies in the Natural Sciences* 46.5, pp. 669–709 (cit. on pp. 4099, 4106).
- Jorge Alberto Barroso and André Nachbin (1997). *Lembrando Leopoldo Nachbin*. Rio de Janeiro: UFRJ (cit. on p. 4100).
- Nicolas Bourbaki (1950). “Sur certains espaces vectoriels topologiques”. *Annales de l’Institut Fourier* 2 (cit. on p. 4100).
- Marcos Cueto (ed.) (1994). *Missionaries of Science – The Rockefeller Foundation and Latin America*. Bloomington: Indiana University Press (cit. on p. 4108).
- Amy Dahan-Dalmedico (1994). “La renaissance des systèmes dynamiques aux États-Unis après la Deuxième Guerre Mondiale: l’action de Solomon Lefschetz”. *Rendiconti del circolo matematico di Palermo (II)* 34, pp. 133–166 (cit. on p. 4105).
- (1996). “L’Eessor des mathématiques appliquées aux États-Unis: l’impact de la Seconde Guerre Mondiale”. *Revue d’histoire des mathématiques* 2, pp. 149–213 (cit. on p. 4105).
- Cândido Lima da Silva Dias (1951). “Espaços vectoriais topológicos e sua aplicação na teoria dos espaços funcionais analíticos”. PhD thesis. São Paulo: USP (cit. on p. 4100).
- Olival Freire and Indianara Silva (forthcoming). “Scientific exchanges between the US and Brazil in the 20th century: Cultural diplomacy and transnational movements”. In: John Krige. *Writing the Transnational History of Science and Technology*. Cambridge, MA: MIT Press (cit. on p. 4108).



- Alexander Grothendieck (1954). *Espaces vectoriels topologiques*. São Paulo: Universidade de São Paulo (cit. on p. 4101).
- André Hirschowitz (1972). “Prolongement analytique en dimension infinie”. *Annales de l’Institut Fourier* 22.2, pp. 255–292 (cit. on p. 4104).
- John Krigé (2006). *American Hegemony and the Postwar Reconstruction of Science in Europe*. Cambridge, MA: MIT Press (cit. on p. 4099).
- John Krigé and Dominique Pestre (eds.) (1997). *Companion to Science in the Twentieth Century*. London and New York: Routledge (cit. on p. 4094).
- Elon Lima (11 May 2016). “Interview done at IMPA by Rogério Siqueira and Tatiana Roque” (cit. on pp. 4096, 4109).
- Gary McGuire (2009). “An Interview with Professor Sean Dineen”. *Irish Mathematical Society Bulletin* (cit. on p. 4103).
- Julio Medaglia (2013). “Balanço da Bossa Nova”. In: *Balanço da Bossa e Outras Bossas*. Ed. by Augusto de Campos. São Paulo: Perspectiva, pp. 67–124 (cit. on p. 4108).
- Leopoldo Nachbin (1946). “Sur la combinaison des topologies métrisables et pseudo-métrisables”. *Comptes Rendus de l’Académie des Sciences de Paris* 223, pp. 938–940 (cit. on p. 4100).
- (1948). *Espaços Vetoriais Topológicos*. Vol. 4. Rio de Janeiro: Notas de Matemática-Universidade do Brasil (cit. on p. 4100).
  - (1950). “A theorem of the Hahn-Banach type for linear transformations”. *Transactions of the American Mathematical Society* 68, pp. 28–46 (cit. on p. 4101).
  - (1954). “Topological vector spaces of continuous functions”. *Proceedings of the Natural Academy of Science* 40, pp. 471–472 (cit. on p. 4100).
  - (1959). “Algebras of finite differential order and the operational calculus”. *Annals of Mathematics* 70, pp. 413–437 (cit. on p. 4101).
  - (1964). *Lectures on the Theory of Distributions*. Vol. 15. Recife: Instituto de Física Matemática da Universidade do Recife (cit. on p. 4102).
  - (1969). *Topology on Spaces of Holomorphic Mappings*. Berlin: Springer (cit. on pp. 4103, 4104).
- Naomi Oreskes and John Krigé (eds.) (2014). *Science and technology in the global Cold War*. Cambridge, MA: MIT Press (cit. on p. 4095).
- Eduardo Ortiz (2003). “La política interamericana de Roosevelt: George D. Birkhoff y la inclusión de América Latina en las redes matemáticas internacionales - Primera Parte”. *Saber y Tiempo: Revista de Historia de la Ciencia* 4.15, pp. 53–111 (cit. on p. 4107).
- Jacob Palis, César Camacho, and Elon Lages Lima (eds.) (2003). *IMPA – 50 anos*. Rio de Janeiro: IMPA (cit. on pp. 4099, 4104, 4108).
- Karen Hunger Parshall (2009). “Marshall Stone and the Internationalization of the American Mathematical Research Community”. *Bulletin of the American Mathematical Society* 46.3, pp. 459–482 (cit. on p. 4099).

- Anne-Sandrine Paumier (2014). “Laurent Schwartz (1915-2002) et la vie collective des mathématiciens”. PhD thesis. Paris: Université Pierre et Marie Curie (cit. on pp. 4102, 4104).
- Mauricio Peixoto (1959). “On Structural Stability”. *Annals of Mathematics* 69, pp. 199–222 (cit. on p. 4109).
- (1961). “Sobre o Problema Fundamental da Teoria das Equações Diferenciais”. *Atas do 3º Colóquio Brasileiro de Matemática - Fortaleza*, pp. 190–194 (cit. on p. 4106).
  - (1962). “Structural Stability on Two-dimensional Manifolds”. *Topology* 1, pp. 101–120 (cit. on p. 4109).
  - (1987). “Acceptance Speech for the TWAS 1986 award in mathematics”. In: *Proceedings of the Sec. Gen. Conf. Third World Ac. Sci.* Ed. by A. M. Faruqui and M. H. A. Hassan. Beijing: World Scientific, pp. 600–614 (cit. on p. 4109).
  - (2000). “Interview done at IMPA by Tatiana Roque” (cit. on p. 4109).
- Rute da Cunha Pires (2006). “A Presença de Bourbaki na Universidade de São Paulo”. PhD thesis. PUC-SP (cit. on p. 4100).
- Dhruv Raina and Ashok Jain (1997). “Big Science and the University in India”. In: John Krigs and Dominique Pestre (eds.) *Companion to Science in the Twentieth Century*. London and New York: Routledge, pp. 859–877 (cit. on pp. 4098, 4099).
- Tatiana Roque (2007). “De Andronov a Peixoto: a noção de estabilidade estrutural e as primeiras motivações da escola brasileira de Sistemas Dinâmicos”. *Revista Brasileira de História da Matemática* 7, pp. 233–246 (cit. on p. 4105).
- (2016). In: Karine Chemla, Renaud Chorlay, and David Rabouin. *The Oxford Handbook of Generality in Mathematics and the Sciences*. Oxford: Oxford University Press. Chap. Different Notions of Genericity in the Classification Problem of Dynamical Systems, pp. 299–324 (cit. on p. 4110).
- Laurent Schwartz (1957). “Théorie des distributions à valeurs vectorielles”. *Annales de l’Institut Fourier* 7, pp. 1–141 (cit. on p. 4101).
- Reinhard Siegmund-Schultze (2001). *Rockefeller and the Internationalization of Mathematics between the Two World Wars*. Basel/Boston: Birkhäuser (cit. on p. 4099).
- José do Carmo Toledo (2012). “Sobre o processo histórico de institucionalização da área de análise matemática no Brasil”. *Revista Brasileira de História da Matemática* 11, pp. 63–87 (cit. on p. 4104).
- Antonio Pedro Tota (2009). *The Seduction of Brazil: The Americanization of Brazil during World War II*. Austin: University of Texas Press (cit. on p. 4107).
- Ludovic Tournès (2010). “La philanthropie américaine et l’Europe: contribution à une histoire transnationale de l’américanisation”. *Bulletin de l’Institut Pierre Renouvin* 31, pp. 173–187 (cit. on p. 4106).
- Lucieli Trivizoli (2011). “Intercâmbios acadêmicos matemáticos entre EUA e Brasil: uma globalização do saber”. PhD thesis. Rio Claro-SP: UNESP (cit. on p. 4099).

- Antonio Augusto Passos Videira (2004). “Pensando no Brasil: O Nacionalismo entre os Físicos Brasileiros no Período entre 1945 e 1955”. *Saber y Tiempo* 5.18, pp. 71–98 (cit. on p. 4097).
- Cassio Leite Vieira and Antonio Augusto Passos Videira (2014). “Carried by history: César Lattes, nuclear emulsions, and the discovery of the pi-meson”. *Physics in Perspective* 16, pp. 3–36 (cit. on p. 4094).

Received 2017-11-29.

TATIANA ROQUE  
INSTITUTO DE MATEMÁTICA–UFRJ

and

ARCHIVES POINCARÉ  
[tati@im.ufrj.br](mailto:tati@im.ufrj.br)

## ON FRANCO–GERMAN RELATIONS IN MATHEMATICS, 1870–1920

DAVID E. ROWE

### Abstract

The first ICMs took place during a era when the longstanding rivalry between France and Germany strongly influenced European affairs. Relations between leading mathematicians of these two countries were also colored by this tense political atmosphere. This brief account highlights what was at stake by focusing on events in Paris and Göttingen from the period 1870 to 1920.

### Introduction

Last year the Institut Henri Poincaré commemorated the hundredth anniversary of the death of Gaston Darboux, one of the greatest mathematicians of his time. On that occasion I tried to give an idea of how Darboux was viewed by some of his contemporaries who lived outside of France. In this paper, I will expand on that theme in order give a somewhat broader picture of Franco-German mathematical relations during the period bounded by two wars.

As the leading French geometer of his generation, Darboux was admired by many distinguished foreign mathematicians who knew his work well, including Sophus Lie, Julius Weingarten, Luigi Bianchi, and of course Felix Klein. Darboux met Klein and Lie already in 1870, and he corresponded with both quite regularly for many years afterward. He also wrote a warm obituary for Lie, [Darboux \[1899\]](#) after the latter's death in 1899. So it might seem at first rather surprising that in 1917 it was not Klein, but rather his Göttingen colleague, David Hilbert, who wrote an obituary for Darboux, [Hilbert \[1917\]](#). Perhaps even more surprising, given the ongoing slaughter on the battlefields, is that any mathematician in Germany would have chosen to write in honor of an esteemed French colleague at that time. In his autobiography [Schwartz \[2001\]](#), Laurent Schwartz commented that this would have been unthinkable in France, though I believe Schwartz was probably wrong

when he wrote that chauvinism during the First World War was greater in France than in Germany.

Many French mathematicians would have known about Hilbert's tribute to Darboux because it was reprinted in 1935 in the third volume of his collected works. Some would have read it long before, since Mittag-Leffler published a French translation in *Acta Mathematica* already in 1919. I will return to Hilbert's wartime éloge for Darboux momentarily, but first let me say a few words about relations between French and German mathematicians in the wake of the Franco-Prussian War. Given the fact that Émile Picard's father died during the siege of Paris and that Paul Appell's family fled from Strasbourg to Nancy, one can easily imagine the impact the war had on their views of the new German state. Henri Poincaré and his family worried that a similar fate would befall Nancy when it was under occupation. A German military official was then stationed at their home, which gave the 16-year-old Poincaré the chance to pick up spoken German. In his *Dernières pensées* he wrote:

When asked to justify rationally our love of country, we can be quite embarrassed, but our mind imagines our defeated armies, France invaded, we feel altogether nauseous, tears begin to flow and we listen no further. And if there are those today who repeat so many sophisms, it is most likely due to their lack of imagination. They are unable to imagine by themselves all this suffering, and if misfortune or some divine punishment fixed their eyes upon it, their soul would revolt as does our own. [Poincaré \[1913\]](#)

In Darboux's case, I suppose he probably felt no differently, though he came, of course, from an older generation.

The Franco-Prussian War surely did have an impact on mathematical relations afterward, but of course its political fallout was nothing like the damage caused by the Great War. Klein already wrote a friendly letter to Darboux in February 1871, after which their mathematical correspondence took up a whole series of common interests (see [Tobies \[2016\]](#)). One can also read Hermite's words of praise after he visited Göttingen in 1877 to attend the Gauss celebration. Probably no French mathematician could match Hermite's enthusiasm for German mathematics, despite his difficulties with the language (see [Archibald \[2002\]](#)). By the 1880s, a handful of younger French mathematicians were going abroad to study at leading German universities. One of these was Paul Painlevé, who spent a year in Göttingen attending lectures offered by H. A. Schwarz and Klein. During the First World War when he served briefly as War Minister, Painlevé spoke out strongly against the German war machine. His words were long remembered by German mathematicians after the war.

## Klein and Lie in Paris

Just before the Franco-Prussian War broke out in mid-July 1870, Klein and Lie sent a status report on French mathematics to the Mathematics Club at the University of Berlin. This contains many remarkable things, particularly when we consider that Klein was only twenty-one years old when he wrote it. The report also contains these remarks about Darboux's new journal, *Bulletin des Sciences Mathématiques et Astronomiques*:

We believe that such a journal is a very useful, but also very difficult undertaking whose goal can be fully achieved only if it has a large number of contributors who are well versed in the areas on which they report. The Bulletin is not yet in this fortunate situation. Nor, indeed, is it difficult to find evidence, in the issues which have appeared so far, of a number of flawed reviews. But the personality of the editor, G. Darboux, a man whom we consider exceptionally gifted (and whose gifts are precisely suited to this cause), seems to us to ensure that the Bulletin will continue to improve with time.<sup>1</sup>

Klein and Lie found that Darboux's reviews “stand out in their expertise and clear exposition,” and they compared these favorably with those written by Jules Hoüel for the *Nouvelles Annales*. They went on to underscore their support for Darboux's undertaking, noting that his goal was “to familiarize French mathematicians with the modern branches of geometry and algebra, which have been relatively unknown in France up to now.”

Many of you will know the story about how Darboux travelled to Fontainebleau to free Sophus Lie from prison. The Norwegian had been detained there at the outset of the war on suspicions that he might have been a German spy. Lie happened to be carrying letters from Klein, written in what seemed like a strange German code language with words like Linien- und Kugelkomplex, etc. Darboux wrote about this incident in his obituary for Lie, noting that he was relieved on meeting him to see that his friend was not at all angry with the police who had arrested him [Darboux \[1899\]](#). Soon after his release, Lie wrote to a friend, “the sun has never seemed to me to have shone so clearly, the trees have never been so green as those I saw yesterday as a free man on my way to the Fontainebleau station” ([Stubhaug \[2002\]](#), p. 147).

## Lie's Line-to-Sphere Mapping

Only shortly before this time, Lie had found his famous line-to-sphere mapping, a contact transformation with many interesting properties. A pretty example comes from the image

---

<sup>1</sup>Appendix in *Letters from Felix Klein to Sophus Lie, 1870-1877*, Heidelberg: Springer-Verlag, scheduled to appear in 2019. For a detailed account of the first five years of Darboux's *Bulletin*, see [Croizat \[2016\]](#), 470-550.

of a quadric surface given by one of its families of generators. These lines map to a family of spheres that envelopes a Dupin cyclide. One can picture this most easily by taking three skew lines in space which then map to three spheres. The set of lines that meet these mutually skew lines form the generators of a quadric surface, and since Lie's mapping is a contact transformation these lines go over to a one-parameter family of spheres tangent to three fixed spheres. Since the second system of generators has the same property, the analogy with a Dupin cyclide becomes clear: these are surfaces enveloped by two families of spheres. Moreover, this mapping has the property that the asymptotic curves of the first surface go over to the curvature lines of the second. In this case, the generators themselves are the asymptotic curves, and these then correspond to the circles of tangency of the Dupin surface (see [Lie and Scheffers \[1896\]](#), pp. 470-475). There is an important connection here with Darboux's mathematics that I should briefly mention.

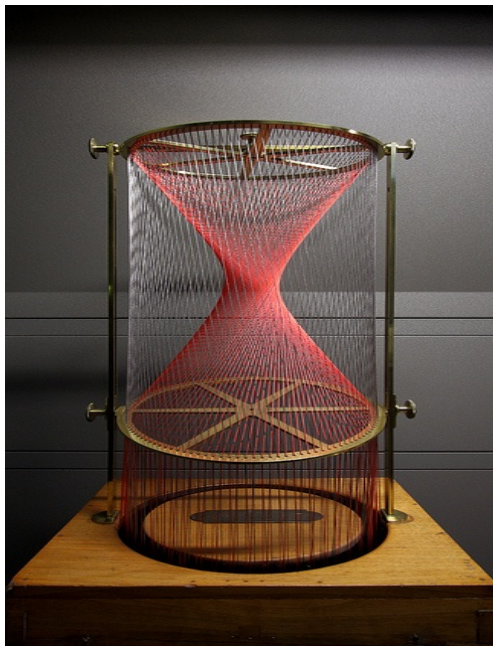


Figure 1: Historical model of a hyperboloid of one sheet by Theodor Olivier.

In 1864 Darboux and Theodore Moutard began work on generalized cyclides, which they studied in the context of inversive geometry.<sup>2</sup> Klein and Lie learned about this new

<sup>2</sup>For a detailed account of this theory and related work by Darboux and others, see [Croizat \[2016\]](#).



Figure 2: Historical model of a cyclide from the Brill collection.

French theory when they met Darboux just before the Franco-Prussian War broke out. Darboux later developed the theory of generalized cyclides by introducing pentaspherical coordinates [Darboux \[1873\]](#).<sup>3</sup> These cyclides are special quartic surfaces with the property that they meet the plane at infinity in a double curve, namely the imaginary circle that lies on all spheres. Darboux also found that their lines of curvature are algebraic curves of degree eight. This finding set up one of Lie's earliest discoveries, communicated to Darboux at that time. This came from Lie's line-to-sphere mapping, when he considered the caustic surface enveloped by lines in a congruence of the second order and class [Rowe \[1989\]](#).

A few years earlier, these special quartics had been studied by the Berlin mathematician Kummer, and they soon came to be called Kummer surfaces. Lie found that these Kummer surfaces will map to the generalized cyclides of Darboux, and since the curvature lines of the latter were known, he immediately deduced that the asymptotic curves on a Kummer surface are algebraic of degree sixteen. In early July 1870, Lie communicated these findings to the Norwegian Scientific Society in Christiania, but this note was only published by Ludwig Sylow in 1899, the year of Lie's death ([Lie \[1934\]](#), pp. 86-87). Lie and Klein discussed this breakthrough in detail, as Klein gradually came to understand

<sup>3</sup>Darboux had already worked out many of these ideas when Klein and Lie met him, but it took him another two years to develop the whole theory in detail and publish it in [Darboux \[1873\]](#). For further background on his early career and mathematical research, see [Croizat \[2016\]](#).





Figure 3: Klein's model of a Kummer surface designed in 1871. Courtesy of the Collection of Mathematical Models, Göttingen University.

Lie's line-to-sphere mapping. He then quickly realized that Lie's claim was correct because he had already come across these same curves of degree sixteen in his own work. Klein had found these curves while studying quadratic line complexes that share the same Kummer surface as their surface of singularity, but he had not realized that they were its asymptotic curves. Little more than a week later, Klein had to flee from Paris, but he soon reported to Lie that he was able to trace the paths of these asymptotic curves and to describe their singularities. He did so by studying a physical model of a Kummer surface made by his friend Albert Wenker.

Klein imparted this information to Lie in a letter from 29 July, possibly one of those that left the sentries in Fontainebleau suspicious when Lie tried to tell them this was only mathematics. By this time, Klein realized that what he had told Lie earlier in Paris about the singularities of these asymptotic curves was, in fact, incorrect. After giving the necessary corrections, he wrote:

I came across these things by means of Wenker's model, on which I wanted to sketch asymptotic curves. To give you a sort of intuitive idea how such curves look, I enclose a sketch. The Kummer surface contains hyperboloid parts, like those sketched; these are bounded by two of the six conics ( $K_1$  and  $K_2$ ) and extend from one double point ( $d_1$ ) to another ( $d_2$ ). Two of the curves are drawn more boldly; these are the two that not only belong to linear complexes but also are curves with four-point contact. They pass through  $d_1$



Figure 4: Klein's sketch of the asymptotic curves between two double points on a Kummer surface from his letter to Sophus Lie.

and  $d_2$  readily, whereas the remaining curves have cusps there. This is also evident from the model. At the same time, one sees how  $K_1$  and  $K_1$  are true enveloping curves.<sup>4</sup>

By the “hyperboloid parts” on a Kummer surface, Klein meant those places where the curvature was negative. Only in these regions were the asymptotic curves real and hence visible. He later reproduced the same figure in the note that he and Lie sent to Kummer for publication in the *Monatsberichte* of the Prussian Academy Klein and Lie [1870].

<sup>4</sup>Klein to Lie, 29 July 1870, *Letters from Felix Klein to Sophus Lie, 1870–1877*, Heidelberg: Springer-Verlag, scheduled to appear in 2019.

## Lie on Parisian Mathematics in 1882

These were obviously exciting times, both mathematically and politically, and for many years afterward Klein and Lie hoped to meet again in Paris. In 1882, that chance finally looked possible, but then Klein suffered a major collapse in his ongoing effort to compete with Henri Poincaré, who published five long papers on the theory of automorphic functions in Mittag-Leffler's new journal, *Acta Mathematica*. So Lie visited Paris on his own, but he reported on his various conversations with French mathematicians in three letters written to Klein. I would like to cite just a few passages from these letters, which give a vivid picture of how Darboux and others were seen by Lie at the time. Private correspondence has, of course, the decided advantage that people will write things they would never put in print, least of all in an obituary for a distinguished figure.

So here is a little gossip from Paris in 1882 [Rowe \[2018b\]](#):

I have spoken now with Hermite about all kinds of things. He has a very amiable nature, but I still don't know how much of it is genuine. People here say it is certain that he can't read a word of German, which indeed explains a number of things. The most remarkable thing he said was the following (which I communicate to you in confidence): Mittag-Leffler told him that the German mathematicians hate the French mathematicians. Nor did he want to hear anything of my protests against this. That is certainly strong. He was eager to hear about friction between German mathematicians, whereas he described the situation in Paris as idyllic in this regard. Probably it is no better in Paris than in Germany.

In another letter, Lie wrote that he regretted telling Klein about Mittag-Leffler's remark, which, in any case, was clearly an overstatement. The rivalries within Germany were at this time in many cases more significant than those between leading mathematicians of these two countries. Klein continually advised young Germans to visit Paris, and quite a few of his students and protégés did so, including Ferdinand Lindemann, Walther Dyck, Eduard Study, and Hilbert. We should also keep in mind that Darboux and Jordan had both met Klein and Lie back in 1870 when their ideas had considerable influence on Klein's Erlangen Program (see [Rowe \[1989\]](#)). This circumstance makes the following remarks by Lie quite surprising:

Poincaré mentioned on one occasion that all of mathematics was a matter of groups. I told him about your [Erlangen] Program, which he did not know. Halphen, Darboux, and Stephanos spoke with the highest praise about you. Until now I have spoken very little with Jordan, whose mother died recently.

And then a little later, Lie wrote:

. . . In the meantime I have spoken at length with Jordan. He finds your investigations difficult to understand. Poincaré said that at first it was hard for him to read your work, but that now it goes very easily. A number of mathematicians, such as Darboux and Jordan, say that you make great demands on the reader in that you often do not supply proofs. I am trying to report on this as correctly as possible.

Lie also commented on how leading French mathematicians reacted to his own work:

So far as my own things go, I am more or less satisfied. Darboux has studied my work with remarkable thoroughness. This is good insofar as he has given gradually more lectures on my theories at the Sorbonne, for example on line and sphere geometry, contact transformations, and first-order partial differential equations. The trouble is that he continually plunders my work. He makes inessential changes and then publishes these without mentioning my name. Now he is starting on the surfaces of constant curvature. I must therefore rework my papers from Christiania [present day Oslo] for the *Mathematische Annalen* just as soon as possible.

Finally, there is this remark about Victor Mannheim, the inventor of the modern slide rule:

Mannheim is friendly as always. He is really a good fellow and warns me constantly about Darboux, for which he really has good reason. But I must speak with Darboux, as he is the one who understands me the best. And for the pleasure I must pay something. In any case, he is promoting mathematical science.

## Franco-German Relations after the Outbreak of the Great War

Darboux later had several dealings with Klein, as both were highly involved in promoting various national and international projects. A year after the Paris ICM, Darboux succeeded Joseph Bertrand as perpetual secretary for the mathematical sciences section of the academy. By then, he and Klein had emerged as the two most active and visible mathematicians in their respective countries. Their contacts ended, however, with the outbreak of the Great War. Let us recall how, right at its outset, 93 German intellectuals attached their names to a manifesto that proudly announced full support of all actions taken by the German army, beginning with its invasion of Belgium. Quite a number of the signatories were prominent natural scientists – including Max Planck, Fritz Haber, Walther Nernst, Ernst Fischer, and Ernst Haeckel – whereas only one mathematician appeared on the list:

Felix Klein. We know rather little about how these names were collected, so it is unclear whether Hilbert actually withheld his support. I'm unaware of any evidence that he was contacted at all. On the other hand, if he had been asked to lend his name, he probably would have refused. Klein was reached by telephone and gave his support without ever having read the document. When it was released, the French Academy dropped his name from its membership roles. There was debate, in fact, over whether all Germans should be dismissed from the academy, but such action was not taken, so Hilbert remained a foreign member. To the best of my knowledge, no actions were taken by any of the German academies against French members.

At this time, Klein had begun his wartime lectures on the history of mathematics in the nineteenth century. These later circulated in mimeographed form and were eventually published in [Klein \[1926\]](#), one year after Klein's death. They offer a highly personal view, as seen from Klein's own vantage point within the Göttingen tradition. National rivalries were also a major theme, and Klein underscored the significance of the École Polytechnique as a model for several polytechnical institutes that followed as well as for applied mathematics in general. He alluded to Jacobi's remarkable lecture in praise of this new institution, a speech delivered at a time when Parisian mathematics stood at its peak. Klein had cultivated friendly relations with Darboux, Poincaré, and other leading French mathematicians [Tobies \[2016\]](#), but his high respect for French achievements in no way diminished his nationalism. Like nearly all Germans of his generation, Klein celebrated the Battle of Sedan as the key event that led to the unification of Germany. No doubt, he attached significance to the fact that the crowning of King Wilhelm of Prussia as Emperor took place in the Hall of Mirrors in Versailles rather than in his palace in Berlin.

Before the war, Darboux and Klein had worked together to help found the International Association of Academies. Such cooperation was obviously unthinkable in wartime, when the French and German scientific communities felt only hostility for one another. One of the most prominent Italian mathematicians, Vito Volterra, sided strongly with the French cause (see [Mazliak and Tazzioli \[2009\]](#).) In November 1916, Darboux wrote to the physicist Arthur Schuster, Secretary of the Royal Society, suggesting a meeting of leading scientists from the Entente powers to address what should be done with regard to international relations after the war ([Lehto \[1998\]](#), p. 16). Since he died in February 1917, nothing came of this initiative, though his successor, Émile Picard, took up this matter before the fighting had come to an end. Picard's attitude toward German scientists was similar to Clemenceau's view of German politicians, and if we remember that Briand and Stresemann were able to overcome the intransigence that hampered the early years of the Weimar Republic, we might think of Hilbert as the leading representative of rapprochement in the world of mathematicians. His call to break the counter-boycott of the Bologna ICM in 1928 led to a direct conflict with Bieberbach, but even more with Brouwer, who had become his arch-nemesis ([Blumenthal \[1935\]](#), p. 427). These events only underscore

Hilbert's longstanding commitment to internationalism, but let me return to his portrait of Darboux.

### Hilbert's Tribute to Darboux

After Darboux's death in February 1917, he received a fitting eulogy from the Göttingen Scientific Society. Darboux belonged to that body as a foreign member since 1901 when he succeeded Hermite. As I noted at the outset, it would seem at first surprising that Hilbert wrote this obituary, [Hilbert \[1917\]](#), considering that Klein had known Darboux far better, both personally and mathematically, than had his younger colleague. In fact, this was an altogether unusual tribute to a foreign scholar, and it caused an immediate stir within Göttingen academic circles.

We should first of all note that the Göttingen Scientific Society did not ordinarily honor its deceased foreign members in this way. Neither Klein nor Hilbert wrote an obituary for Hermite or for any other foreign member of the society, so there was certainly no compulsion for Hilbert to take up his pen to eulogize Darboux. One can hardly escape the conclusion that his motivation was, in large part, political, though personal gratitude could also have played a role as well. In 1905, Darboux and Klein had been charged with the difficult task of judging who should be awarded the first Bolyai Prize. What made this decision difficult was the personal and national prestige involved, since only two names needed to be taken into consideration: Poincaré and Hilbert. Klein naturally favored Hilbert, just as Darboux supported Poincaré, who was awarded the prize. But Darboux apparently agreed that he would back Hilbert's nomination for the second prize in 1910, and on that occasion Poincaré himself wrote the report in support of this decision.

Still, Hilbert surely had other reasons for writing this obituary, and the more likely motives would have been connected to his special place within the larger academic community in Göttingen. Hilbert's outspoken internationalist views had led to many open clashes, both within the philosophical faculty as well as the scientific society. If we take into account that Hilbert's text was presented at the society's annual public session, held on 12 May 1917, then his political motivation becomes even more obvious. Clearly, he knew that his speech was bound to provoke great controversy, and just as surely this was his very intention. According to his biographer, Constance Reid, when word got out in Göttingen about the Darboux *Nachruf*, an angry band of students gathered in front of Hilbert's house to demand that he withdraw the text ([Reid \[1970\]](#), p. 145). Reid's rather romantic account was largely based on oral interviews, so we cannot be too confident about the details of this incident, particularly her claim that Hilbert threatened to resign his position if he did not receive an official apology from the Rector of the university for the behavior of these students. Still, even if the details cannot be corroborated, there is every reason

to believe that something like this happened. In fact, two years later Hilbert did seriously consider resigning his professorship to accept a chair in Bern. Without doubt, his *éloge* for Darboux was intended not only to honor a great mathematician; it was also meant as a direct provocation to all those in the Göttingen Scientific Society who saw French scholars as their mortal enemies.

Hilbert began by praising Darboux and Camille Jordan for their universality, which he claimed had opened the way for a younger generation of mathematicians who no longer felt hemmed in by the special disciplines that dominated most research during the late nineteenth century. This universal outlook had become a watchword for Göttingen mathematics, so by identifying Darboux with it, Hilbert underscored the intellectual affinities that linked him with them. He then recalled how Darboux, in his plenary address at the 1908 ICM in Rome, had compared his own era with the new trends that were unfolding since the turn of the century. For Hilbert, it had been personally gratifying that Darboux brought up his famous speech on “Mathematical Problems” from the Second ICM held in Paris in 1900. He was also pleased to recall how this older representative of French mathematics, though by now only an outside observer, had spoken up in support of these radically new developments. In short, he saw Darboux as a progressive spirit.

Following these introductory remarks, Hilbert alluded to the various phases in Darboux’s career, starting with the impression he made already as a student on his countrymen. Here he recalled the well-known anecdote about how, after scoring first among all candidates for both *écoles*, Darboux chose to attend the *École Normale*. In Hilbert’s telling, though, we can easily hear echoes of an anti-militarist theme. He noted that Darboux grew up in modest circumstances and that he lost his father at a young age. About his decision not to attend the *École Polytechnique*, Hilbert wrote that Darboux chose “to decline the sword and gold-embroidered cloak of an officer or civil engineer in preference for the more humble title of a professor and the less distinguished teaching profession . . . , Hilbert [1917, p. 366].” He added that this “was something that had never before occurred and that awakened general astonishment” at the time, citing an article by the “then famous French Goethe-expert Jean-Jacques Weiss” Weiss [1861]. Probably no one in the audience knew this name, but one can almost imagine their faces when they heard the most famous living German mathematician refer to some obscure Frenchman as a famous expert on Goethe. Hilbert went on to say that Weiss wanted to record Darboux’s decision in order to show “how at least once something like this had occurred on our planet.”

Not surprisingly, Hilbert mentioned Darboux’s influence on Lie and Klein, but he also briefly described three dissertations written in Göttingen that were inspired by lesser known works by Darboux. When, toward the end, he turned to Darboux’s four-volume *Theorie des surfaces*, his praise for it was almost boundless. He called this not only a classic work for surface theory but also an invaluable tool for studying mechanics, calculus of variations, partial differential equations, and invariant theory. Moreover, in Hilbert’s



view, no one before Darboux had recognized the deep connections between these fields of central importance for contemporary research. Darboux's treatise, he wrote, belongs in the library of every mathematician, like other great works by French authors, such as Jordan's *Cours d'Analyse*, Picard's *Traité d'Analyse*, and Poincaré's *Mécanique celeste*. The message could hardly have been clearer – these works belong to all mathematicians because the world of mathematics knows no national boundaries. Hilbert even alluded to the relevance of Darboux's work for Einstein's new theory of gravitation, truly a new theory in 1917.

### Einstein and Hilbert as Leading Internationalists

Einstein first met Hilbert when he came to Göttingen in 1915 to give a series of lectures. The not yet famous physicist knew that Hilbert was a brilliant mathematician, but he now came to realize he was also an outspoken internationalist who was unafraid to clash with opponents. Emmy Noether's attempt to habilitate had begun at this time, and Einstein was well aware of Hilbert's efforts to promote her case. She taught special courses under his name, but could only gain an official appointment after the war ended, since the Prussian Ministry of Education quashed all proposals to allow women such rights before then.<sup>5</sup>

In the final year of the war, Einstein contacted Hilbert to propose that they join hands with like-minded colleagues from other countries in order to make the case for peace and moral progress. He began this appeal with these words:

Countless times in these desolate years of general nationalistic delusion, men of science and the arts issued statements to the public that have already inflicted incalculable damage to the feeling of solidarity that had been developing with such promise before the war . . . The hue and cry of straight-laced preachers and servants of the bleak principle of power is becoming so loud and public opinion is being misled to such a degree by methodical silencing of the press that those with better intentions, feeling wrtechedly isolated, do not dare to raise their voices. (Einstein to Hilbert, before 27 April 1918, Schulmann, Kox, Janssen, and Illy [1998]).

After consulting with some friends, Hilbert replied that in his opinion the time was not yet ripe for Einstein's "well-meaning and appealing undertaking." In fact, he warned that any such declarations "would be tantamount to self-denunciations, which all our enemies in the faculties would be extremely glad to cite." For them, "the very word 'international' is like a red flag for a bull." Hilbert also cautioned Einstein against:

---

<sup>5</sup>On the events and conflicts associated with this effort, see Tollmien [1990]. Noether's wartime contributions to general relativity are discussed in Rowe [n.d.].



firing off our gunpowder at the wrong time and possibly also at the wrong persons. . . . I would like to recommend waiting until the mad hurricane has spent itself and reason has the opportunity of returning – and this time is sure to come. We would have to restrict ourselves to the German professors, since they alone are thoroughly known to us here and also have most to do with it. Other peoples must wash their own dirty laundry. (Hilbert to Einstein, 1 May 1918, [Schulmann, Kox, Janssen, and Illy \[1998\]](#)).

Nothing came of this venture, but the exchange reveals that Hilbert had a far better feel for academic opinion in Germany at this time than did Einstein. He was also a firm advocate of academic freedom and a defender of those, such as Leonard Nelson, who came under attack during wartime for their pacifist views. Around the same time that he wrote this letter to Einstein, Hilbert informed Klein that he refused to attend future meetings of the Göttingen Academy, so long as no one besides him was willing to protest against the behavior of its secretary, Edward Schröder. Hilbert was incensed that Schröder had taken it upon himself to inform military authorities about the pacifist views of a colleague in physics. After the war, he took steps to force Schröder's resignation, though without success; the latter remained secretary of the philological-historical section until 1924.

These few remarks offer a glimpse of the atmosphere in Göttingen during wartime as well as some of the events that help to place Hilbert's obituary for Darboux in its original context. Let me end by quoting Hilbert again, this time from the year 1909 when Poincaré came to Göttingen to deliver the first series of Wolfskehl lectures. Here are a few words taken from Hilbert's welcoming address on that occasion [Rowe \[2018a, p. 197\]](#):

You know, highly honored colleague, as do we all, how steady and close the mathematical interests of France and Germany have been and continue to be. Even when we recall only quickly the developments of the recent past, and out of the rich and many-voiced concert of mathematical science we take hold of the two fundamental tones of number theory and function theory, then we think perhaps of Jacobi, who had in Hermite the outstanding heir to his arithmetical ideas. And Hermite, who unfolded the flag of arithmetic in France, had our Minkowski, who brought it back to Germany again. Or if we only think of the names Cauchy, Riemann, Weierstrass, Poincaré, Klein, and Hadamard, these names build a chain whose links join one another in succession. The mathematical threads tying France and Germany are, like no two other nations, diverse and strong, so that from a mathematical perspective we may view Germany and France as a single land.

## References

- Thomas Archibald (2002). “Charles Hermite and German Mathematics in France”. *Mathematics Unbound. The Evolution of an International Mathematical Research Community (1800-1945)*. K. H. Parshall and A. C. Rice, eds., Providence, RI: American Mathematical Society, pp. 123–137 (cit. on p. [4114](#)).
- Otto Blumenthal (1935). “Lebensgeschichte”. *David Hilbert, Gesammelte Abhandlungen* 3, pp. 388–429 (cit. on p. [4122](#)).
- Barnabé Croizat (2016). “Gaston Darboux: naissance d’un mathématicien, genèse d’un professeur, chronique d’un rédacteur”. PhD thesis. Université Lille 1 - Sciences et Technologies (cit. on pp. [4115–4117](#)).
- Gaston Darboux (1873). *Sur une Classe Remarquable de Courbes et de Surfaces Algébriques et sur la Théorie des Imaginaires*. Gauthier-Villars, Paris (cit. on p. [4117](#)).
- (1899). “Sophus Lie”. *Bulletin of the American Mathematical Society* 5.7, pp. 367–370 (cit. on pp. [4113](#), [4115](#)).
- David Hilbert (1917). “Gaston Darboux, Nachrichten der Königlichen Gesellschaft der Wissenschaften zu Göttingen, Geschäftliche Mitteilungen”. French translation in *Acta Mathematica* 42 (1919): 269–273. Reprinted in *David Hilbert Gesammelte Abhandlungen*, Bd. 3, 365–369, pp. 71–75 (cit. on pp. [4113](#), [4123](#), [4124](#)).
- (1935). *David Hilbert Gesammelte Abhandlungen*.
- Felix Klein (1926). *Vorlesungen über die Entwicklung der Mathematik im 19. Jahrhundert*. Vol. 1. Springer, Berlin. MR: [529278](#) (cit. on p. [4122](#)).
- Felix Klein and Sophus Lie (1870). “Über die Haupttangentialkurven der Kummerschen Fläche vierten Grades mit sechzehn Knotenpunkten”. *Monatsberichte der Preussischen Akademie der Wissenschaften*. Reprinted in S. Lie, *Gesammelte Abhandlungen*, Bd. 1, pp. 97–104 (cit. on p. [4119](#)).
- Olli Lehto (1998). *Mathematics without borders*. A history of the International Mathematical Union. Springer-Verlag, New York, pp. xvi+399. MR: [1488698](#) (cit. on p. [4122](#)).
- Sophus Lie (1934). *Gesammelte Abhandlungen*. Bd I. Hrsg. von der Akademie der Wissenschaften zu Leipzig und dem Norwegischen Mathematischen Verein durch Friedrich Engel und Poul Heegaard. B. G. Teubner, Leipzig, H. Aschehoug & Co., Oslo, pp. x+476. MR: [0122662](#) (cit. on p. [4117](#)).
- Sophus Lie and Georg Scheffers (1896). *Geometrie der Berührungstransformationen*. With editorial assistance by Georg Scheffers. Leipzig: Teubner, pp. xi+694. MR: [0460049](#) (cit. on p. [4116](#)).
- Laurent Mazliak and Rossana Tazzioli (2009). *Mathematicians at war*. Vol. 22. Archimedes: New Studies in the History and Philosophy of Science and Technology. Volterra and his French colleagues in World War I, With a preface by Ivor Grattan-Guinness. Springer, New York, pp. x+194. MR: [2583733](#) (cit. on p. [4122](#)).

- Henri Poincaré (1913). *Dernières pensées*. Ernest Flammarion, Paris (cit. on p. [4114](#)).
- Constance Reid (1970). *Hilbert*. New York: Springer-Verlag (cit. on p. [4123](#)).
- David E. Rowe (n.d.). “Emmy Noether’s Role in the Relativity Revolution”. To appear in *The Mathematical Intelligencer* (cit. on p. [4125](#)).
- (1989). “Klein, Lie, and the Geometric Background of the Erlangen Program”. In: *The history of modern mathematics. Vol. I, Proceedings of the symposium held at Vassar College, Poughkeepsie, New York, June 20–24, 1989*. Ed. by David E. Rowe and John McCleary. Modern Mathematics: Ideas and their reception. Academic Press, Inc., Boston, MA, pp. xvi+453. MR: [1037792](#) (cit. on pp. [4117](#), [4120](#)).
  - (2018a). “Poincaré Week in Göttingen, 22–28 April 1909”. In: *A Richer Picture of Mathematics. The Göttingen Tradition and Beyond*. New York: Springer, pp. 195–202 (cit. on p. [4126](#)).
  - (2018b). “Three letters from Sophus Lie to Felix Klein on Mathematics in Paris”. In: *A Richer Picture of Mathematics. The Göttingen Tradition and Beyond*. New York: Springer, pp. 105–109 (cit. on p. [4120](#)).
- Robert Schulmann, Anne J Kox, Michel Janssen, and József Illy (1998). *The collected papers of Albert Einstein. The Berlin years: Correspondence 1914–1918*. Vol. 8. Princeton University Press (cit. on pp. [4125](#), [4126](#)).
- Laurent Schwartz (2001). *A mathematician grappling with his century*. Translated from the 1997 French original by Leila Schneps. Birkhäuser Verlag, Basel, pp. viii+490. MR: [1821332](#) (cit. on p. [4113](#)).
- Arild Stubhaug (2002). *The mathematician Sophus Lie*. It was the audacity of my thinking, Translated from the 2000 Norwegian original by Richard H. Daly. Springer-Verlag, Berlin, pp. xii+555. MR: [1876528](#) (cit. on p. [4115](#)).
- Renate Tobies (2016). “Felix Klein und französische Mathematiker”. *Silvia Schöneburg, Hrsg.: Mathematik von einst für jetzt; Festschrift für Karin Richter, Thomas Krohn* (cit. on pp. [4114](#), [4122](#)).
- Cordula Tollmien (1990). “Biographie der Mathematikerin Emmy Noether (1882 – 1935) und zugleich ein Beitrag zur Geschichte der Habilitation von Frauen an der Universität Göttingen”. *Göttinger Jahrbuch* 38, pp. 153–219 (cit. on p. [4125](#)).
- Jean-Jacques Weiss (1861). “France, 19 Novembre 1861”. *Journal des débats politiques et littéraires* (cit. on p. [4124](#)).

Received 2017-09-01.

## Index of Authors

Abramovich, Dan	Vol. 2, 523	Cantat, Serge	Vol. 2, 619
Ambainis, Andris	Vol. 4, 3265	Caporaso, Lucia	Vol. 2, 635
Andersen, Jørgen	Vol. 3, 2541	Careaga, Julio	Vol. 4, 3489
Andoni, Alexandr	Vol. 4, 3287	Cartis, Coralia	Vol. 4, 3711
Andreatta, Fabrizio	Vol. 2, 249	Castro, Manuel J.	Vol. 4, 3515
André, Yves	Vol. 2, 277	Chelkak, Dmitry	Vol. 4, 2801
Arakawa, Tomoyuki (荒川 知幸)	Vol. 2, 1263	Chen, Jungkai A. (陳榮凱)	Vol. 2, 653
Araujo, Carolina	Vol. 2, 547	Chen, Meng (陈猛)	Vol. 2, 653
Argyros, Spiros A.	Vol. 3, 1477	Córdoba, Diego	Vol. 3, 2193
Aschenbrenner, Matthias	Vol. 2, 1	Degond, Pierre	Vol. 4, 3925
de la Asunción, Marc	Vol. 4, 3515	Delort, Jean-Marc	Vol. 3, 2241
Babai, László	Vol. 4, 3319	DeMarco, Laura	Vol. 3, 1867
Balogh, József	Vol. 4, 3059	Demeter, Ciprian	Vol. 3, 1539
Bartels, Arthur	Vol. 2, 1041	Deng, Yu	Vol. 3, 2137
Bedrossian, Jacob	Vol. 3, 2137	De Philippis, Guido	Vol. 3, 2215
Belavin, Alexander	Vol. 3, 2567	Díaz, Lorenzo	Vol. 3, 1887
Bergeron, Nicolas	Vol. 2, 831	Diehl, Stefan	Vol. 4, 3489
Berndtsson, Bo	Vol. 2, 859	Di Francesco, Philippe	Vol. 3, 2581
Bertozzi, Andrea	Vol. 4, 3865	Dinh, Tien-Cuong	Vol. 3, 1561
Birkar, Caucher	Vol. 2, 565	Dolbeault, Jean	Vol. 3, 2261
Bishop, Christopher	Vol. 3, 1511	van den Dries, Lou	Vol. 2, 1
Bochi, Jairo	Vol. 3, 1825	Duminil-Copin, Hugo	Vol. 4, 2829
van der Boor, Mark	Vol. 4, 3893	Du, Qiang (杜强)	Vol. 4, 3541
Borst, Sem	Vol. 4, 3893	Ekholm, Tobias	Vol. 2, 1063
Bosch Casabò, Marianna	Vol. 4, 4015	El Karoui, Noureddine	Vol. 4, 2857
Boucksom, Sébastien	Vol. 2, 591	Esedoğlu, Selim	Vol. 4, 3947
Bourgade, Paul	Vol. 4, 2759	Esteban, Maria J.	Vol. 3, 2261
Bowen, Lewis	Vol. 3, 1847	Exel, Ruy	Vol. 3, 1583
Bresch, Didier	Vol. 3, 2167	Fall, Mouhamed	Vol. 3, 1613
Bühlmann, Peter	Vol. 4, 2785	Fargues, Laurent	Vol. 2, 291
Bürger, Raimund	Vol. 4, 3489	Fayad, Bassam	Vol. 3, 1909
		Fernández Nieto, Enrique D.	Vol. 4, 3515

Finkelberg, Michael	Vol. 2, 1283	Koropecski, Andres	Vol. 3, 1995
Fujiwara, Koji (藤原 耕二)	Vol. 2, 1087	Krikorian, Raphaël	Vol. 3, 1909
Futorny, Vyacheslav	Vol. 2, 1303	Kucharz, Wojciech	Vol. 2, 719
Gallardo, José M.	Vol. 4, 3515	Kurdyka, Krzysztof	Vol. 2, 719
Garnier, Josselin	Vol. 4, 2877	Landim, Claudio	Vol. 3, 2617
Geiß, Christof	Vol. 2, 99	Lassas, Matti	Vol. 4, 3751
Gelander, Tsachik (צחיק גלנדר)	Vol. 2, 1321	Lasserre, Jean	Vol. 4, 3773
Giga, Yoshikazu (儀我美一)	Vol. 3, 2287	Le, Can	Vol. 4, 2925
Giles, Michael	Vol. 4, 3571	van Leeuwen, Johan S. H.	Vol. 4, 3893
González Vida, José M.	Vol. 4, 3515	Levina, Elizaveta	Vol. 4, 2925
Gouëzel, Sébastien	Vol. 3, 1933	Lipshitz, Robert	Vol. 2, 1153
Gould, Nicholas I. M.	Vol. 4, 3711	Liverani, Carlangelo	Vol. 3, 2643
Gubinelli, Massimiliano	Vol. 3, 2311	Logunov, Alexander (Александр Логунов)	
Guillarmou, Colin	Vol. 3, 2339		Vol. 3, 2391
Hacking, Paul	Vol. 2, 671	Lopes Filho, Milton	Vol. 3, 2519
Haydon, Richard	Vol. 3, 1477	Loss, Michael	Vol. 3, 2261
He, Xuhua (何旭华)	Vol. 2, 1345	Macias, Jorge	Vol. 4, 3515
Hochman, Michael	Vol. 3, 1949	Mądry, Aleksander	Vol. 4, 3361
van der Hoeven, Joris	Vol. 2, 1	Majdoub, Mohamed (محمّد المجدوب)	
Hryniewicz, Umberto	Vol. 2, 941		Vol. 3, 2413
Huh, June	Vol. 4, 3093	Malinnikova, Eugenia (Евгения	
Indyk, Piotr	Vol. 4, 3287	Малинникова)	Vol. 3, 2391
Ioana, Adrian	Vol. 3, 1639	Malliaris, Maryanthe	Vol. 2, 83
Iovita, Adrian	Vol. 2, 249	Manolescu, Ciprian	Vol. 2, 1175
Iyama, Osamu	Vol. 2, 125	Martel, Yvan	Vol. 3, 2439
Jabin, Pierre-Emmanuel	Vol. 3, 2167	Masmoudi, Nader (المصمودي نادر)	
Jackson, Stephen	Vol. 2, 25		Vol. 3, 2137
James, Richard	Vol. 4, 3967	Máthé, András	Vol. 3, 1713
Jiang, Kai (蒋凯)	Vol. 4, 3591	Matomäki, Kaisa	Vol. 2, 321
Jin, Shi	Vol. 4, 3611	Mayboroda, Svitlana (Світлана Майборода)	
Johnson, William B.	Vol. 3, 1673		Vol. 3, 1691
Kalai, Yael	Vol. 4, 3337	Maynard, James	Vol. 2, 345
Kashaev, Rinat	Vol. 3, 2541	Mejías, Camilo	Vol. 4, 3489
Kassel, Fanny	Vol. 2, 1115	Miller, Jason	Vol. 4, 2945
Kawahigashi, Yasuyuki (河東泰之)		Mishra, Siddhartha	Vol. 4, 3641
	Vol. 3, 2597	Mj, Mahan	Vol. 2, 885
Keel, Sean	Vol. 2, 671	Möller, Martin	Vol. 3, 2017
Keevash, Peter	Vol. 4, 3113	Montanari, Andrea	Vol. 4, 2973
Kenyon, Richard	Vol. 4, 3137	Morales, Tomás	Vol. 4, 3515
Kerz, Moritz	Vol. 2, 163	Morris, Robert	Vol. 4, 3059
Keum, JongHae (금종해)	Vol. 2, 699	Mouhot, Clément	Vol. 3, 2467
Khanin, Konstantin	Vol. 3, 1973	Mukherjee, Debankur	Vol. 4, 3893
Kiselev, Alexander (Александр Киселев)		Munshi, Ritabrata	Vol. 2, 363
	Vol. 3, 2363	Nassiri, Meysam (میسام نصیری)	
Koenigsmann, Jochen	Vol. 2, 45		Vol. 3, 1995
Kohlenbach, Ulrich	Vol. 2, 61	Natale, Sonia	Vol. 2, 173
Koltchinskii, Vladimir	Vol. 4, 2903	Navas, Andrés	Vol. 3, 2035

Némethi, András	Vol. 2, 749	Shcherbina, Tatyana	Vol. 3, 2687
Nonnenmacher, Stéphane	Vol. 3, 2495	Singer, Amit	Vol. 4, 3995
Nussenzveig Lopes, Helena	Vol. 3, 2519	Smith, Ivan	Vol. 2, 969
Ortega, Sergio	Vol. 4, 3515	Steinberg, Benjamin	Vol. 3, 1583
Osin, Denis	Vol. 2, 919	Steurer, David	Vol. 4, 3389
Pak, Igor	Vol. 4, 3153	Sun, Song (孙崧)	Vol. 2, 993
Panin, Ivan	Vol. 2, 201	Szegedy, Balázs	Vol. 4, 3213
Pappas, Georgios	Vol. 2, 377	Tachikawa, Yuji	Vol. 3, 2709
Parés, Carlos	Vol. 4, 3515	Tang, Tao (汤涛)	Vol. 4, 3669
Park, Byeong	Vol. 4, 2995	Tardos, Gábor	Vol. 4, 3235
Petermichl, Stefanie	Vol. 3, 1733	Tayachi, Slim (سليم طيانشي)	Vol. 3, 2413
Pilloni, Vincent	Vol. 2, 249	Taylor, Jonathan	Vol. 4, 3019
von Plato, Jan	Vol. 4, 4057	Thom, Andreas	Vol. 3, 1779
Poltoratski, Alexei	Vol. 3, 1753	Thomas, Rekha	Vol. 4, 3819
Poonen, Bjorn	Vol. 2, 399	Thorne, Jack	Vol. 2, 415
Popa, Mihnea	Vol. 2, 781	Tiep, Pham Huu	Vol. 2, 223
Postnikov, Alexander	Vol. 4, 3181	Toint, Philippe	Vol. 4, 3711
Potrie, Rafael	Vol. 3, 2063	Toninelli, Fabio	Vol. 3, 2733
Prasad, Dipendra	Vol. 2, 1367	Tornberg, Anna-Karin	Vol. 4, 3691
Przytycki, Feliks	Vol. 3, 2087	Tóth, Bálint	Vol. 4, 3039
Radford, Luis	Vol. 4, 4037	Trélat, Emmanuel	Vol. 4, 3843
Radziwiłł, Maksym	Vol. 2, 321	Tsimerman, Jacob	Vol. 2, 435
Raghavendra, Prasad	Vol. 4, 3389	Uribe, Bernardo	Vol. 2, 1217
Razenshteyn, Ilya	Vol. 4, 3287	Vassilevska Williams, Virginia (Виргиния Василевска Уилиямс)	Vol. 4, 3447
Reid, Alan	Vol. 2, 1193	Vershynin, Roman	Vol. 4, 2925
Rindler, Filip	Vol. 3, 2215	Viazovska, Maryna	Vol. 2, 455
Roque, Tatiana	Vol. 4, 4075	Viehmann, Eva	Vol. 2, 1425
Rossman, Benjamin	Vol. 4, 3425	Walsh, Miguel	Vol. 2, 467
Rowe, David	Vol. 4, 4095	Wienhard, Anna	Vol. 2, 1013
Ruiz Baier, Ricardo	Vol. 4, 3489	Willwacher, Thomas	Vol. 2, 1241
Sagastizábal, Claudia	Vol. 4, 3797	Winter, Wilhelm	Vol. 3, 1801
Salomão, Pedro A. S.	Vol. 2, 941	Wormald, Nicholas	Vol. 4, 3245
Samotij, Wojciech	Vol. 4, 3059	Xu, Chenyang (许晨阳)	Vol. 2, 807
Sarkar, Sucharit	Vol. 2, 1153	You, Jiangong (尤建功)	Vol. 3, 2113
Schiffmann, Olivier	Vol. 2, 1393	Yun, Zhiwei (恽之玮)	Vol. 2, 1447
Schlein, Benjamin	Vol. 3, 2669	Zhang, Pingwen (张平文)	Vol. 4, 3591
Schramm, Tselil (צליל שרם)	Vol. 4, 3389	Zhang, Wei (张伟)	Vol. 2, 487
Shcherbina, Mariya	Vol. 3, 2687		



**World Scientific**  
[www.worldscientific.com](http://www.worldscientific.com)  
11060 hc

ISBN 978-981-3272-87-3 (set)



9 789813 272873

ISBN 978-981-3272-93-4 (v4)



9 789813 272934